

# Galaxy workflow guide for variant detection

---

1	Introduction.....	4
1.1	Reference materials .....	5
1.2	Outline of tutorial.....	5
2	Getting Started.....	6
2.1	Accessing Galaxy .....	6
2.2	Run “pe-sync: Paired-end synchronization” .....	7
2.3	Review Run “pe-sync: Paired-end synchronization” output.....	8
2.4	Run FastQC.....	9
3	Mapping with BWA.....	11
3.1	Map with BWA for illumina .....	12
4	Quality Assessment / Quality Control .....	16
4.1	Determine insert size distribution .....	17
4.2	Review insert size distribution plot.....	18
5	GATK Phase 1: Data Pre-processing.....	19
5.1	Removal of ambiguously-mapped and low-quality reads.....	22
5.2	Sorting Reads and updating mate-pair information.....	24
5.3	Removal of duplicates.....	25
	INDEL Realignment.....	26
5.4	Create Targets for Realignment.....	26
5.5	Realign reads around INDELS .....	29
5.6	Remove duplicates after INDEL realignment .....	30
	Base Quality Recalibration.....	31
5.7	Count Covariates (before base recalibration) .....	31
5.8	Analyze Covariates (before base recalibration) .....	33
5.9	Review Covariate plots (before base quality recalibration) .....	34
5.10	Recalibrate base quality scores.....	35
5.11	Count Covariates (after base recalibration) .....	36
5.12	Analyze Covariates (after base recalibration) .....	38
5.13	Review Covariate plots (after base quality recalibration) .....	39
6	GATK Phase 2: Variant Discovery.....	42
6.1	Variant detection using Unified Genotyper .....	43
6.2	Review Unified Genotyper results (Raw Variants) .....	47
7	GATK Phase 3: Preliminary Analysis .....	50
	Variant Recalibration .....	52
7.1	Select SNPs.....	52
7.2	Recalibrate SNPs .....	55
7.3	Apply recalibration .....	61
7.4	Review Variant Recalibration Models.....	62

7.5	Review Recalibrated Variants (SNPs) .....	63
7.6	Select INDELs.....	66
7.7	Recalibrate INDELs .....	68
7.8	Apply recalibration .....	72
7.9	Review Recalibrated Variants (INDELs).....	73
7.10	Combine SNPs and INDELs .....	76
	Variant Annotation .....	77
7.11	Annotate variants using SnpEff.....	77
7.12	Review Annotated Variants .....	79

## 1 Introduction

This guide was prepared using DNA sample from a European CEPH family (NA10858) obtained from the Coriell Cell Repository (Camden, NJ, USA). It outlines a typical variant detection analysis using GATK's best practices on the Galaxy framework. The sample dataset was obtained using Hybrid-capture with RNA baits. Some of the GATK tools such as VariantRecalibrator\_tk, as part of the input, VCF files containing known common variants "*true sites*", utilized for statistical training purposes. Data from 1000 Genomes project is used for training. It is publicly available on the 1000Genomes website (<http://www.1000genomes.org/data#DataAccess>) .

The following input files were used:

### **Sequence data:** fastq files

L7\_R1\_CAGATC\_Index\_7\_groomed.fastq  
L7\_R2\_CAGATC\_Index\_7\_groomed.fastq

### **Exon data:** bed files

tutorial\_exons.bed

### **Recalibration data (1000G data):** vcf files

dbsnp\_137.hg19.vcf  
Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf  
hapmap\_3.3.hg19.vcf  
1000G\_omni2.5.hg19.vcf  
1000G\_phase1.snps.high\_confidence.hg19.vcf

## ★ FASTQ Files

FASTQ format is a text-based format for storing a biological sequence and its corresponding quality scores. See [http://en.wikipedia.org/wiki/FASTQ\\_format](http://en.wikipedia.org/wiki/FASTQ_format).

## ★ Bed Files

BED format provides a way for defining genomic regions. We will use BED format to define target regions e.g., exons being targeted for sequence capture. The first three required fields specify: name of chromosome, start position and end position. For more information on BED format see <http://genome.ucsc.edu/FAQ/FAQformat.html#format1>.

## ★ VCF Files

Variant Call Format (VCF), not to be confused with the standard file format for storing contact information, is a specification for storing sequence variations. For more information on VCF format see [http://en.wikipedia.org/wiki/Variant\\_Call\\_Format](http://en.wikipedia.org/wiki/Variant_Call_Format).

## 1.1 Reference materials

- M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. [Nat. Genet. 43, 491\(2011\)](https://doi.org/10.1038/ng.1175).
- Genome Analysis Toolkit (GATK) website: <http://www.broadinstitute.org/gatk/>
- Summary of best practices for variant detection:  
<http://www.broadinstitute.org/gatk/guide/best-practices>
- BWA manual: <http://bio-bwa.sourceforge.net/>
- SAMtools: <http://samtools.sourceforge.net/>
- Picard-Tools: <http://picard.sourceforge.net/>
- SNP effect predictor: <http://snpeff.sourceforge.net/>
- Galaxy screencasts: [galaxycast.org](http://galaxycast.org)

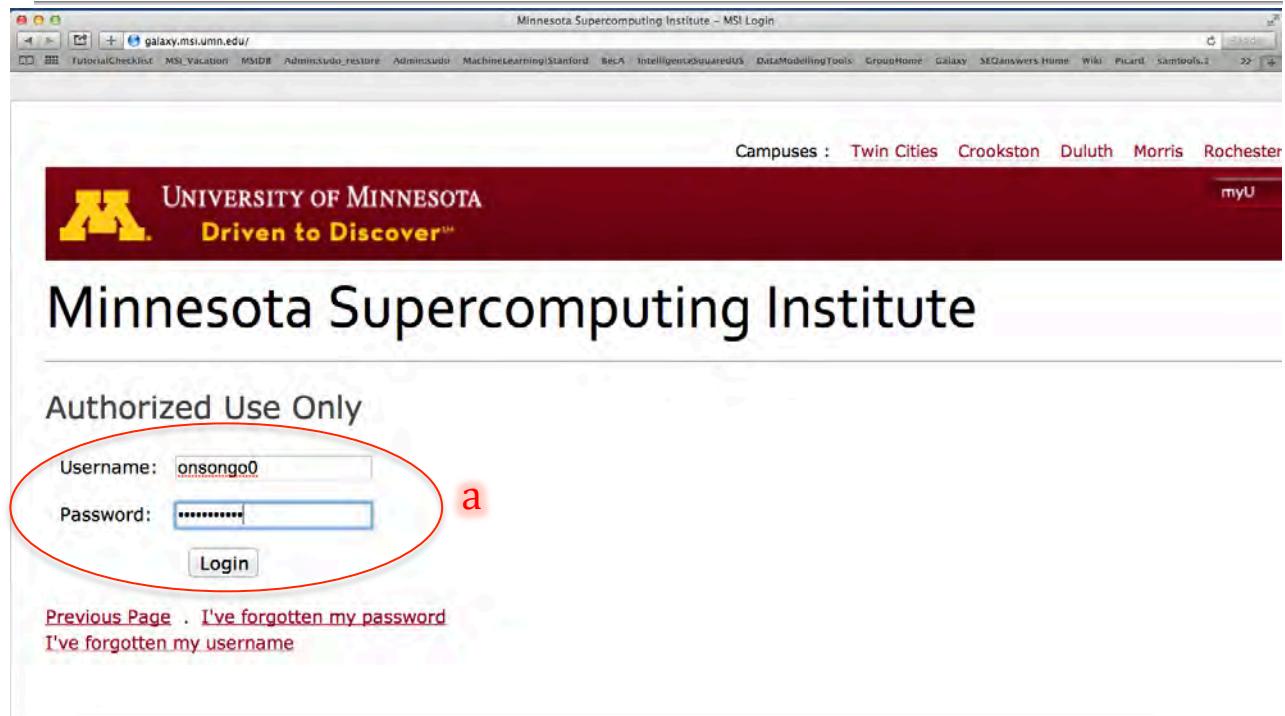
## 1.2 Outline of tutorial

- 1 Introduction
- 2 Getting Started
- 3 Mapping with BWA
- 4 Quality Assessment / Quality Control
- 5 GATK Phase 1: Data Pre-processing
  - INDEL Realignment
  - Base Quality Recalibration
- 6 GATK Phase 2: Variant Discovery
- 7 GATK Phase 3: Preliminary Analysis
  - Variant Recalibration
  - Variant Annotation

## 2 Getting Started

### 2.1 Accessing Galaxy

- Log in with your username and password



The screenshot shows the Galaxy / UMN interface. The left side features a "Tools pane" with a list of available tools, including "Get Data", "Send Data", "ENCODE Tools", "Lift-Over", "Text Manipulation", "Filter and Sort", "Join, Subtract and Group", "Convert Formats", "Extract Features", "Fetch Sequences", "Fetch Alignments", "Get Genomic Scores", "Operate on Genomic Intervals", "Statistics", "Wavelet Analysis", "Graph/Display Data", "Regional Variation", "Multiple regression", "Multivariate Analysis", "Evolution", "Motif T", "Multiple Alignments", "Metagenomic analyses", "Metagenomics Mothur", "FASTA manipulation", "NCBI BLAST+", and "NGS: QC and manipulation". The center pane displays the large Minnesota "M" logo. Below it, text reads: "To facilitate the communications between you, the galaxy-umn project team and other users, we would like to ask you to use the help forum pages to post your questions, report issues or make suggestions. <http://galaxy.msi.umn.edu/help>". At the bottom, it says: "This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.". The right side features a "History pane" which is currently empty, displaying the message: "Your history is empty. Click 'Get Data' on the left pane to start". A red box labeled "History pane" is drawn around this area.

## 2.2 Run “pe-sync: Paired-end synchronization”

The *pe-sync* tool checks to make sure the forward and reverse reads are synchronized.

- a) Load the *pe-sync* tool from the tool pane: “MSI -> pe-sync: Paired-end synchronization check”
- b) Set input files: From the dropdown menu under
  - Input 1:** select “L7\_R1\_CAGATC\_Index\_7\_groomed.fastq”
  - Input 2:** select “L7\_R2\_CAGATC\_Index\_7\_groomed.fastq”
- c) Click “Execute”

This screenshot shows the Galaxy interface. On the left, the 'Tools' panel is open, with the 'MSI' section highlighted by a red box and labeled 'a'. Under 'MSI', there is a description of the 'pe-sync' tool. In the center, there is a large maroon 'M' logo. Below it, a message says: 'If you have questions or concerns, please e-mail help@msi.umn.edu.' On the right, the 'History' panel shows several items, including 'Variant\_Detection\_RISS' (132.8 MB), 'Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf', 'hapmap\_3.3.hg19.vcf', 'dbsnp\_137.hg19.vcf', '1000G\_phase1snps.high\_confidence.hg19.vcf', '1000G\_omni2.5.hg19.vcf', 'tutorial\_exons.bed', 'L7\_R2\_CAGATC\_Index\_7\_groomed.fastq', and 'L7\_R1\_CAGATC\_Index\_7\_groomed.fastq'. The 'L7\_R1\_CAGATC\_Index\_7\_groomed.fastq' entry is highlighted with a red box and labeled 'c'.

This screenshot shows the Galaxy interface with the 'pe-sync' tool selected. The 'Input 1:' dropdown is set to '1: L7\_R1\_CAGATC\_Index\_7\_groomed.fastq' and the 'Input 2:' dropdown is set to '2: L7\_R2\_CAGATC\_Index\_7\_groomed.fastq'. Both dropdowns are highlighted with a red box and labeled 'b'. Below these dropdowns is a blue 'Execute' button, which is circled with a red circle and labeled 'c'. The rest of the interface is identical to the one in the previous screenshot.

## 2.3 Review Run “pe-sync: Paired-end synchronization” output

- In the history pane click on the *eye icon* next to the *pe-sync* output file to display output on the center pane
- Verify data is synchronized

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", "User", and "Using 1.2 GB". The left sidebar under "Tools" lists various NGS and SNP/WGA analysis tools. A green message box in the center states: "The following job has been successfully added to the queue: 9: pe-sync report for L7\_R1\_CAGATC\_Index\_7\_groomed.fastq and L7\_R2\_CAGATC\_Index\_7\_groomed.fastq You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered." To the right, the "History" pane lists several jobs:

- Variant\_Detection\_RISS (132.8 MB)
- 9: pe-sync report for L7\_R1\_CAGATC\_Index\_7\_groomed.fastq and L7\_R2\_CAGATC\_Index\_7\_groomed.fastq (status: running, circled with red 'a')
- 8: Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf (status: running)
- 7: hapmap\_3.3.hg19.vcf (status: running)
- 6: dbsnp\_137.hg19.vcf (status: running)

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", "User", and "Using 1.2 GB". The left sidebar under "Tools" lists various NGS and SNP/WGA analysis tools. A message box in the center states: "Casava 1.7 read id style PASSED full check" (circled with red 'b'). To the right, the "History" pane lists several jobs:

- Variant\_Detection\_RISS (132.8 MB)
- 9: pe-sync report for L7\_R1\_CAGATC\_Index\_7\_groomed.fastq and L7\_R2\_CAGATC\_Index\_7\_groomed.fastq
- 8: Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf (status: running)

## 2.4 Run FastQC

- Load the FastQC tool from the tool pane: "NGS: QC and manipulation -> FastQC: Read QC..."
- Set the input file to select "L7\_R1\_CAGATC\_Index\_7\_groomed.fastq" from the dropdown menu under "Short read data from your current history"
- Click "Execute"

**Galaxy / UMN**

Analyze Data Workflow Shared Data Visualization Help User Using 1.2 GB

Tools

**NCBI BLAST+**

**NGS: QC and manipulation**

- Trim sequences
- Reverse-Complement
- Rename sequences
- Compute quality statistics
- Draw nucleotides distribution chart
- Collapse sequences
- Clip adapter sequences
- Barcode Splitter
- Remove sequencing artifacts
- FASTQ to FASTA converter
- Filter by quality
- Quality format converter (ASCII-Numeric)
- Draw quality score box plot
- FASTQC: FASTQ/SAM/NIM
- FastQC:Read QC reports using FastQC**

Casava 1.7 read id style  
PASSED full check

History

- Variant\_Detection\_RISS 132.8 MB
- 9: pe-sync report for L7\_R1\_CAGATC\_Index\_7\_groomed.fastq and L7\_R2\_CAGATC\_Index\_7\_groomed.fastq
- 8: Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf
- 7: hapmap\_3.3.hg19.vcf
- 6: dbsnp\_137.hg19.vcf
- 5: 1000G\_phase1.snps.high\_confidence.hg19.vcf
- 4: 1000G\_omni2.5.hg19.vcf
- 3: tutorial\_exons.bed
- 2: L7\_R2\_CAGATC\_Index\_7\_groomed.fastq
- 1: L7\_R1\_CAGATC\_Index\_7\_groomed.fastq

**Galaxy / UMN**

Analyze Data Workflow Shared Data Visualization Help User Using 1.2 GB

Tools

**NCBI BLAST+**

**NGS: QC and manipulation**

- Trim sequences
- Reverse-Complement
- Rename sequences
- Compute quality statistics
- Draw nucleotides distribution chart
- Collapse sequences
- Clip adapter sequences
- Barcode Splitter
- Remove sequencing artifacts
- FASTQ to FASTA converter

**FastQC:Read QC (version 0.52)**

**Short read data from your current history**  
1: L7\_R1\_CAGATC\_Index\_7\_groomed.fastq

**Title for the output file – to remind you what the job was for:**  
FastQC

Letters and numbers only please – other characters will be removed

**Contaminant list:**  
Selection is Optional

tab delimited file with 2 columns: name and sequence.  
For example: Illumina Small RNA RT Primer  
CAAGCCAAAGACGGCATACGA

**Execute**

History

- Variant\_Detection\_RISS 132.8 MB
- 9: pe-sync report for L7\_R1\_CAGATC\_Index\_7\_groomed.fastq and L7\_R2\_CAGATC\_Index\_7\_groomed.fastq
- 8: Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf
- 7: hapmap\_3.3.hg19.vcf
- 6: dbsnp\_137.hg19.vcf
- 5: 1000G\_phase1.snps.high\_confidence.hg19.vcf
- 4: 1000G\_omni2.5.hg19.vcf

- d) In the history pane click on the name of the *FastQC* output file to display the tool in the center window to re-run the program
- e) Click the blue arrowed circle to display the *FastQC* tool in the center window to re-run the program
- f) Set the input file: select “L7\_R2\_CAGATC\_Index\_7\_groomed.fastq” from the dropdown menu under “Short read data from your current history”
- g) Click “Execute”

The following job has been successfully added to the queue:

10:  
FastQC\_L7\_R1\_CAGATC\_Index\_7\_groomed.fastq

You can check the status of queued jobs and view the resulting data by refreshing the History pane.

**d**

The following job has been successfully added to the queue:

10:  
FastQC\_L7\_R1\_CAGATC\_Index\_7\_groomed.fastq

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

**e**

FastQC:Read QC (version 0.52)

**f**

**Short read data from your current history:**  
2: L7\_R2\_CAGATC\_Index\_7\_groomed.fastq

**Title for the output file – to remind you what the job was for:**  
FastQC

Letters and numbers only please – other characters will be removed

**Contaminant list:**  
Selection is Optional  
tab delimited file with 2 columns: name and sequence.  
For example: Illumina Small RNA RT Primer  
CAAGCAGAAGACGGCATACGA

**Execute** **g**

**History**

Variant\_Detection\_RISS  
132.8 MB

10:  
FastQC\_L7\_R1\_CAGATC\_Index\_7\_groomed.fastq.html  
8.3 KB  
format: html, database: hg19\_canonical

HTML file

9: pe-sync report for  
L7\_R1\_CAGATC\_Index\_7\_groomed.fastq and  
L7\_R2\_CAGATC\_Index\_7\_groomed.fastq

8:  
Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf

### 3 Mapping with BWA

For mapping illumina data to a reference genome, BWA is the recommended aligner. Among its many advantages such as accuracy and speed, it emits BAM files natively. GATK only supports the BAM format for mapped reads. For more information on BAM file format see <http://samtools.sourceforge.net/SAM1.pdf>.

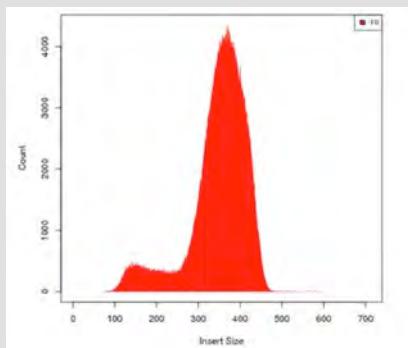
Other aligners besides BWA can be used, provided that the BAM files satisfy the GATK's formatting requirements (see the GATK's website for more details). This section illustrates how to specify the required fields to ensure that your BAM files adhere to the GATK's formatting requirements.

#### ★ Reference sequence

Reference sequences must be sorted in the order of one of the official b3x (e.g., b36, b37) or hg1x (e.g., hg18, hg19) references. A reference sequence adhering to this requirement is available in Galaxy (hg19\_canonical).

#### ★ Insert Size

When using BWA, users will need to specify a maximum expected insert size and median insert size. The maximum expected insert size is used to determine if a read pair is mapped properly. BWA should be able to infer this information from aligned reads and only uses the specified field if there are not enough good alignments. For this tutorial data, we will use 1000 for maximum insert size and 400 for median insert size. This information can sometimes be obtained from the sequencing center but later in the tutorial we will show how to plot an insert size distribution histogram from mapped reads to determine these values. Below is a typical insert size distribution histogram.



#### ★ Read Groups

Many downstream processes can take in multiple alignment files, or merged alignment files, that mix reads that were run in different lanes, on different sequencers, on different dates, etc. In order to detect systematic biases that may be introduced by any of these factors, and know which reads came from where, the BAM file specification allows the assignment of "Read Groups" to any collection of reads that logically were produced together. The GATK requires that all BAM files must list the read groups with sample names in the header and every read must belong to a read group. Consequently, when running BWA the full parameters list must be used with read groups, library name, sample name and platform used to produce the reads specified. Even though not required, we encourage you to specify optional parameters such as *sequencing center (CN)* and *date run was produced (DT)*. If these features are not specified, there will be no way to go back and determine if systematic biases occurred at a given center or on a given date.

### 3.1 Map with BWA for illumina

- Load BWA tool from the tool pane: "NGS: Mapping -> Map with BWA for illumina"
- Select a reference genome -> hg19\_canonical
- Is this library mate-paired? -> Paired-end
- Forward FASTQ file, forward reads -> L7\_R1\_CAGATC\_Index\_7\_groomed.fastq
- Reverse FASTQ file, reverse reads -> L7\_R2\_CAGATC\_Index\_7\_groomed.fastq
- BWA settings to use -> Full Parameter List

The screenshot shows the Galaxy interface with the following details:

- Tools Panel:** Shows various NGS tools including "NGS: Assembly", "NGS: Mapping", "Lastz paired reads map short paired reads against reference sequence", "Lastz map short reads against reference sequence", "Map with Bowtie for SOLiD", "Map with Bowtie for Illumina", "ILLUMINA", and "Map with BWA for Illumina". The "Map with BWA for Illumina" tool is highlighted with a red oval and an arrow pointing to it from the left.
- Report View:** Displays a "FastQC Report" for "L7\_R2\_CAGATC\_Index\_7\_groomed.fastq" dated Sat 16 Nov 2013.
- Summary View:** Lists three checked items: "Basic Statistics", "Per base sequence quality", and "Per sequence quality scores".
- History View:** Shows a list of recent history items including "Variant\_Detection\_RISS", "FastQC\_L7\_R2\_CAGATC\_Index\_7\_groomed.fastq.html", "FastQC\_L7\_R1\_CAGATC\_Index\_7\_groomed.fastq.html", "pe-sync report for L7\_R1\_CAGATC\_Index\_7\_groomed.fastq and L7\_R2\_CAGATC\_Index\_7\_groomed.fastq", and "Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf".

The screenshot shows the Galaxy interface with the "Map with BWA for Illumina" tool configuration dialog open. The configuration steps are labeled as follows:

- Reference genome selection: "hg19\_canonical" is selected in the dropdown menu.
- Library type selection: "Paired-end" is selected in the dropdown menu.
- Forward FASTQ file selection: "L7\_R1\_CAGATC\_Index\_7\_groomed.fastq" is selected in the dropdown menu.
- Reverse FASTQ file selection: "L7\_R2\_CAGATC\_Index\_7\_groomed.fastq" is selected in the dropdown menu.
- BWA settings selection: "Full Parameter List" is selected in the dropdown menu.

The History panel on the right shows the same items as the previous screenshot.

- g) Maximum insert size for a read to be considered as being mapped properly (sampe -a): -> 1000
- h) Specify the read group for this file -> Yes
- i) Read group identifier (ID). -> NA\_10858\_400
- j) Sequencing center that produced the read (CN): -> UMGC

**Galaxy / UMN**

Analyze Data Workflow Shared Data Visualization Help User Using 1.2 GB

Tools

NGS: Assembly  
NGS: Mapping

Lastz paired reads map short paired reads against reference sequence  
Lastz map short reads against reference sequence  
Map with Bowtie for SOLiD  
Map with Bowtie for Illumina  
ILLUMINA  
Map with BWA for Illumina  
Bowtie2 is a short-read aligner  
Map with BFAST  
SSAHA2 pairwise sequence alignment program  
Megablast compare short reads against htgs, nt, and wgs databases  
Parse blast XML output  
Map with PerlM for SOLiD and Illumina  
Re-align with CRMA

Maximum insert size for a read pair to be considered as being mapped properly (sampe -a): **1000** g

For paired-end reads only. Only used when there are not enough good alignments to infer the distribution of insert sizes

Maximum occurrences of a read for pairing (sampe -o): **100000**

For paired-end reads only. A read with more occurrences will be treated as a single-end read. Reducing this parameter helps faster pairing

Specify the read group for this file? (samse/sampe -r): **Yes** h

Read group identifier (ID). Each @RG line must have a unique ID. The value of ID is used in the RG tags of alignment records. Must be unique among all read groups in header section: **NA\_10858\_400** i

Required if RC specified. Read group IDs may be modified when merging SAM files in order to handle collisions.

Sequencing center that produced the read (CN): **UMGC** j

Optional

History

Variant\_Detection\_RISS  
133.5 MB

11: FastQC\_L7\_R2\_CAGATC\_Index\_7\_groomed.fastq.html  
10: FastQC\_L7\_R1\_CAGATC\_Index\_7\_groomed.fastq.html  
9: pe-sync report for L7\_R1\_CAGATC\_Index\_7\_groomed.fastq and L7\_R2\_CAGATC\_Index\_7\_groomed.fastq  
8: Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf  
7: hapmap\_3.3.hg19.vcf  
6: dbsnp\_137.hg19.vcf  
5: 1000G\_phase1.snp.high\_confidence.hg19.vcf  
4: 1000G\_omni2.5.hg19.vcf  
3: tutorial\_exons.bed  
2: L7\_R2\_CAGATC\_Index\_7\_groomed.fastq

- k) Description (DS): -> Coriell\_HapMap\_400bpInsert
- l) Date that run was produced (DT): -> 2013-11-20

**WARNING!!** When entering the date, DO NOT use any other characters between the year, date and month e.g., 2013\_11\_20 WILL NOT WORK. GATK expects a DATE **data type** specified using dashes as shown (2011-05-11). Using any other characters will result in GATK producing an error message.

- m) Library name (LB): -> NA\_10858

The screenshot shows the Galaxy web interface with the following details:

- Tools Panel:** On the left, a sidebar lists various bioinformatics tools under categories like NGS: Assembly, NGS: Mapping, Lastz, Map with Bowtie, and Illumina.
- Workflow Step:** The main area shows a step titled "Variant\_Detection\_RISS". The step has a size of "133.5 MB" and a status of "Using 1.2 GB".
- Configuration Fields:**
  - Description (DS):** A text input field containing "Coriell\_HapMap\_400bpIn". This field is circled in red with the letter "k" above it.
  - Date that run was produced (DT):** A text input field containing "2013-11-20". This field is circled in red with the letter "l" above it.
  - Library name (LB):** A text input field containing "NA\_10858". This field is circled in red with the letter "m" above it.
  - Programs used for processing the read group (PG):** An empty text input field.
- History:** A panel on the right shows a history of previous steps and artifacts, including "FastQC\_L7\_R2\_CAGATC\_Index\_7\_groomed.fastq.html", "FastQC\_L7\_R1\_CAGATC\_Index\_7\_groomed.fastq.html", and several VCF files (e.g., "Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf", "hapmap\_3.3.hg19.vcf", "dbsnp\_137.hg19.vcf").

- n) Predicted median insert size (PI): -> 400
- o) Platform/technology used to produce the reads (PL): -> ILLUMINA
- p) Platform unit (PU): -> HWUSI-EAS1737:7
- q) Sample (SM): -> NA\_10858
- r) Click "Execute"

The screenshot shows the Galaxy / UMN web interface. On the left, a sidebar lists various NGS tools. The main area displays the configuration for the 'NGS: Mapping' tool. Several input fields are highlighted with red circles and letters:

- Predicted median insert size (PI):** 400 (highlighted with a red circle labeled 'n')
- Platform/technology used to produce the reads (PL):** ILLUMINA (highlighted with a red circle labeled 'o')
- Platform unit (PU):** HWUSI-EAS1737:7 (highlighted with a red circle labeled 'p')
- Sample (SM):** NA\_10858 (highlighted with a red circle labeled 'q')
- Execute** button (highlighted with a red circle labeled 'r')

On the right, the 'History' panel shows a list of completed workflows and their outputs.

## 4 Quality Assessment / Quality Control

### ★ Insert Size Distribution

In section 3, we had to input, as parameters, the maximum insert size and median insert size into BWA when mapping reads to a reference genome. These values can be obtained from the sequencing center. Alternatively, one can first map reads with BWA using “Commonly Used” parameters and use the resultant output as input to a tool in Galaxy (available under Picard-Tools) to plot an insert size distribution histogram.

In addition to providing input parameter values to BWA, the insert size distribution histogram serves as an addition verification step for data integrity. Recall, BWA should be able to infer insert sizes from aligned reads and only uses supplied information if there are not enough good alignments. Generating this insert size distribution plot thus provides additional confirmation that the appropriate insert size was used. A distribution histogram differing widely from the expect insert size distribution should serve as a red flag.

## 4.1 Determine insert size distribution



- Load insert size metrics tool from the tool pane: "NGS: Picard (beta)-> Insertion size metrics for PAIRED data"
- SAM/BAM dataset to generate statistics for: -> "...Map with BWA for Illumina on data ....."
- Click "Execute"

**Galaxy / UMN**

Analyze Data Workflow Shared Data Visualization Help User Using 219.7 MB

Tools

- NCBI BLAST+
- NGS: QC and manipulation
- NGS: Picard (beta)**
- FASTQ to BAM creates an unaligned BAM file
- SAM to FASTQ creates a FASTQ file
- BAM Index Statistics
- SAM/BAM Alignment Summary Metrics
- SAM/BAM GC Bias Metrics
- Estimate Library Complexity
- Insertion size metrics for PAIRED data**
- SAM/BAM Hybrid Selection Metrics for targeted resequencing data
- Add or Replace Groups

If you have questions or concerns, please e-mail help@msi.umn.edu.

MSI Galaxy Updates

- Galaxy updated to version 2013.11.04

History

- Variant\_Detection\_RISS 352.5 MB
- 12: Map with BWA for Illumina on data 2 and data 1: mapped reads
- 11: FastQC\_L7\_R2\_CAGATC\_Index\_7\_groomed.fastq.html
- 10: FastQC\_L7\_R1\_CAGATC\_Index\_7\_groomed.fastq.html
- 9: pe-sync report for L7\_R1\_CAGATC\_Index\_7\_groomed.fastq\_and L7\_R2\_CAGATC\_Index\_7\_groomed.fastq
- 8: Mills\_and\_1000G\_gold\_standard.in

**Galaxy / UMN**

Analyze Data Workflow Shared Data Visualization Help User Using 219.7 MB

Tools

- NCBI BLAST+
- NGS: QC and manipulation
- NGS: Picard (beta)**
- FASTQ to BAM creates an unaligned BAM file
- SAM to FASTQ creates a FASTQ file
- BAM Index Statistics
- SAM/BAM Alignment Summary Metrics
- SAM/BAM GC Bias Metrics
- Estimate Library Complexity
- Insertion size metrics for PAIRED data
- SAM/BAM Hybrid Selection Metrics for targeted resequencing data
- Add or Replace Groups
- Reorder SAM/BAM
- Replace SAM/BAM Header
- Paired Read Mate Fixer for paired data

Insertion size metrics (version 1.56.0)

**SAM/BAM dataset to generate statistics for:** 12: Map with BWA for Illumina on data 2 and data 1: mapped reads  
If empty, upload or import a SAM/BAM dataset.

**Title for the output file:** Insertion size metrics  
Use this remind you what the job was for

**Deviations:** 10.0  
See Picard documentation: Generate mean, sd and plots by trimming the data down to MEDIAN + DEVIATIONS\*MEDIAN\_ABSOLUTE\_DEVIATION

**Histogram width:** 0  
Explicitly sets the histogram width option – leave 0 to ignore

**Minimum percentage:** 0.05  
Discard any data categories (out of FR, TANDEM, RF) that have fewer than this percentage of overall reads

**Metric Accumulation Level:** All reads (default)  
Sample  
Library  
Read group  
Level(s) at which metrics will be accumulated

**Execute** C

History

- Variant\_Detection\_RISS 352.5 MB
- 12: Map with BWA for Illumina on data 2 and data 1: mapped reads
- 11: FastQC\_L7\_R2\_CAGATC\_Index\_7\_groomed.fastq.html
- 10: FastQC\_L7\_R1\_CAGATC\_Index\_7\_groomed.fastq.html
- 9: pe-sync report for L7\_R1\_CAGATC\_Index\_7\_groomed.fastq\_and L7\_R2\_CAGATC\_Index\_7\_groomed.fastq
- 8: Mills\_and\_1000G\_gold\_standard.in
- 7: hapmap\_3.3.hg19.vcf
- 6: dbsnp\_137.hg19.vcf

## 4.2 Review insert size distribution plot

- In the history pane click on the “eye” next to the name of the insert size metric tool output file
- If desired: right click on the image plot to download it as an image

The screenshot shows the Galaxy interface. On the left, a sidebar titled "Tools" lists various bioinformatics tools. In the center, a green success message box displays:

- A checkmark icon.
- The text: "The following job has been successfully added to the queue:"
- The job ID: "13: InsertSize\_Insertion size metrics.html"
- A note: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered."

On the right, the "History" pane lists several completed jobs:

- "Variant\_Detection\_RISS" (status: "finished", size: 352.5 MB)
- "13: InsertSize\_Insertion size metrics.html" (status: "finished", circled with red "a")
- "12: Map with BWA for Illumina on data 2 and data 1: mapped reads" (status: "finished")
- "11: FastQC\_L7\_R2\_CAGATC\_Index\_7\_gr" (status: "finished")

The screenshot shows the Galaxy interface. On the left, the "Tools" sidebar includes "Insert size metrics for PAIRED data". In the center, a histogram titled "Insert Size Histogram for All\_Reads" is displayed, showing a sharp peak around 400. A context menu is open over the plot, with the option "Save Image to the Desktop" highlighted with a red circle "b".

Below the histogram, a message states: "The following output files were created (click the filename to view/download a copy):".

Two files are listed:

- [CollectInsertSizeMetrics.log](#)
- [CollectInsertSizeMetrics.metrics.txt](#)

On the right, the "History" pane shows the same completed jobs as the previous screenshot, with the "13: InsertSize\_Insertion size metrics.html" job still circled with red "a".

## 5 GATK Phase 1: Data Pre-processing

The [Genome Analysis ToolKit](#) (GATK) is an open source java programming suite for NGS data handling and variant detection that was created in support of the 1000 Genomes Project. A best-practices pipeline for variant calling based on the GATK was published by MA DePristo et al. in [\*Nat. Genet.\* 43, 491\(2011\)](#). Updates regularly appear on the Broad's website: <http://www.broadinstitute.org/gatk/>. It contains three main phases: Data pre-processing, variant discovery and preliminary analysis.

### ★ GATK Phase 1: Data Pre-processing

In order to identify variants relative to a reference, mapping of the reads to the reference must be performed. We have already discussed mapping raw reads to the genome using BWA. However, the raw alignments generated by BWA, or by any genome mapping algorithm, are not of sufficient quality to identify real biological variants because they contain numerous systematic errors that must be filtered. In this phase, several steps are taken to clean up raw BAM files to get them ready for genotype calling.

### ★ GATK Phase 2: Variant Discovery

Once BAM files are cleaned of systematic artifacts, we can proceed with calling SNPs and Indels. (Structural Variants are outside the scope of this tutorial). The GATK's Unified Genotyper is the current industry standard, but is limited to diploid organisms in un-pooled samples. Accommodation of polyploid genomes and pooled samples has recently been introduced to newer versions of the GATK.

### ★ GATK Phase 3: Preliminary Analysis

Even when using filtered BAM files, some systematic machine artifacts pollute the raw variant calls produced by NGS pipelines. External verification tests have revealed profiles of unreliable calls (e.g., calls in regions of unusually high local depth of coverage that probably represent collapsed repeat regions, calls only supported by reads on one strand and not the other, or calls violating Hardy-Weinberg Equilibrium). Hence, in this phase, the GATK attempts to partition the raw calls into confidence classes or tranches, based on their inherent characteristics.

## ★ GATK Phase 1 details

The types of systematic biases that must be corrected in raw BAM files include:

- **Removal of ambiguously-mapped and low-quality reads**

BWA assigns a mapping quality (MAPQ) value of 0 to non-uniquely mapped reads. Read pairs that map equally well to multiple locations have *at best* a 50% chance of being mapped to the correct location, and hence are not typically suitable for variant detection. Additionally, reads that don't map in proper pairs (possibly involved in large structural rearrangements) also may reduce confidence for SNP and small indel calls. We use SAMTools (see <http://samtools.sourceforge.net/>) to remove these classes of problematic reads.

- **Sorting Reads and updating mate-pair information**

In addition to the input constraints by GATK, read files must be sorted in coordinate order with respect to the reference. We will use Picard-Tools to sort reads and ensure all the mate-pair information is in sync between each read and its mate pair.

- **Removal of PCR duplicates**

Most library preps, especially those that involve sequence capture, involve several rounds of PCR. Allele frequencies and genotype calls can be skewed if certain individual sequence fragments are preferentially amplified relative to others, as shown below. Hence all paired end fragments mapping to the exact same genomic coordinates should be reduced to one copy. We use Picard-Tools instead of SAMTools to remove duplicates because it considers both reads in the pair.



- **Indel Realignment**

Indels pose difficult challenges to mapping algorithms, especially in sequence regions with simple sequence repeats. Incorrect mapping across indels often leads to false-positive SNPs nearby as the image below illustrates. The GATK realigner target creator systematically goes through a BAM file and identifies all sequence positions where at least one read has an indel. The GATK indel realigner will then check every read at each flagged position and determine through a likelihood ratio test whether it better matches the reference sequence or the alternate indel call, resulting in a cleaned up BAM file as below. NOTE: inclusion of known indel sites (e.g., from the 1000 Genomes Project) in addition to novel sites will improve performance.

- **Base quality recalibration**

Sequencing machines simply do not report accurate base call qualities. The phred-scale quality they report for all bases in a run is directly testable empirically, when running on a well-characterized population. Since it is estimated that 99% of all variants in the Caucasian population have been deposited in dbSNP, the vast majority of mapping differences *not in dbSNP* should simply be sequencing errors. So, if we check base calls that had a raw quality of Q20, we would expect to find about 1 novel mapping discrepancy approximately every 100 such basecalls checked – but the real numbers often reveal systematic biases.

- **Base quality recalibration across covariates**

The GATK allows one to explore a breakdown of empirical vs. reported quality values across many covariates. For example, an Illumina run may systematically *differentially under- or over-estimate* base quality across the length of the read as shown below. As is typical in these runs, the reported base call qualities are least accurate in the beginning and the end of each read. The GATK's base quality recalibration routines can simultaneously correct for several different covariates at once (e.g., cycle, dinucleotides, homopolymer runs), but simultaneous optimization of many covariates can be difficult in practice. (i.e., If there are any dependencies among the covariates entered, "fixing" or over-fitting one may have an adverse effect on the others.)

### Phred Quality Scores

The base call quality scores (Q values) being recalibrated should not be confused with the other two Q scores used to assess mapping quality and variant call quality. The Phred quality score was developed by the program **Phred** to help in automation of DNA sequencing (see [http://en.wikipedia.org/wiki/Phred\\_quality\\_score](http://en.wikipedia.org/wiki/Phred_quality_score) for more details). It is a logarithmic link to error probabilities ( $Q = -10 \log_{10}P$ ) and can be used to assess the quality of any measure with error probabilities. In addition to being used to report base call qualities, Q scores is also used to measure mapping quality and variant call quality in NGS based variant detection.

- **Base Quality**

The base call quality Q score is a measure of how confident a sequencing machine is that the correct base call was made. For example, if 1 in 10 base calls are wrong, the probability of error is 1 in 10 ( $P = 0.1$ ). Recall,  $Q \text{ score} = -10 \log_{10}(0.1) = -10 \log_{10}(10^{-1}) = -10(-1) = 10$ . So the Q score for a base call with a 1 in 10 chance of being wrong is Q10. In a similar manner, if  $P=0.01$  (i.e., 1 error in 100 bases) this implies Q20.

- **Mapping Quality**

Mapping quality Q score is a measure of how well a sequenced read maps to a reference genome. A read that uniquely maps to a reference genome will have a higher Q score value relative to a read that maps equally well to several locations in the genome. Factors that contribute to the likelihood of mapping error and hence reduce reported mapping quality include: (1) the number of alternative equal-scoring mappings in the genome and (2) the number of high-quality basecall mismatches with the reference.

- **Variant Call Quality**

Variant call quality Q score is measure of the likelihood of the variant call being correct. A heterozygous (A/T) call with 200 bases matching the reference allele and 199 bases matching the alternate allele is expected to have a higher Q score value relative to a variant call with 1 base matching the reference allele and 1 base matching the alternate allele. A single wrong base call will completely change the second variant call (with only 2 supporting bases, 1 for the reference allele and the 1 for the alternate allele).

### Using the “Operate on Genomic intervals” (-L) analysis option

The GATK offers a -L analysis option (“Operate on Genomic intervals”) that restricts analysis to a specific part of the genome. This option can be very useful in reducing computation time when users have large datasets but are only interested in a small part of the genome. For example, a user with Whole Genome Sequence data might be interested in variants in a few genes. Restricting analysis to the genes of interest will significantly reduce computation time. You generally might NOT want to restrict analysis to the region of interest. For certain tools such as the *base quality score recalibrator* and *variant call recalibrator*, addition data from other regions of the genome will improve accuracy

## 5.1 Removal of ambiguously-mapped and low-quality reads

- In the history pane click on the Options wheel at the top (on the right side of the word "History") and click on "Saved Histories"
- Switch to the History "Variant\_Detection\_RISS"
- Load *Filter SAM* tool from the tool pane: "NGS: SAM Tools -> Filter SAM or BAM files on FLAG MAPQ RG LN or by region"

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". On the far right, it says "Using 219.7 MB". Below the navigation is a table of tools and their details. To the right of the table is a "History" pane. A red circle labeled 'a' highlights the "Saved Histories" button in the "History" pane.

The screenshot shows the "Saved Histories" page. The left sidebar lists various NGS tools. The main area displays a table of saved histories with columns for Name, Datasets, Tags, Sharing, Size on Disk, Created, Last Updated, and Status. A red circle labeled 'b' highlights the "Variant\_Detection\_RISS" entry in the table. A red arrow labeled 'b' points from the sidebar to this entry. The right side of the screen shows a detailed view of the selected history, which is titled "Variant\_Detection\_RISS".

The screenshot shows the "Saved Histories" page. The left sidebar has a red box around the "NGS: SAM Tools" section. A red arrow labeled 'c' points from this section to the "Variant\_Detection\_RISS" entry in the main table. The right side of the screen shows a detailed view of the selected history, which is titled "Variant\_Detection\_RISS".

- d) SAM/BAM dataset to generate statistics for: -> “....Map with BWA for Illumina on data .....
- e) Minimum MAPQ quality score: -> 1
- f) Filter on bitwise flag: -> “yes”
- g) In the *center pane*, scroll down to the section “**Only output alignments with all of these flag bits set:** and check boxes next to
- ✓ The read is mapped in a proper pair
- h) Click “Execute”

**Galaxy / UMN**

Analyze Data Workflow Shared Data Visualization Help User

Tools

NGS: SAM Tools

- rmdup remove PCR duplicates
- MPileup SNP and indel caller
- flagstat provides simple stats on BAM files
- SAM-to-BAM converts SAM format to BAM format
- Generate\_pileup from BAM dataset
- Merge BAM Files merges BAM files together
- Filter SAM on bitwise flag values
- Convert SAM to interval
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- BAM-to-SAM converts BAM format to SAM format
- Filter SAM or BAM files on FLAG MAPQ RG LN or by region
- BAM to fastq Old version – use

**Filter SAM or BAM (version 1.1.0)**

**SAM or BAM File to Filter:** 12: Map with BWA for Illumina on data 2 and data 1: mapped reads **d**

**Header in output:** Include Header

**Minimum MAPQ quality score:** 1 **e**

**Filter on bitwise flag:** yes **f**

**Only output alignments with all of these flag bits set:**

Select All Unselect All

Read is paired  
 Read is mapped in a proper pair **g**  
 The read is unmapped  
 The mate is unmapped  
 Read strand  
 Mate strand  
 Read is the first in a pair  
 Read is the second in a pair  
 The alignment or this read is not primary  
 The read fails platform/vendor quality checks  
 The read is a PCR or optical duplicate

History

Variant\_Detection\_RISS 387.5 MB

13: InsertSize\_Insertion\_size metrics.html

12: Map with BWA for Illumina on data 2 and data 1: mapped reads

11: FastQC\_L7\_R2\_CAGATC\_Index\_7\_groomed.fastq.html

10: FastQC\_L7\_R1\_CAGATC\_Index\_7\_groomed.fastq.html

9: pe-sync report for L7\_R1\_CAGATC\_Index\_7\_groomed.fasta and L7\_R2\_CAGATC\_Index\_7\_groomed.fasta

8: Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf

7: hapmap\_3.3.hg19.vcf

Using 219.7 MB

**Galaxy / UMN**

Analyze Data Workflow Shared Data Visualization Help User

Tools

NGS: SAM Tools

- rmdup remove PCR duplicates
- MPileup SNP and indel caller
- flagstat provides simple stats on BAM files
- SAM-to-BAM converts SAM format to BAM format
- Generate pileup from BAM dataset

**Filter SAM or BAM (version 1.1.0)**

**Selection is Optional** **d**

**Select regions (only used when the input is in BAM format):**

region should be presented in one of the following formats: 'chr1', 'chr2:1,000' and 'chr3:1000-2,000'

**Execute** **h**

**What it does**

History

Variant\_Detection\_RISS 387.5 MB

13: InsertSize\_Insertion\_size metrics.html

12: Map with BWA for Illumina on data 2 and data 1: mapped reads

11:

Using 219.7 MB

## 5.2 Sorting Reads and updating mate-pair information

- Load *paired-read mate fixer* tool from the tool pane: “NGS: Picard (beta) -> Paired Read Mate Fixer for paired data”
- SAM/BAM dataset to fix: -> “....Filter SAM or BAM ...”
- Sort order: -> Coordinate sort
- Output BAM instead of SAM: -> check (✓)
- Click “Execute”

The screenshot shows the Galaxy interface with the following details:

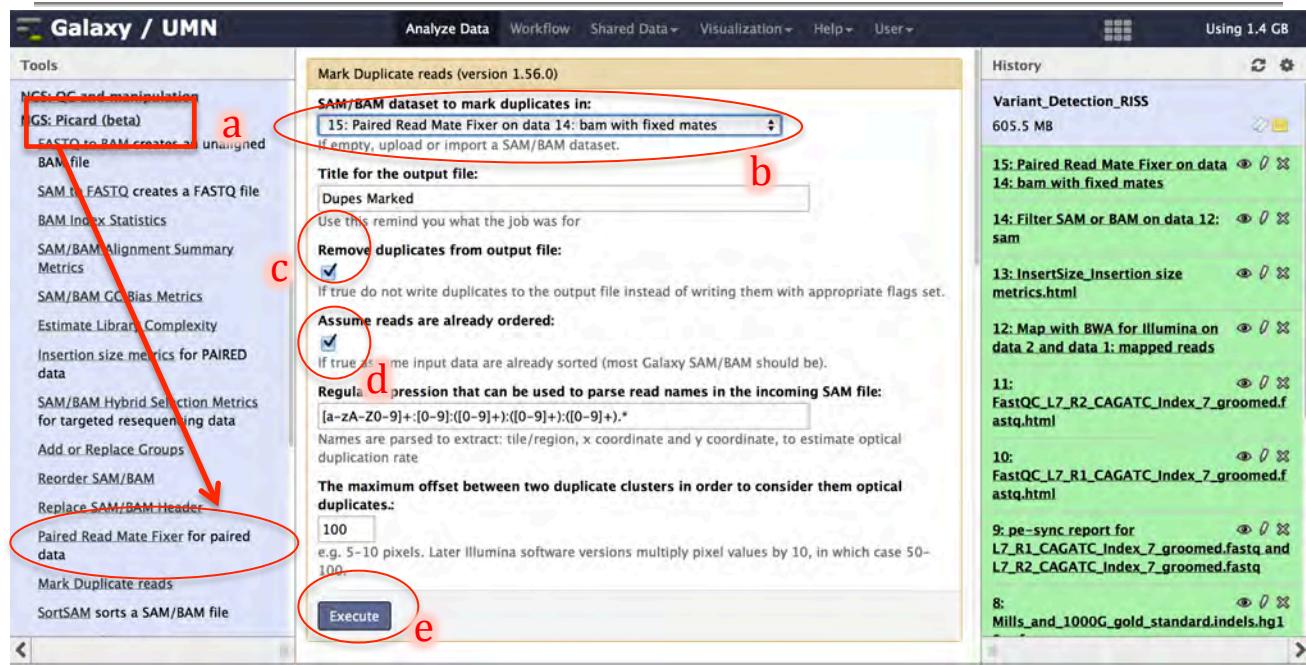
- Tools Panel:** Shows various NGS tools, with "Paired Read Mate Fixer for paired data" circled in red and labeled 'a'.
- Job Queue:** A green box indicates a successful job addition: "The following job has been successfully added to the queue: 14: Filter SAM or BAM on data 12: sam". Below it, instructions say: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered."
- History Panel:** Lists several completed jobs, including "Variant\_Detection\_RISS", "14: Filter SAM or BAM on data 12:", "13: InsertSize\_Insertion\_size", "12: Map with BWA for Illumina on data 2 and data 1:", "11: FastQC\_L7\_R2\_CAGATC\_Index\_7\_groomed.fasta.html", "10: FastQC\_L7\_R1\_CAGATC\_Index\_7\_groomed.fasta.html", "9: pe-sync report for L7\_R1\_CAGATC\_Index\_7\_groomed.fastq and L7\_R2\_CAGATC\_Index\_7\_groomed.fastq", and "8: Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf".

The screenshot shows the Galaxy interface with the following details:

- Tools Panel:** Shows various NGS tools, with "Paired Read Mate Fixer for paired data" circled in red and labeled 'a'.
- Tool Configuration:** The "Paired Read Mate Fixer (version 1.56.0)" dialog is open, with the following settings highlighted:
  - b:** "SAM/BAM dataset to fix" dropdown set to "14: Filter SAM or BAM on data 12: sam".
  - c:** "Sort order" dropdown set to "Coordinate sort".
  - d:** "Output BAM instead of SAM" checkbox checked (indicated by a red circle).
  - e:** "Execute" button.
- History Panel:** Lists several completed jobs, including "Variant\_Detection\_RISS", "14: Filter SAM or BAM on data 12:", "13: InsertSize\_Insertion\_size", "12: Map with BWA for Illumina on data 2 and data 1:", "11: FastQC\_L7\_R2\_CAGATC\_Index\_7\_groomed.fasta.html", "10: FastQC\_L7\_R1\_CAGATC\_Index\_7\_groomed.fasta.html", "9: pe-sync report for L7\_R1\_CAGATC\_Index\_7\_groomed.fastq and L7\_R2\_CAGATC\_Index\_7\_groomed.fastq", and "8: Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf".

### 5.3 Removal of duplicates

- Load *mark duplicates* tool from the tool pane: "NGS: Picard (beta) -> Mark Duplicate reads"
- SAM/BAM dataset to mark duplicates in: -> "...Paired Read Mate Fixer on data..."
- Remove duplicates from output file: -> check (✓)
- Assume reads are already ordered: -> check (✓)
- Click "Execute"



# INDEL Realignment

## 5.4 Create Targets for Realignment

- Load *realigner target creator* tool from the tool pane: "NGS: GATK Tools -> Realigner Target Creator for use in local realignment"
- BAM file: -> "...MarkDups\_Dupes Marked.bam"
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Click the "Add new Binding for reference-ordered data" button

The screenshot shows the Galaxy interface with the following details:

- Tools Panel:** NGS: SAM Tools, NGS: GATK Tools (highlighted with a red box), ALIGNMENT UTILITIES, REALIGNMENT, BASE RECALIBRATION.
- Job Queue:** A green success message box displays:
  - The following job has been successfully added to the queue:
  - 16: MarkDups\_Dupes Marked.bam
  - 17: MarkDups\_Dupes Marked.html

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.
- History:** A list of completed jobs including:
  - Variant\_Detection\_RISS
  - 17: MarkDups\_Dupes Marked.html
  - 16: MarkDups\_Dupes Marked.bam
  - 15: Paired Read Mate Fixer on data
  - 14: bam with fixed mates
  - 14: Filter SAM or BAM on data 12: sam
  - 13: InsertSize\_Insertion size metrics.html

The screenshot shows the Realigner Target Creator tool configuration window:

- Choose the source for the reference list:** Locally cached (highlighted with a red box).
- BAM file:** 16: MarkDups\_Dupes Marked.bam (highlighted with a red box).
  - I,--input\_file <input\_file>
- Using reference genome:** Homo sapiens hg19\_canonical (GATK) (highlighted with a red box).
  - R,--reference\_sequence <reference\_sequence>
- Binding for reference-ordered data:** Add new Binding for reference-ordered data (highlighted with a red box).
- Basic or Advanced GATK options:** Basic (highlighted with a red box).
- Basic or Advanced Analysis options:** Basic (highlighted with a red box).

The right side of the screen shows the Galaxy History pane with the same list of completed jobs as the previous screenshot.

- e) Under “Binding Type:” select “dbSNP”
- f) Under “ROD file:” select “...dbsnp\_137.hg19.vcf”
- g) Click the “Add new Binding for reference-ordered data” button again
- h) Under “Binding Type:” select “INDELS”
- i) Under “ROD file:” select “...Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf”
- j) On the drop down menu below “Basic or Advanced GATK options:” select “Advanced”
- k) Click the “Add new Operate on Genomic intervals” button

Galaxy / UMN

Tools

NGS: SAM Tools  
NGS: GATK Tools  
ALIGNMENT UTILITIES  
Depth of Coverage on BAM files  
Print Reads from BAM files  
REALIGNMENT  
Realigner Target Creator for use in local realignment  
Indel Realigner – perform local realignment

Binding for reference-ordered data

Binding for reference-ordered data 1

Binding Type: dbSNP e

ROD file: 6: dbsnp\_137.hg19.vcf f

Add new Binding for reference-ordered data g

Remove Binding for reference-ordered data 1

History

Variant\_Detection\_RISS 641.3 MB

17: MarkDups\_Dupes.Marked.html ① 0 ✘  
16: MarkDups\_Dupes.Marked.bam ① 0 ✘  
15: Paired Read Mate Fixer.on data ① 0 ✘  
14: bam with fixed mates  
14: Filter SAM or BAM on data 12: ① 0 ✘ sam

Galaxy / UMN

Tools

NGS: SAM Tools  
NGS: GATK Tools  
ALIGNMENT UTILITIES  
Depth of Coverage on BAM files  
Print Reads from BAM files  
REALIGNMENT  
Realigner Target Creator for use in local realignment  
Indel Realigner – perform local realignment  
BASE RECALIBRATION  
Count Covariates on BAM files  
Table Recalibration on BAM files

Binding for reference-ordered data 2

Binding Type: INDELS h

ROD file: 8: Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf i

Add new Binding for reference-ordered data 2

Basic or Advanced GATK options: Advanced j

Pedigree files

-ped,--pedigree <pedigree>

Add new Pedigree file

History

Variant\_Detection\_RISS 641.3 MB

17: MarkDups\_Dupes.Marked.html ① 0 ✘  
16: MarkDups\_Dupes.Marked.bam ① 0 ✘  
15: Paired Read Mate Fixer.on data ① 0 ✘  
14: bam with fixed mates  
14: Filter SAM or BAM on data 12: ① 0 ✘ sam  
13: InsertSize\_Insertion.size metrics.html  
12: Map with BWA for Illumina on data 2 and data 1: mapped reads

Galaxy / UMN

Tools

NGS: SAM Tools  
NGS: GATK Tools  
ALIGNMENT UTILITIES  
Depth of Coverage on BAM files  
Print Reads from BAM files  
REALIGNMENT  
Realigner Target Creator for use in local realignment  
Indel Realigner – perform local realignment  
BASE RECALIBRATION  
Count Covariates on BAM files

How strict should we be in validating the pedigree information:

STRICT

Read Filters

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

Add new Read Filter

Operate on Genomic intervals

-L,--Intervals <Intervals>

Add new Operate on Genomic intervals k

Exclude Genomic intervals

-XL,--excludeIntervals <excludeIntervals>

Add new Exclude Genomic intervals

History

17: MarkDups\_Dupes.Marked.html ① 0 ✘  
16: MarkDups\_Dupes.Marked.bam ① 0 ✘  
15: Paired Read Mate Fixer.on data ① 0 ✘  
14: bam with fixed mates  
14: Filter SAM or BAM on data 12: ① 0 ✘ sam  
13: InsertSize\_Insertion.size metrics.html  
12: Map with BWA for Illumina on data 2 and data 1: mapped reads

- I) Under “Genomic intervals:” select the file “tutorial\_exons.bed”  
 m) Scroll down and and the drop down menu below “Basic or Advanced Analysis options:” select “Advanced”  
 n) Fraction of base qualities needing to ..... to have high entropy (mismatchFraction): -> 0  
 o) Click “Execute”

Analyze Data Workflow Shared Data Visualization Help User Using 1.4 GB

**Tools**

- NGS: SAM Tools
- NGS: GATK Tools
- ALIGNMENT UTILITIES
- Depth of Coverage on BAM files
- Print Reads from BAM files
- REALIGNMENT
- Realigner Target Creator for use in local realignment

-rf,--read\_filter <read\_filter>  
  
**Operate on Genomic intervals**  
 -L,--intervals <Intervals>  
**Operate on Genomic intervals 1**  
**Genomic intervals:**  
 3: tutorial\_exons.bed

History

- 17: MarkDups\_Dupes Marked.html
- 16: MarkDups\_Dupes Marked.bam
- 15: Paired Read Mate Fixer on data
- 14: bam with fixed mates
- 14: Filter SAM or BAM on data 12: sam

Analyze Data Workflow Shared Data Visualization Help User Using 1.4 GB

**Tools**

- NGS: SAM Tools
- NGS: GATK Tools
- ALIGNMENT UTILITIES
- Depth of Coverage on BAM files
- Print Reads from BAM files
- REALIGNMENT
- Realigner Target Creator for use in local realignment
- Indel Realigner – perform local realignment
- BASE CALIBRATION
- Count Covariates on BAM files
- Table Recalibration on BAM files
- Analyze Covariates – draw plots
- GENOTYPING
- Unified Genotyper SNP and indel caller
- ANNOTATION
- Variant Annotator
- FILTRATION
- Variant Filtration on VCF file

Disable experimental low-memory sharding functionality:  
 --disable\_experimental\_low\_memory\_sharding

Makes the GATK behave non deterministically, that is, the random numbers generated will be different in every run:  
 -ndrs,--nonDeterministicRandomSeed

**Basic or Advanced Analysis options:**  
 Advanced

Window size for calculating entropy or SNP clusters (windowSize):  
 10  
 -window,--windowSize <windowSize>

Fraction of base qualities needing to mismatch for a position to have high entropy (mismatchFraction):  
 0

to disable set to <= 0 or > 1 (-mismatch,--mismatchFraction <mismatchFraction>)

Minimum reads at a locus to enable using the entropy calculation (minReadsAtLocus):  
 4  
 -minReads,--minReadsAtLocus <minReadsAtLocus>

Maximum interval size:  
 500  
 -maxInterval,--maxIntervalSize <maxIntervalSize>

0

History

- 17: MarkDups\_Dupes Marked.html
- 16: MarkDups\_Dupes Marked.bam
- 15: Paired Read Mate Fixer on data
- 14: bam with fixed mates
- 14: Filter SAM or BAM on data 12: sam
- 13: InsertSize\_Insertion\_size metrics.html
- 12: Map with BWA for Illumina on data 2 and data 1: mapped reads
- 11: FastQC\_L7\_R2\_CAGATC\_Index\_7\_groomed.fastq.html
- 10: FastQC\_L7\_R1\_CAGATC\_Index\_7\_groomed.fastq.html
- 9: pe-sync\_report for L7\_R1\_CAGATC\_Index\_7\_groomed.fastq and L7\_R2\_CAGATC\_Index\_7\_groomed.fastq
- 8:

## 5.5 Realign reads around INDELS

- Load *indel realigner* tool from the tool pane: "NGS: GATK Tools -> Indel Realigner - perform local realignment"
- BAM file: -> "...MarkDups\_Dupes Marked.bam"
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Restrict realignment to provided intervals: -> "...Realigner Target Creator on data....."
- Click "Execute"

The screenshot shows the Galaxy interface with the following details:

- Tools Panel:** Shows the "NGS: GATK Tools" category selected. Within it, the "Indel Realigner – perform local realignment" tool is highlighted with a red box and labeled 'a'.
- Job Queue:** A green box displays a success message: "The following job has been successfully added to the queue: 18: Realigner Target Creator on data 6, data 16, and others (GATK intervals)". Below it, another message says: "19: Realigner Target Creator on data 6, data 16, and others (log)". A note states: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered."
- History:** On the right, the history pane lists several completed jobs, including "Variant\_Detection\_RISS", "19: Realigner Target Creator on data 6, data 16, and others (log)", "18: Realigner Target Creator on data 6, data 16, and others (GATK intervals)", "17: MarkDups\_Dupes Marked.html", "16: MarkDups\_Dupes Marked.bam", and "15: Paired Read Mate Fixer on data 14: bam with fixed mates".

The screenshot shows the configuration of the "Indel Realigner – perform local realignment" tool in the Galaxy interface. The parameters are set as follows:

- BAM file:** Set to "16: MarkDups\_Dupes Marked.bam" (circled in red, labeled 'b').
- Using reference genome:** Set to "Homo sapiens hg19\_canonical (GATK)" (circled in red, labeled 'c').
- Restrict realignment to provided intervals:** Set to "18: Realigner Target Creator on data 6, data 16, and others (GATK intervals)" (circled in red, labeled 'd').
- Execute:** The "Execute" button at the bottom left is highlighted with a red circle and labeled 'e'.

## 5.6 Remove duplicates after INDEL realignment

- Load *mark duplicates* tool from the tool pane: "NGS: Picard (beta) -> Mark Duplicate reads"
- SAM/BAM dataset to mark duplicates in: -> "...Indel Realigner.... (BAM)"
- Remove duplicates from output file: -> check (✓)
- Assume reads are already ordered: -> check (✓)
- Click "Execute"

The screenshot shows the Galaxy interface with the following details:

- Tools Panel:** Shows various bioinformatics tools. The "Mark Duplicate reads" tool is circled with a red arrow labeled "a".
- Job Queue:** A green box indicates a job has been successfully added to the queue: "20: Indel Realigner on data 18 and data 16 (BAM)".
- History Pane:** Shows a list of completed jobs, including "Variant\_Detection\_RISS" and several "Indel Realigner" steps.

The screenshot shows the "Mark Duplicate reads" tool configuration window:

- Input:** "SAM/BAM dataset to mark duplicates in:" dropdown is set to "20: Indel Realigner on data 18 and data 16 (BAM)".
- Output:** "Title for the output file:" is "Dupes Marked".
- Duplicates:** "Remove duplicates from output file:" checkbox is checked (circled with red arrow "b").
- Assumptions:** "Assume reads are already ordered:" checkbox is checked (circled with red arrow "c").
- Regular Expression:** "Regular expression that can be used to parse read names in the incoming SAM file:" is "[a-zA-Z0-9]+:[0-9]+:[0-9]+:[0-9]+.\*".
- Offset:** "The maximum offset between two duplicate clusters in order to consider them optical duplicates:" is set to "100".
- Execute Button:** A blue button labeled "Execute" is at the bottom (circled with red arrow "e").
- History Pane:** Shows a list of completed jobs, including "Variant\_Detection\_RISS" and several "Indel Realigner" steps.

# Base Quality Recalibration

## 5.7 Count Covariates (before base recalibration)

- Load *count covariates* tool from the tool pane: “NGS: GATK Tools -> Count Covariates on BAM files”
- BAM file: -> “...MarkDups\_Dupes Marked.bam” (be sure to select the file generated after indel realignment)
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Covariates to be used in the recalibration: -> check boxes next to
  - ✓ ReadGroupCovariate
  - ✓ QualityScoreCovariate
  - ✓ CycleCovariate
  - ✓ DinucCovariate
- Click the “Add new Binding for reference-ordered data” button

**a**

The following job has been successfully added to the queue:  
22: MarkDups\_Dupes Marked.bam  
23: MarkDups\_Dupes Marked.html

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

**b**

**c**

**d**

**e**

- f) Binding Type: -> dbSNP
- g) ROD file: -> dbsnp\_137.hg19.vcf
- h) Click "Execute"

The screenshot shows the Galaxy software interface with the 'Variant\_Detection\_RISS' tool selected. The left sidebar lists various tools under categories like NGS: SAM Tools, NGS: GATK Tools, ALIGNMENT UTILITIES, REALIGNMENT, BASE RECALIBRATION, GENOTYPING, and ANNOTATION. The main panel displays the configuration for the 'Variant\_Detection\_RISS' tool. It includes fields for 'Binding Type' (set to 'dbSNP', labeled f), 'ROD file' (containing '6. dbsnp\_137.hg19.vcf', labeled g), and an 'Execute' button (highlighted with a red circle, labeled h). A warning message at the bottom states: "⚠ This calculation is critically dependent on being able to skip over known variant sites. Please provide a dbSNP ROD or a VCF file containing known sites of genetic variation." The right panel shows the 'History' of previous runs, listing steps such as 'MarkDups\_Dupes\_Marked.html', 'MarkDups\_Dupes\_Marked.bam', 'Indel Realigner on data\_18\_and\_data\_16.log', 'Indel Realigner on data\_18\_and\_data\_16.BAM', 'Realigner Target Creator on data\_6, data\_16, and others.log', 'Realigner Target Creator on data\_6, data\_16, and others.GATK intervals', 'MarkDups\_Dupes\_Marked.html', 'MarkDups\_Dupes\_Marked.bam', 'Paired Read Mate Fixer on data.bam', and 'bam with fixed mates'. The top right corner indicates 'Using 1.4 GB'.

## 5.8 Analyze Covariates (before base recalibration)

- Load *analyze covariates* tool from the tool pane: "NGS: GATK Tools -> Analyze Covariates - draw plots"
- Covariates table recalibration file: -> "Count covariates...."
- Click "Execute"

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". On the left, a "Tools" sidebar lists various NGS tools under categories like "NGS: SAM Tools", "NGS: GATK Tools", and "REALIGNMENT". A red box labeled "a" highlights the "NGS: GATK Tools" category. A red arrow points from this box to the "Analyze Covariates – draw plots" tool, which is also highlighted with a red circle. The main content area displays a green success message: "The following job has been successfully added to the queue: 24: Count Covariates on data 6 and data 22 (Covariate File)". Below this, another message says: "25: Count Covariates on data 6 and data 22 (log)". A note states: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered." To the right, the "History" pane shows a list of completed jobs, including "Variant\_Detection\_RISS", "24: Count Covariates on data 6 and data 22 (Covariate File)", "25: Count Covariates on data 6 and data 22 (log)", and others. The total memory usage is listed as "Using 1.4 GB".

This screenshot shows the "Analyze Covariates (version 0.0.5)" dialog box. The "Tools" sidebar on the left is identical to the previous screenshot. The main dialog has a yellow header. A red box labeled "b" highlights the "Covariates table recalibration file:" dropdown, which contains "24: Count Covariates on data 6 and data 22 (Covariate File)". Below this, a red box labeled "c" highlights the "Execute" button. The "Basic or Advanced options:" section includes a dropdown menu set to "Basic". To the right, the "History" pane shows the same list of completed jobs as the previous screenshot, with "Using 1.4 GB" memory usage.

## 5.9 Review Covariate plots (before base quality recalibration)

Before recalibrating the base quality scores, it is prudent to first examine if reported and empirical scores agree. If they do, there is no need to recalibrate.

- In the history pane click the eye icon next to the name of the *analyze covariates* HTML output file to display the file in the center pane
- Click on the pdf output file "NA\_10858\_400.QualityScoreCovariate.dat.quality\_emp\_v\_stated.pdf" showing the difference between reported and empirical base calls as a function of the different covariate
- Inspect plot to determine if recalibration is necessary

The figure consists of three vertically stacked screenshots of the Galaxy web interface, showing the workflow and analysis results.

**Screenshot 1 (Top): History Pane**

- The title bar says "Galaxy / UMN".
- The "Tools" sidebar includes sections for NGS: SAM Tools, NGS: GATK Tools, ALIGNMENT UTILITIES, REALIGNMENT, and BASE RECALIBRATION.
- The main area shows a success message: "The following job has been successfully added to the queue: 26: Analyze Covariates on data 24 (HTML)".
- The "History" pane lists several jobs:
  - Variant\_Detection\_RISS (726.5 MB)
  - 27: Analyze Covariates on data 24 (log) (circled with red 'a')
  - 26: Analyze Covariates on data 24 (HTML) (circled with red 'a')
  - 25: Count Covariates on data 6 and data 22 (log)

**Screenshot 2 (Middle): Galaxy - GATK Output**

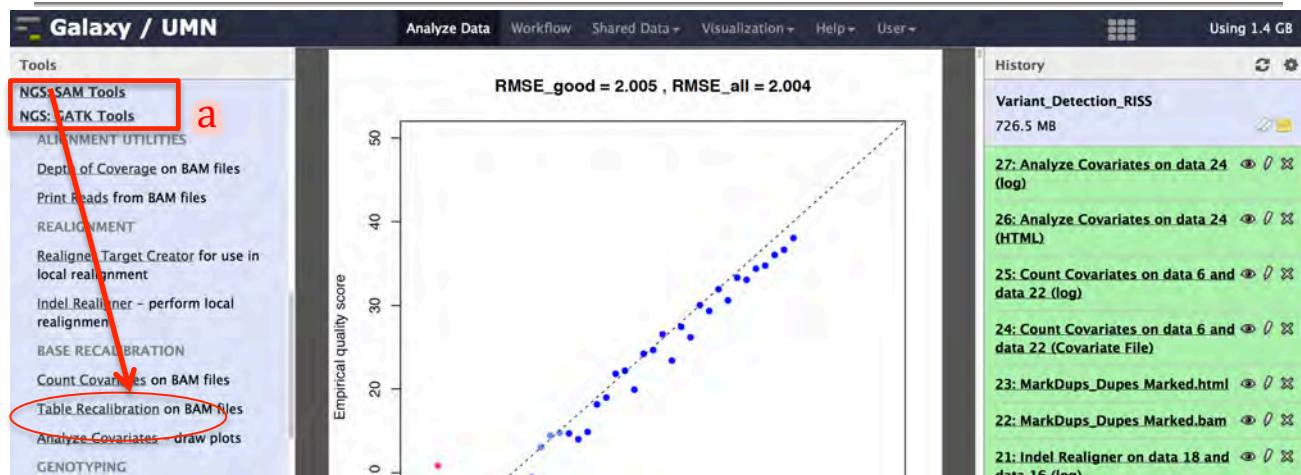
- The title bar says "Galaxy / UMN".
- The "Tools" sidebar includes sections for NGS: SAM Tools, NGS: GATK Tools, ALIGNMENT UTILITIES, REALIGNMENT, and BASE RECALIBRATION.
- The main area shows a list of generated files:
  - NA\_10858\_400.CycleCovariate.dat
  - NA\_10858\_400.CycleCovariate.dat.Cycle\_hist.pdf
  - NA\_10858\_400.CycleCovariate.dat.qual\_diff\_v\_Cycle.pdf
  - NA\_10858\_400.CycleCovariate.dat.reported\_qual\_v\_Cycle.pdf
  - NA\_10858\_400.DinucCovariate.dat
  - NA\_10858\_400.DinucCovariate.dat.Dinuc\_hist.pdf
  - NA\_10858\_400.DinucCovariate.dat.qual\_diff\_v\_Dinuc.pdf
  - NA\_10858\_400.DinucCovariate.dat.reported\_qual\_v\_Dinuc.pdf
  - NA\_10858\_400.QualityScoreCovariate.dat
  - NA\_10858\_400.QualityScoreCovariate.dat.quality\_emp\_hist.pdf (circled with red 'b')
  - NA\_10858\_400.QualityScoreCovariate.dat.quality\_emp\_v\_stated.pdf (circled with red 'b')
  - NA\_10858\_400.QualityScoreCovariate.dat.quality\_rep\_hist.pdf
- The "History" pane lists the same jobs as Screenshot 1.

**Screenshot 3 (Bottom): Covariate Plot**

- The title bar says "Galaxy / UMN".
- The "Tools" sidebar includes sections for NGS: SAM Tools, NGS: GATK Tools, ALIGNMENT UTILITIES, REALIGNMENT, and BASE RECALIBRATION.
- The main area displays a scatter plot titled "RMSE\_good = 2.005 , RMSE\_all = 2.004". The x-axis is "Reported quality score" and the y-axis is "Empirical quality score", both ranging from 0 to 50. A dashed diagonal line represents the identity line. Blue dots represent data points, and a red arrow points to one specific point at approximately (25, 25). A red 'c' is placed near the bottom right of the plot area.
- The "History" pane lists the same jobs as Screenshot 1.

## 5.10 Recalibrate base quality scores

- Load *Table Recalibration* tool from the tool pane: "NGS: GATK Tools -> Table Recalibration on BAM files"
- Covariates table recalibration file: -> "...Count covariates...."
- BAM file: -> "...MarkDups\_Dupes Marked.bam" (be sure to select the file generated after indel realignment)
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Click "Execute"



The screenshot shows the Galaxy interface with the 'Table Recalibration' tool configuration. The configuration steps are highlighted with red circles:

- b**: Covariates table recalibration file: 24: Count Covariates on data 6 and data 22 (Covariate File)
- c**: BAM file: 22: MarkDups\_Dupes Marked.bam
- d**: Using reference genome: Homo sapiens hg19\_canonical (GATK)
- e**: Execute button

The history panel on the right shows the workflow steps taken.

## 5.11 Count Covariates (after base recalibration)

- Load *count covariates* tool from the tool pane: "NGS: GATK Tools -> Count Covariates on BAM files"
- BAM file: -> "...Table Recalibrated....."
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Covariates to be used in the recalibration: -> check boxes next to
  - ReadGroupCovariate
  - QualityScoreCovariate
  - CycleCovariate
  - DinucCovariate
- Click on the "Add new Binding for reference-ordered data" button

The screenshot shows the Galaxy interface with the following details:

- Tools Panel:** Shows the "NGS: GATK Tools" section, with the "Count Covariates on BAM files" tool highlighted by a red box labeled 'a'.
- Job Queue:** A green success message box indicates that the job has been added to the queue: "28: Table Recalibration on data 24 and data 22 (BAM)".
- History Panel:** Displays a list of completed jobs, including "Variant\_Detection\_RISS", "29: Table Recalibration on data 24 and data 22 (log)", "28: Table Recalibration on data 24 and data 22 (BAM)", "27: Analyze Covariates on data 24 (log)", "26: Analyze Covariates on data 24 (HTML)", and "25: Count Covariates on data 6 and data 22 (log)".

The screenshot shows the configuration of the "Count Covariates on BAM files" tool in Galaxy:

- Tool Configuration:**
  - BAM file:** Set to "28: Table Recalibration on data 24 and data 22 (BAM)" (circled in red, labeled 'b').
  - Using reference genome:** Set to "Homo sapiens hg19\_canonical (GATK)" (circled in red, labeled 'c').
  - Covariates to be used in the recalibration:** Checkboxes are selected for "ReadGroupCovariate", "QualityScoreCovariate", "CycleCovariate", and "DinucCovariate" (circled in red, labeled 'd').
  - Binding for reference-ordered data:** A button labeled "Add new Binding for reference-ordered data" is highlighted with a red circle (labeled 'e').
- History Panel:** Shows a list of completed jobs, identical to the one in the previous screenshot.

- f) Binding Type: -> dbSNP
- g) ROD file: -> dbsnp\_137.hg19.vcf
- h) Click "Execute"

The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy / UMN, Analyze Data, Workflow, Shared Data, Visualization, Help, User, Using 1.4 GB.
- Left Sidebar (Tools):**
  - NGS: SAM Tools
  - NGS: GATK Tools
    - ALIGNMENT UTILITIES
    - Depth of Coverage on BAM files
    - Print Reads from BAM files
  - REALIGNMENT
  - Realigner Target Creator for use in local realignment
  - IndelRealigner - perform local realignment
  - BASE RECALIBRATION
  - Count Covariates on BAM files
  - Table Recalibration on BAM files
  - Analyze Covariates - draw plots
  - GENOTYPING
- Middle Panel (Variant\_Detection\_RISS):**
  - Binding for reference-ordered data**: -knownSites, --knownSites <knownSites>
  - Binding for reference-ordered data 1**
  - Binding Type:** dbSNP (highlighted with a red circle, labeled f)
  - ROD file:** 6: dbsnp\_137.hg19.vcf (highlighted with a red circle, labeled g)
  - Remove Binding for reference-ordered data 1**
  - Add new Binding for reference-ordered data**
  - Basic or Advanced GATK options:** Basic
  - Basic or Advanced Analysis options:** Basic
  - Execute** (highlighted with a red circle, labeled h)
- Right Panel (History):**
  - Variant\_Detection\_RISS (726.5 MB)
  - 29: Table Recalibration on data 24 and data 22 (log)
  - 28: Table Recalibration on data 24 and data 22 (BAM)
  - 27: Analyze Covariates on data 24 (log)
  - 26: Analyze Covariates on data 24 (HTML)
  - 25: Count Covariates on data 6 and data 22 (log)
  - 24: Count Covariates on data 6 and data 22 (Covariate File)

## 5.12 Analyze Covariates (after base recalibration)

- Load *analyze covariates* tool from the tool pane: "NGS: GATK Tools -> Analyze Covariates - draw plots"
- Covariates table recalibration file: -> "Count covariates...." (be sure to select the file generated after base quality recalibration)
- Click "Execute"

The following job has been successfully added to the queue:

30: Count Covariates on data 6 and data 28 (Covariate File)  
 31: Count Covariates on data 6 and data 28 (log)

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

**History**

- Variant\_Detection\_RISS  
780.9 MB
- 31: Count Covariates on data 6 and data 28 (log)
- 30: Count Covariates on data 6 and data 28 (Covariate File)
- 29: Table Recalibration on data 24 and data 22 (log)
- 28: Table Recalibration on data 24 and data 22 (BAM)
- 27: Analyze Covariates on data 24 (log)
- 26: Analyze Covariates on data 24 (HTML)

**Analyze Covariates (version 0.0.5)**

**Covariates table recalibration file:**  
30: Count Covariates on data 6 and data 28 (Covariate File) **b**

**Basic or Advanced options:**  
Basic **c**

**What it does**  
Create collapsed versions of the recal csv file and call R scripts to plot residual error versus the various covariates.  
For more information on base quality score recalibration using the GATK, see this tool specific page.  
To learn about best practices for variant detection using GATK, see this overview.  
If you encounter errors, please view the GATK FAQ.

**History**

- Variant\_Detection\_RISS  
780.9 MB
- 31: Count Covariates on data 6 and data 28 (log)
- 30: Count Covariates on data 6 and data 28 (Covariate File)
- 29: Table Recalibration on data 24 and data 22 (log)
- 28: Table Recalibration on data 24 and data 22 (BAM)
- 27: Analyze Covariates on data 24 (log)
- 26: Analyze Covariates on data 24 (HTML)

## 5.13 Review Covariate plots (after base quality recalibration)

- In the history pane click the eye icon next to the name of the *analyze covariates* HTML output file to display the file in the center pane
- Click on the pdf output file "NA\_10858\_400.QualityScoreCovariate.dat.quality\_emp\_v\_stated.pdf" showing the difference between reported and empirical base calls as a function of the different covariate

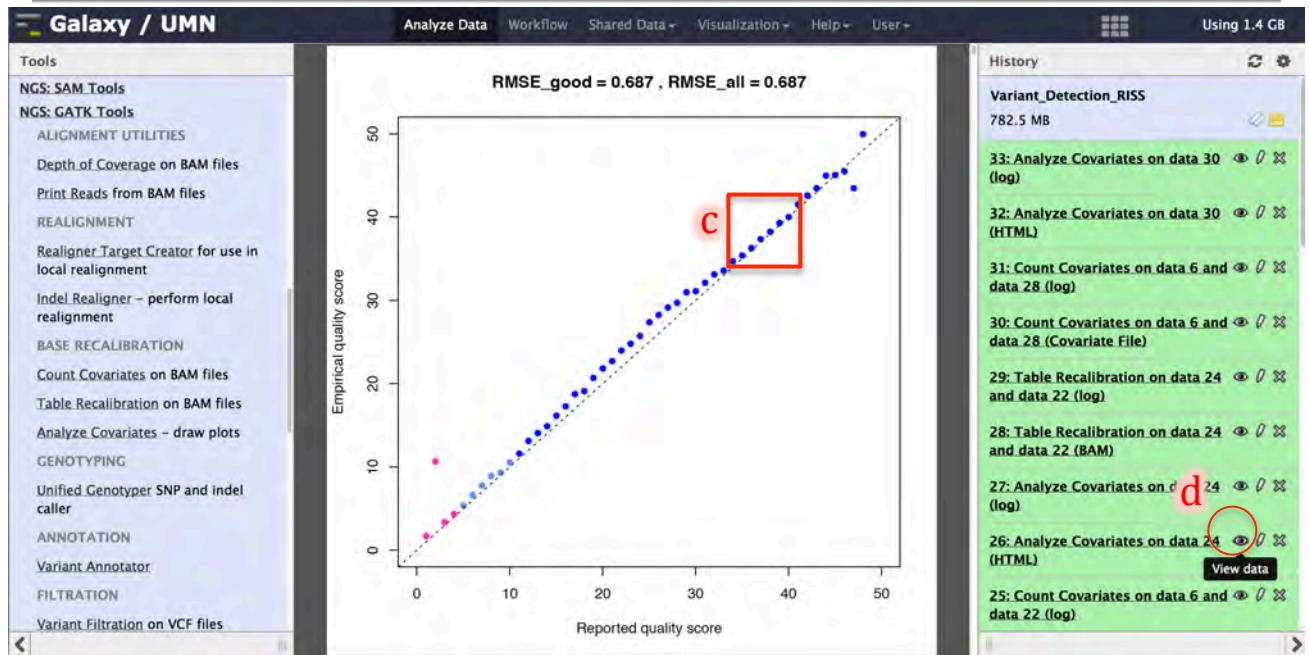
The screenshot shows the Galaxy interface with the following details:

- Tools Panel:** Shows categories like NGS: SAM Tools, NGS: GATK Tools, ALIGNMENT UTILITIES, Depth of Coverage on BAM files, Print Reads from BAM files, REALIGNMENT, Realigner Target Creator for use in local realignment, and Indel Realigner – perform local.
- Job Queue:** A green box indicates a successful job addition: "The following job has been successfully added to the queue: 32: Analyze Covariates on data 30 (HTML)".
- History Panel:** Shows the following jobs:
  - Variant\_Detection\_RISS (782.5 MB)
  - 33: Analyze Covariates on data 30 (log) (highlighted with a red circle 'a')
  - 32: Analyze Covariates on data 30 (HTML) (highlighted with a red circle 'a')
  - 31: Count Covariates on data 6 and data 28 (log)

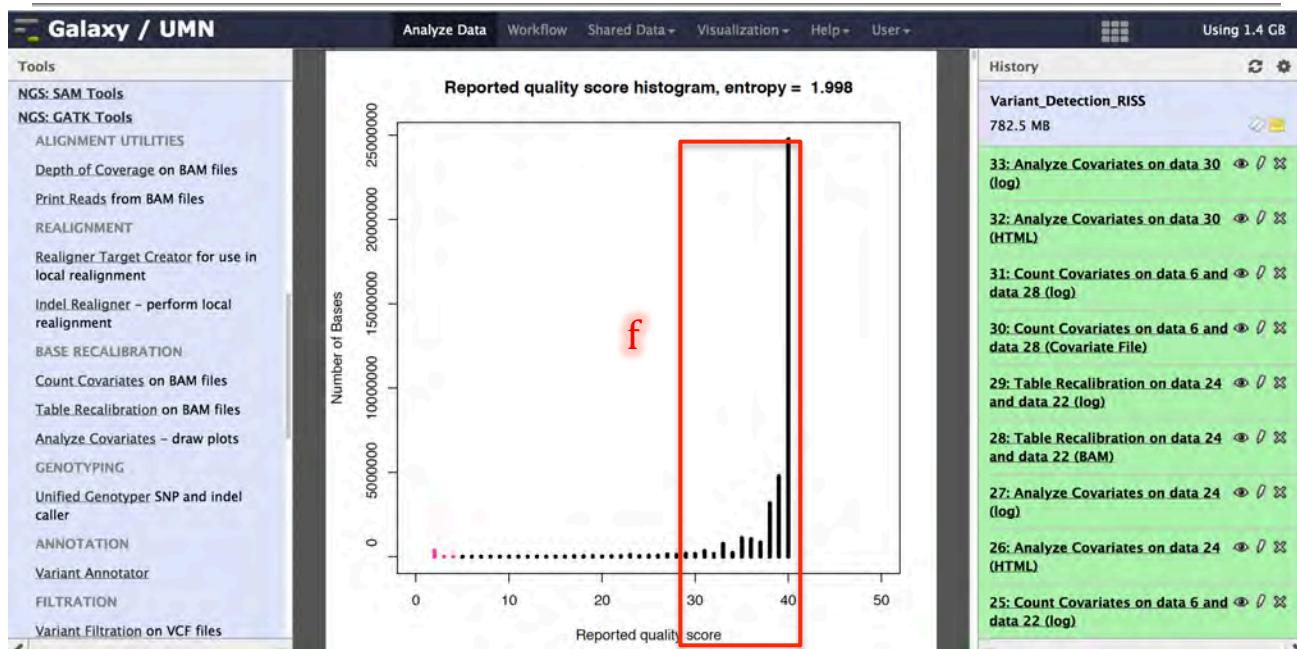
The screenshot shows the Galaxy interface with the following details:

- Tools Panel:** Shows categories like NGS: SAM Tools, NGS: GATK Tools, ALIGNMENT UTILITIES, Depth of Coverage on BAM files, Print Reads from BAM files, REALIGNMENT, Realigner Target Creator for use in local realignment, Indel Realigner – perform local realignment, and BASE RECALIBRATION.
- Output List:** A list titled "Galaxy - GATK Output" contains the following items:
  - NA\_10858\_400.CycleCovariate.dat
  - NA\_10858\_400.CycleCovariate.dat.Cycle\_hist.pdf
  - NA\_10858\_400.CycleCovariate.dat.qual\_diff\_v\_Cycle.pdf
  - NA\_10858\_400.CycleCovariate.dat.reported\_qual\_v\_Cycle.pdf
  - NA\_10858\_400.DinucCovariate.dat
  - NA\_10858\_400.DinucCovariate.dat.Dinuc\_hist.pdf
  - NA\_10858\_400.DinucCovariate.dat.qual\_diff\_v\_Dinuc.pdf
  - NA\_10858\_400.DinucCovariate.dat.reported\_qual\_v\_Dinuc.pdf
  - NA\_10858\_400.QualityScoreCovariate.dat
  - NA\_10858\_400.QualityScoreCovariate.dat.quality\_emp\_hist.pdf (highlighted with a red box 'b')
  - NA\_10858\_400.QualityScoreCovariate.dat.quality\_emp\_v\_stated.pdf (highlighted with a red box 'b')
  - NA\_10858\_400.QualityScoreCovariate.dat.quality\_rep\_hist.pdf
- History Panel:** Shows the following jobs:
  - Variant\_Detection\_RISS (782.5 MB)
  - 33: Analyze Covariates on data 30 (log)
  - 32: Analyze Covariates on data 30 (HTML) (highlighted with a red circle 'a')
  - 31: Count Covariates on data 6 and data 28 (log)
  - 30: Count Covariates on data 6 and data 28 (Covariate File)

- c) Examine plot to evaluate recalibration. Observe most improvement is in the Q30 - Q40 range
- d) To determine why most improvement is in the Q30 - Q40 range, examine input data by clicking the eye icon next to the name of the *analyze covariates* HTML output file generated before base recalibration "Analyze...."
- e) Click on the histogram showing distribution of reported quality score before recalibration "NA\_10858\_400.QualityScoreCovariate.dat.quality\_rep\_hist.pdf"



- f) Observe input data consisted of scores in the Q30 to Q40 range. Recalibration only as good as training data



## 6 GATK Phase 2: Variant Discovery

### ★ GATK Phase 2 details

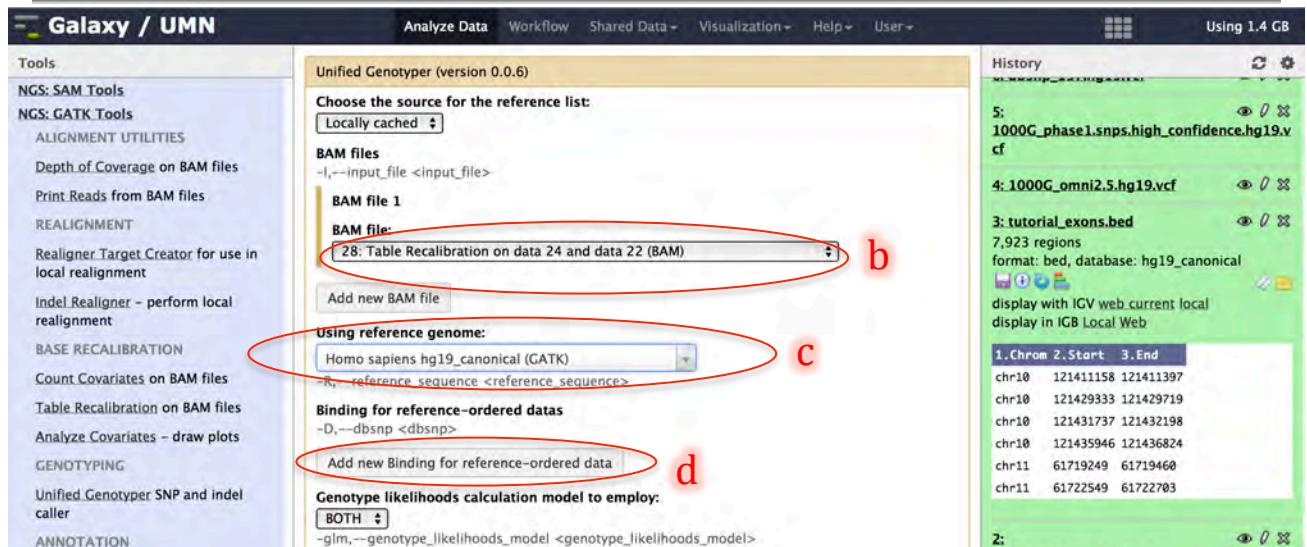
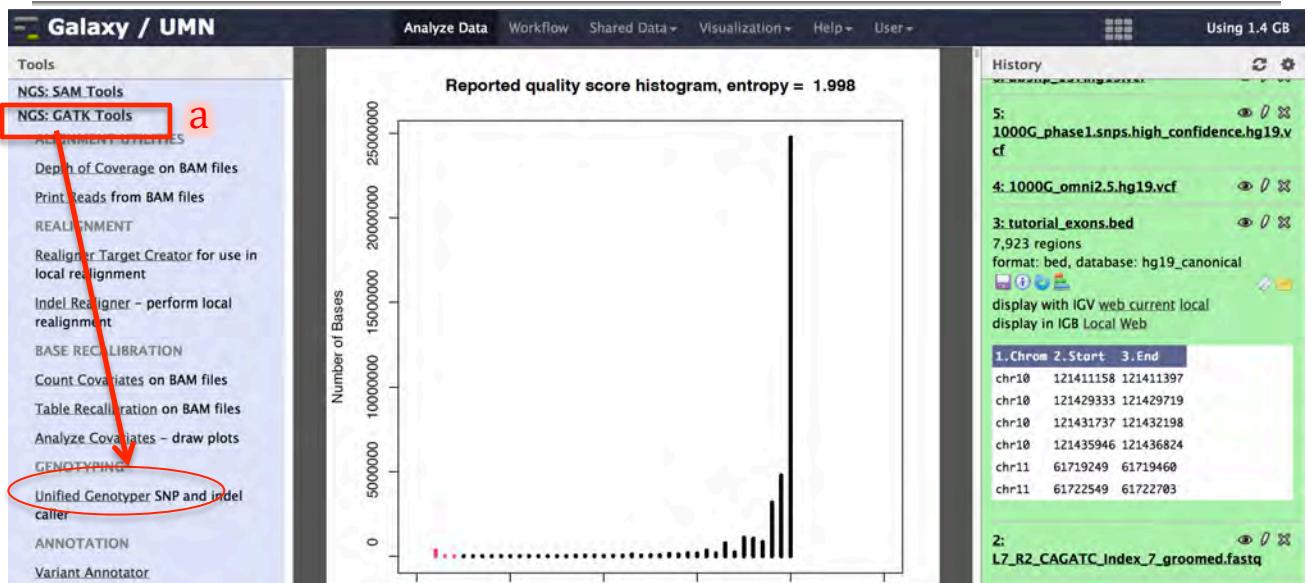
The GATK's Unified Genotyper employs a Bayesian model to compute the likelihood for each of the 10 possible bi-allelic diploid genotypes (AA, AC, AG, AT, CC, CG, CT, GG, GT, TT), as described in the equation below. The likelihood is computed across the entire pileup of bases at a position, taking into consideration the associated quality scores. Only "good bases" are considered – typically those satisfying a minimum base quality, read mapping quality, and pair mapping quality. In the formula below,  $L(G|D)$  is computed over all 10 possible genotypes. See [http://www.broadinstitute.org/gatk/gatkdocs/org\\_broadinstitute\\_sting\\_gatk\\_walkers\\_genotyper\\_UnifiedGenotyper.html](http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_genotyper_UnifiedGenotyper.html) for more information.

#### • Parameter selection and considerations

1. It is useful to provide a reference dbSNP VCF file to the Unified Genotyper, as this will automatically transfer rsIDs from dbSNP onto known variants detected in your samples.
2. Genotype likelihood calculations can be performed for SNP, INDEL or BOTH.
3. Two separate variant call phred-scale quality values are reported: a threshold for *high-quality calling* variants and a potentially less stringent threshold for emitting/outputting lower-confidence borderline calls.
4. The final GATK phase discussed in the next section needs information about each of the variant calls in order to rank the confidence of each one (e.g., FisherStrand to assess strand bias). You will save time and effort by allowing the Unified Genotyper to track these ahead of time under "Annotation Types", rather than do this later.

## 6.1 Variant detection using Unified Genotyper

- Navigate back to Galaxy and load *Unified Genotyper* tool from the tool pane: "NGS: GATK Tools -> Unified Genotyper SNP and indel caller"
- BAM file: -> "...Table Recalibrated.....(BAM)"
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Click the "Add new Binding for reference-ordered data" button



- e) Binding Type: -> dbSNP
- f) ROD file: -> dbsnp\_137.hg19.vcf
- g) Genotype likelihoods calculation model to employ: -> BOTH
- h) The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be called: -> 20
- i) The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be emitted (and filtered if less than the calling threshold): -> 20
- j) On the drop down menu below “**Basic or Advanced GATK options:**” select “Advanced”
- k) Click the “Add new Operate on Genomic intervals” button

- l) Under “**Genomic intervals:**” select the file “tutorial\_exons.bed”  
 m) Basic or Advanced Analysis options: -> Advanced  
 n) Annotation Types: -> check boxes next to  
     ✓ FisherStrand  
     ✓ HaplotypeScore  
     ✓ HomopolymerRun  
     ✓ MappingQualityRankSumTest  
     ✓ QualByDepth  
     ✓ ReadPosRankSumTest

**Galaxy / UMN**

Tools

NGS: SAM Tools

NGS: GATK Tools

ALIGNMENT UTILITIES

Depth of Coverage on BAM files

Print Reads from BAM files

REALIGNMENT

Realigner Target Creator for use in local realignment

Indel Realigner – perform local realignment

BASE RECALIBRATION

Count Covariates on BAM files

Analyze Data Workflow Shared Data Visualization Help User

Operate on Genomic intervals

Genomic intervals: 3: tutorial\_exons.bed

Add new Operate on Genomic intervals

History

5: 1000G\_phase1.snps.high\_confidence.hg19.vcf

4: 1000G\_omni2.5.hg19.vcf

3: tutorial\_exons.bed

7,923 regions

format: bed, database: hg19\_canonical

display with IGV web current local

display in IGB Local Web

1.Chrom 2.Start 3.End

chr10 121411158 121411397

**Galaxy / UMN**

Tools

NGS: SAM Tools

NGS: GATK Tools

ALIGNMENT UTILITIES

Depth of Coverage on BAM files

Print Reads from BAM files

REALIGNMENT

Realigner Target Creator for use in local realignment

Analyze Data Workflow Shared Data Visualization Help User

--disable\_experimental\_low\_memory\_sharding

Makes the GATK behave non deterministically, that is, the random numbers generated will be different in every run:

-ndrs,--nonDeterministicRandomSeed

Basic or Advanced Analysis: Advanced

Non-reference probability calculation model to employ:

EXACT

-pnrm,--p\_nonref\_model <p\_nonref\_model>

History

5: 1000G\_phase1.snps.high\_confidence.hg19.vcf

4: 1000G\_omni2.5.hg19.vcf

3: tutorial\_exons.bed

7,923 regions

format: bed, database: hg19\_canonical

display with IGV web current local

display in IGB Local Web

**Galaxy / UMN**

Tools

NGS: SAM Tools

NGS: GATK Tools

ALIGNMENT UTILITIES

Depth of Coverage on BAM files

Print Reads from BAM files

REALIGNMENT

Realigner Target Creator for use in local realignment

Indel Realigner – perform local realignment

BASE RECALIBRATION

Count Covariates on BAM files

Table Recalibration on BAM files

Analyze Covariates – draw plots

GENOTYPING

Unified Genotyper SNP and indel caller

ANNOTATION

Variant Annotator

FILTRATION

Variant Filtration on VCF files

Analyze Data Workflow Shared Data Visualization Help User

Annotation Types:

Select All Unselect All

AlleleBalance  
 AlleleBalanceBySample  
 BaseCounts  
 BaseQualityRankSumTest  
 ChromosomeCounts  
 DepthOfCoverage  
 DepthPerAlleleBySample  
 FisherStrand  
 GQContent  
 HaplotypeScore  
 HardyWeinberg  
 HomopolymerRun  
 InbreedingCoeff  
 IndelType  
 LowMQ  
 MVLikelihoodRatio  
 MappingQualityRankSumTest  
 MappingQualityZero  
 MappingQualityZeroBySample  
 MappingQualityZeroFraction  
 NBaseCount  
 QualByDepth  
 RMSMappingQuality  
 ReadDepthAndAllelicFractionBySample  
 ReadPosRankSumTest  
 SampleList

History

5: 1000G\_phase1.snps.high\_confidence.hg19.vcf

4: 1000G\_omni2.5.hg19.vcf

3: tutorial\_exons.bed

7,923 regions

format: bed, database: hg19\_canonical

display with IGV web current local

display in IGB Local Web

1.Chrom 2.Start 3.End

chr10 121411158 121411397

chr10 121429333 121429719

chr10 121431737 121432198

chr10 121435946 121436824

chr11 61719249 61719460

chr11 61722549 61722703

2: L7\_R2\_CAGATC\_Index\_7\_groomed.fastq

1: L7\_R1\_CAGATC\_Index\_7\_groomed.fastq

o) Annotation Interfaces/Groups: -> check box next to

✓ Standard

p) Click "Execute"

The screenshot shows the Galaxy / UMN interface. In the center, under 'Annotation Interfaces/Groups', there is a list of options. The 'Standard' checkbox is checked and highlighted with a red circle. Other options listed include 'RodRequiringAnnotation', 'Experimental', 'WorkInProgress', 'RankSumTest', and several grouped annotations like 'AlleleBalance', 'AlleleBalanceBySample', etc.

The screenshot shows the Galaxy / UMN interface. In the center, there is a section titled 'Allow the discovery of multiple alleles (SNPs only)'. Below it is a checkbox for 'multiallelic'. At the bottom of this section is a large blue 'Execute' button, which is highlighted with a red circle. To the right of the execute button, there is some text starting with 'p'.

## 6.2 Review Unified Genotyper results (Raw Variants)

- In the history pane click the eye icon next to the name of the *Genotyper* output file to display the file in the center pane
- Click the arrow at the bottom of the *tools pane* to the left of the browser minimize it

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". The right side shows a "History" panel with a list of jobs:

- Variant\_Detection\_RISS (782.6 MB)
- 36: Unified Genotyper on data 28, data 6, and data 3 (log) (a)
- 35: Unified Genotyper on data 28, data 6, and data 3 (metrics)
- 34: Unified Genotyper on data 28, data 6, and data 3 (VCF)
- 33: Analyze Covariates on data 30 (log)
- 32: Analyze Covariates on data 30 (HTML)

The central area displays a message about a successfully added job:

The following job has been successfully added to the queue:  
 34: Unified Genotyper on data 28, data 6, and data 3 (VCF)  
 35: Unified Genotyper on data 28, data 6, and data 3 (metrics)  
 36: Unified Genotyper on data 28, data 6, and data 3 (log)

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

The left sidebar lists various tools under the "Tools" heading, including NGS: SAM Tools, NGS: GATK Tools, ALIGNMENT UTILITIES, Depth of Coverage on BAM files, Print Reads from BAM files, REALIGNMENT, Realigner Target Creator for use in local realignment, Indel Realigner - perform local realignment, BASE RECALIBRATION, Count Covariates on BAM files, and Table Recalibration on BAM files.

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". The right side shows a "History" panel with a list of jobs:

- Variant\_Detection\_RISS (782.6 MB)
- 36: Unified Genotyper on data 28, data 6, and data 3 (log)
- 35: Unified Genotyper on data 28, data 6, and data 3 (metrics)
- 34: Unified Genotyper on data 28, data 6, and data 3 (VCF)
- 33: Analyze Covariates on data 30 (log)
- 32: Analyze Covariates on data 30 (HTML)
- 31: Count Covariates on data 6 and data 28 (log)
- 30: Count Covariates on data 6 and data 28 (Covariate File)
- 29: Table Recalibration on data 24 (log)
- 28: Table Recalibration on data 24 (HTML)

The central area displays a large amount of command-line code for the Unified Genotyper tool:

```

Chrom Pos
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the sample">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=2 or greater quality score)">
##FORMAT=<ID=CQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes (0 = homozygous reference, 1 = heterozygous, 2 = homozygous alternate)">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the sample">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the sample">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test for each base quality level">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DP,Number=0,Type=Integer,Description="Approximate read depth; some reads may have zero depth due to low quality scores or other filtering">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to determine if the observed number of insertions and deletions is significantly different than expected under the null hypothesis of random distribution of indels">
##INFO=<ID=HR,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of V">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most one haplotype">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient estimated from the sample">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test for each mapping quality level">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test for each read position">
##UnifiedGenotyper="analysis_type=UnifiedGenotyper input_file=[/galaxy/PRODUCTION/database/tmp/chr1.vcf]
##contig=<ID=chr1,length=249250621,assembly=hg19>
##contig=<ID=chr10,length=135534747,assembly=hg19>

```

The left sidebar lists various tools under the "Tools" heading, including NGS: SAM Tools, NGS: GATK Tools, ALIGNMENT UTILITIES, Depth of Coverage on BAM files, Print Reads from BAM files, REALIGNMENT, Realigner Target Creator for use in local realignment, Indel Realigner - perform local realignment, BASE RECALIBRATION, Count Covariates on BAM files, Analyze Covariates - draw plots, GENOTYPING, Unified Genotyper SNP and indel caller, ANNOTATION, Variant Annotator, and Variant Filtration on VCF files. A red circle labeled 'b' is drawn around the "Variant Filtration on VCF files" tool.

- c) Click the arrow at the bottom of the *history pane* to the right of the browser minimize it
- d) Scroll to the left and right of the *center pane* to view variants. NOTE: Browser not the ideal application for viewing results
- e) Click the arrow at the bottom-left corner of your browser to bring the *tools pane* back to view

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". The main area displays a VCF file content:

```

Chrom          Pos
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HRun,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of Variant Allele In Either Direction">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample w">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##UnifiedGenotyper="analysis_type=UnifiedGenotyper input_file=[/galaxy/PRODUCTION/database/tmp/gatk-nYq7lo/gatk_input_0.bam] read
##contig=<ID=chr1,length=249250621,assembly=hg19>
> ##contig=<ID=chr10,length=135534747,assembly=hg19>

```

The right side of the screen shows the "History" pane with a list of workflow steps. A red circle highlights the right scroll bar of the history pane.

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". The main area displays a VCF file content:

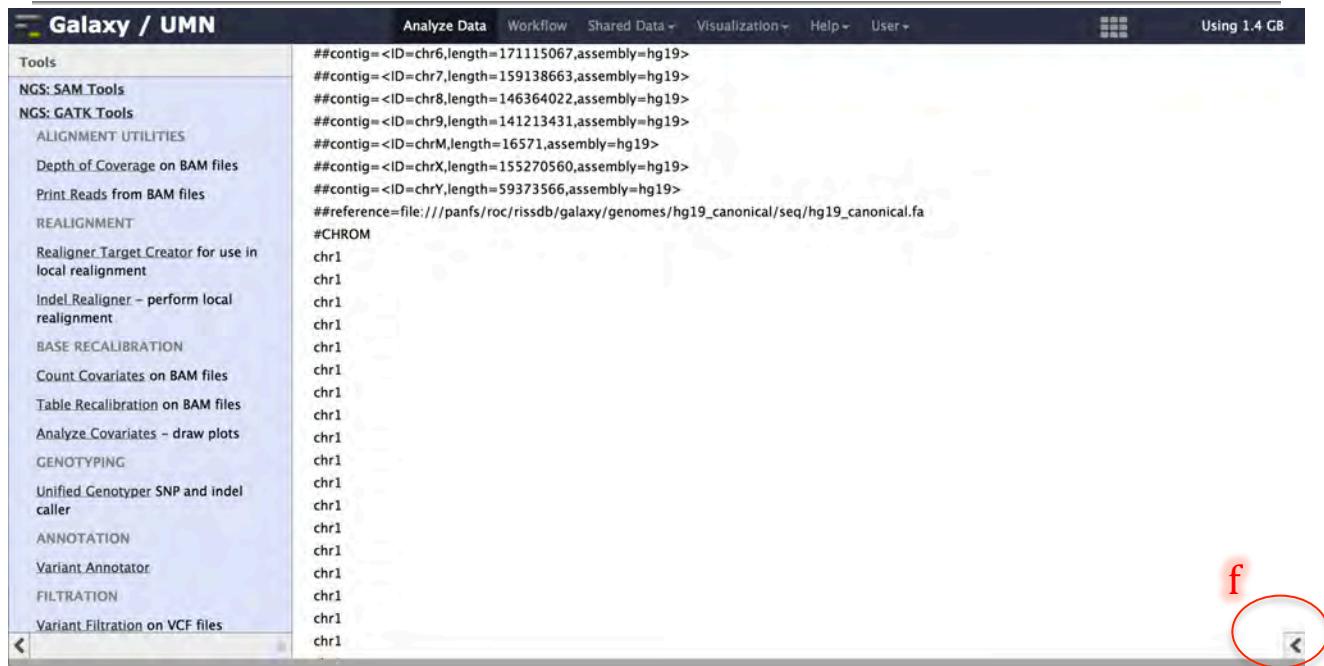
```

##contig=<ID=chr6,length=171115067,assembly=hg19>
##contig=<ID=chr7,length=159138663,assembly=hg19>
##contig=<ID=chr8,length=146364022,assembly=hg19>
##contig=<ID=chr9,length=141213431,assembly=hg19>
##contig=<ID=chrM,length=16571,assembly=hg19>
##contig=<ID=chrX,length=155270560,assembly=hg19>
##contig=<ID=chrY,length=59373566,assembly=hg19>
##reference=file:///panfs/roc/rissdb/galaxy/genomes/hg19_canonical/seq/hg19_canonical.fa
#CHROM          POS    ID
chr1           35251075 rs200004121
chr1           55470811 rs41297877
chr1           55474262 rs33938617
chr1           55474325 rs6682884
chr1           103354115 rs17127203
chr1           103380379 rs112615091
chr1           103444679 rs11164649
chr1           103468336 -
chr1           103471456 rs112482103
chr1           103480117 rs55851925
chr1           103496620 rs7523441
chr1           103496805 rs10612145
chr1           116243868 rs28730711
chr1           116243877 rs7413162
chr1           116260532 rs2997741
chr1           116260544 rs3811001
chr1           116283343 rs9428083
chr1           197297540 rs12042179

```

A red double-headed arrow is positioned over the variant list. A red circle highlights the bottom scroll bar of the center pane.

- f) Click the arrow at the bottom-right corner of your browser to bring the *tools pane* back to view



The screenshot shows the Galaxy / UMN web application. At the top, there is a navigation bar with links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. On the right side of the header, it says "Using 1.4 GB". Below the header, there is a sidebar titled "Tools" containing a list of tool categories and their descriptions. The categories include NGS: SAM Tools, NGS: GATK Tools, ALIGNMENT UTILITIES, REALIGNMENT, BASE recalibration, GENOTYPING, ANNOTATION, and FILTRATION. Under each category, there are specific tool names like Realigner.Target\_Creator, Indel\_Realigner, Count Covariates on BAM files, etc. To the right of the sidebar, the main content area displays a large amount of text, which appears to be a command-line script or configuration file. In the bottom right corner of the main content area, there is a small red circle with a white arrow pointing towards the bottom-right corner of the browser window, indicating where to click to bring the tools pane back into view.

## 7 GATK Phase 3: Preliminary Analysis

### ★ GATK Phase 3 details

#### ★ Variant Recalibration

Any pipeline that takes raw sequencing reads, maps them to a reference and attempts to make genotype calls will have inherent systematic errors leading to false-positive variant calls. *The challenge is to separate true genotype calls from machine artifacts.* The GATK's Variant Quality Score Recalibrator attempts to separate raw variant calls into different confidence levels, or tranches, based on training from “truth” data – usually variants that have been verified. Training is done using a Gaussian Mixture model. A wide variety of external evidence can be used to help train the recalibrator:

- **Known dbSNP rates**

As a result of the 1000 Genomes project, it is estimated that 99% of all variants have been cataloged for Caucasian samples. The numbers are nearly as high for some African and Asian populations. Therefore, it stands to reason that SNPs at known sites are more likely to be real. Even more confidence may be associated with carefully validated sets like the HapMap project or the 1000 Genomes OMNI-chip validation set. If a sample has a very high rate of novel variants, yet comes from a well-sampled population, the quality of those calls is circumspect.

- **Transition (Ti)/Transversion (Tv) rates are non-random**

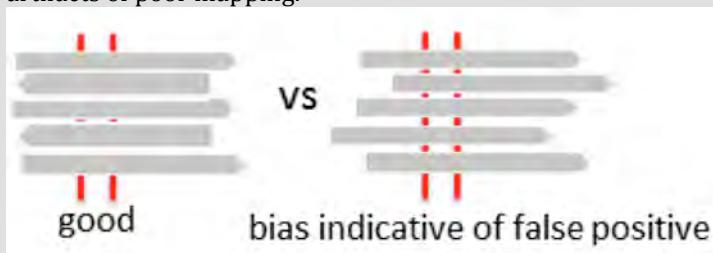
Transition ( $A \leftrightarrow G, C \leftrightarrow T$ ) and transversion ( $A \leftrightarrow C, A \leftrightarrow T, C \leftrightarrow G, G \leftrightarrow T$ ) ratios are not random (0.5). Selection pressure works against transversions in coding DNA and in other structurally or functionally-relevant regions. High throughput validation studies have established that Whole genome sequencing typically yields  $Ti/Tv$  rates  $\sim 2.0\text{-}2.1$ , and exome data around  $\sim 3.0\text{-}3.3$ . Lower rates for samples are highly indicative of poor-quality calls.

- **Population-specific heterozygosity should hold**

If you are running the GATK with multiple samples, you have the opportunity to infer heterozygosity rates. Significant deviations from Hardy-Weinberg equilibrium (e.g., observing all AT calls with no AA or TT calls at a locus) often indicate a systematic problem.

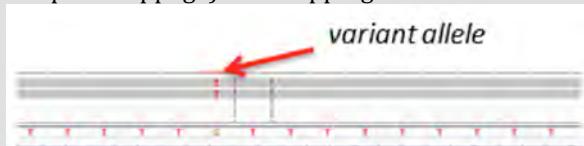
- **FisherStrand (FS)**

Variants identified in regions where nearly all reads are on one strands are more likely to be artifacts of poor mapping.



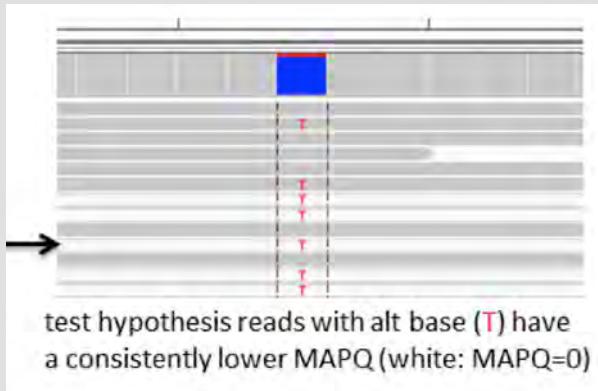
- **HomopolymerRun (HRun)**

Long tracts of single-nucleotide repeats are prone to error arising during library creation (e.g., template slippage) and mapping.



- **MappingQualityRankSumTest (MQRankSum)**

True heterozygous calls should have reference calls and alternate calls with comparable mapping quality. Suspicion is raised if the alternate calls preferentially appear in poorer-quality mapped reads.

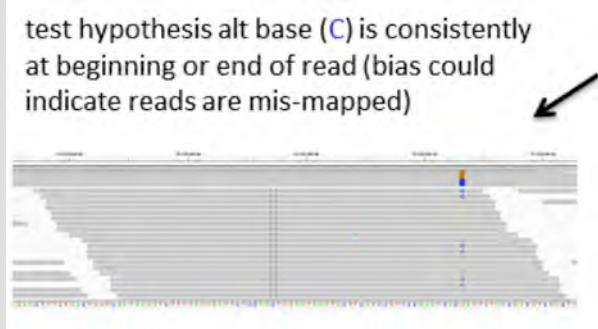


- **QualByDepth (QD)**

One would expect that the deeper coverage you get, the higher the confidence you should have in the variant call. But this isn't strictly correct. It is only true if the majority of the bases are of high quality and they fall in reads that are confidently mapped. Recall, the Unified Genotyper only makes genotype calls based on the "good bases" at a locus. In other words, the low quality bases or bases that occur in poorly mapped reads do not contribute to the raw genotype quality score. So, we should penalize pileups that have a high proportion of "bad bases". This is achieved here by taking the raw confidence assigned by the unified genotyper for a variant site *divided by* the *unfiltered* depth. This measure properly penalizes huge pileups with poorly mapped reads or basecalls (usually due to collapsed repeats).

- **ReadPosRankSum**

One would expect variant calls to be randomly distributed in position along a read, and not preferentially appear near the beginning or end of reads. Bias in the positioning of the alternate could be evidence of mismapping.



## ★ Variant Annotation

Once you obtain a list of variants in VCF format, it is highly desirable to know the potential effect of the variant on surrounding genes. For example, a variant might be upstream, downstream or intronic relative to specific genes, or it might be a synonymous coding SNP, non-synonymous coding or splice-site altering SNP, etc. Additionally, it may be desirable to predict whether an SNP is potentially deleterious based on the conservation level of the affected sequence, and protein 3D structure elements, etc. This information may be obtained using the widely-popular tool annovar (<http://www.openbioinformatics.org/annovar/>). Although this tool is free, licensing prohibits it from being wrapped and re-distributed in Galaxy. Galaxy includes an alternative tool called snpEff that is better integrated with Galaxy and the GATK, and has nearly the same level of functionality.

# Variant Recalibration

## 7.1 Select SNPs

- Load *Select variants* tool from the tool pane: "NGS: GATK Tools -> Select Variants from VCF files"
- Variant file to select: -> "...Unified Genotyper....(VCF)"
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Basic or Advanced Analysis options: -> Advanced

The screenshot shows the Galaxy interface with the 'Tools' pane open. The 'Select Variants from VCF files' tool is highlighted with a red circle and labeled 'a'. The main workspace shows a command-line interface with several '##contig' lines and 'chr1' entries. The right panel, 'History', lists various workflow steps, with the 'Variant\_Detection\_RISS' step expanded, showing its details.

The screenshot shows the Galaxy interface with the 'Select Variants' tool configuration dialog open. The 'Variant file to select' dropdown is set to '34: Unified Genotyper on data 28, data 6, and data 3 (VCF)' and is circled in red and labeled 'b'. The 'Using reference genome' dropdown is set to 'Homo sapiens hg19\_canonical (GATK)' and is circled in red and labeled 'c'. The right panel, 'History', shows the same workflow steps as the previous screenshot.

The screenshot shows the Galaxy interface with the 'Select Variants' tool configuration dialog open, specifically focusing on the 'Basic or Advanced Analysis options' dropdown, which is set to 'Advanced' and circled in red and labeled 'd'. The right panel, 'History', shows the same workflow steps as the previous screenshots.

- e) Select only a certain type of variants from the input file: -> check box next to ✓ SNP
- f) Click "Execute"
- g) Click the pencil icon next to the output file to edit attributes

Select a random subset of variants:  
Use all variants

Don't include loci found to be non-variant after the subsetting procedure:  
-env,--excludeNonVariants

Select only a certain type of variants from the input file:  
Select All Unselect All

INDEL  
 SNP  
 MIXED  
 MNP  
 SYMBOLIC  
 NO\_VARIATION

-selectType,--selectTypeToInclude <selectTypeToInclude>

**Execute**

History

- Variant\_Detection\_RISS 782.6 MB
  - 36: Unified Genotyper on data 28, data 6, and data 3 (log)
  - 35: Unified Genotyper on data 28, data 6, and data 3 (metrics)
  - 34: Unified Genotyper on data 28, data 6, and data 3 (VCF)
  - 33: Analyze Covariates on data 30 (log)
  - 32: Analyze Covariates on data 30 (HTML)
  - 31: Count Covariates on data 6 and data 28 (log)

The following job has been successfully added to the queue:

37: Select Variants on data 34 (Variant File)

38: Select Variants on data 34 (log)

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

- Variant\_Detection\_RISS 782.8 MB
  - 38: Select Variants on data 34 (log) **g** 0
  - 37: Select Variants on data 34 (Variant File) 0
  - 36: Unified Genotyper on data 28, data 6, and data 3 (log)

- h) Enter "SNPs" under **Name**:  
 i) Click "Save"

The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy / UMN, Analyze Data, Workflow, Shared Data, Visualization, Help, User.
- Left Sidebar (Tools):**
  - Count Covariates on BAM files
  - Table Recalibration on BAM files
  - Analyze Covariates – draw plots
  - GENOTYPING**
  - Unified Genotyper SNP and indel caller
  - ANNOTATION
  - Variant Annotator
  - FILTRATION
  - Variant Filtration on VCF files
  - Select Variants from VCF files
  - VARIANT QUALITY SCORE RECALIBRATION
  - Variant Recalibrator
  - Apply Variant Recalibration
  - VARIANT UTILITIES
  - Validate Variants
  - Eval Variants
  - Combine Variants
  - NGS: Variant Detection
- Main Content (Edit Attributes Dialog):**
  - Name:** SNPs (highlighted with a red circle, labeled 'h')
  - Info:** (empty text area)
  - Annotation / Notes:** (empty text area)
  - Database/Build:** Human hg19 in GATK canonical chr... (dropdown menu)
  - Number of comment lines:** 53 (checkbox checked)
  - Buttons:** Save (highlighted with a red circle, labeled 'i'), Auto-detect
  - Description:** This will inspect the dataset and attempt to correct the above column values if they are not accurate.
- Right Sidebar (History):**
  - Variant\_Detection\_RISS
  - 782.8 MB
  - 38: Select Variants on data 34 (log)
  - 37: Select Variants on data 34 (Variant File)
  - 36: Unified Genotyper on data 28, data 6, and data 3 (log)
  - 35: Unified Genotyper on data 28, data 6, and data 3 (metrics)
  - 34: Unified Genotyper on data 28, data 6, and data 3 (VCF)
  - 33: Analyze Covariates on data 30 (log)
  - 32: Analyze Covariates on data 30 (HTML)
  - 31: Count Covariates on data 6 and data 28 (log)
  - 30: Count Covariates on data 6 and data 28 (Covariate File)

## 7.2 Recalibrate SNPs

- Load *variant recalibration* tool from the tool pane: “NGS: GATK Tools -> Variant Recalibrator”
- Variant file to recalibrate: -> “SNPs”
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Click on “Add new Binding for reference-ordered data”

**Screenshot 1: Galaxy / UMN - Variant Recalibrator (Attributes)**

The 'Tools' pane on the left lists various variants-related tools. The 'Variant Recalibrator' tool is highlighted with a red circle (a).

The main panel shows the 'Edit Attributes' step. The 'Name:' field contains 'SNPs'. The 'Info:' and 'Annotation / Notes:' fields are empty.

**Screenshot 2: Galaxy / UMN - Variant Recalibrator (Tool Configuration)**

The 'Tools' pane on the left lists various variants-related tools. The 'Variant Recalibrator' tool is highlighted with a red circle (a).

The main panel shows the 'Variant Recalibrator (version 0.0.4)' tool configuration. Step (b) highlights the 'Variant file to recalibrate' dropdown, which is set to '37: SNPs'. Step (c) highlights the 'Using reference genome' dropdown, which is set to 'Homo sapiens hg19\_canonical (GATK)'. Step (d) highlights the 'Binding for reference-ordered data' button.

- e) Binding Type: -> HapMap
- f) ROD file: -> hapmap\_3.3.hg19.vcf
- g) Use as training/truth/known sites: -> Set training/truth/known sites
  - Is Training Site: -> ✓
  - Is Truth Site: -> ✓
- h) prior probability of being true: -> 15.0
- i) Click on “Add new Binding for reference-ordered data”
- j) Binding Type: -> OMNI
- k) ROD file: -> 1000G\_omni2.5.hg19.vcf
- l) Use as training/truth/known sites: -> Set training/truth/known sites
  - Is Training Site: -> ✓
  - Is Truth Site: -> ✓
- m) prior probability of being true: -> 12.0
- n) Click on “Add new Binding for reference-ordered data”

The screenshot shows the Galaxy interface with the 'Binding for reference-ordered data' tool selected. The 'Binding Type' dropdown is set to 'HapMap' (labeled e). The 'ROD file' input field contains '7: hapmap\_3.3.hg19.vcf' (labeled f). The 'Set training/truth/known sites' dropdown is open, showing 'Is Known Site' (unchecked), 'Is Training Site' (checked), and 'Is Truth Site' (checked) (labeled g). The 'prior probability of being true' input field is set to '15.0' (labeled h). The 'Annotations which should used for calculations' section has 'Select All' and 'Unselect All' buttons.

The screenshot shows the Galaxy interface with the 'Binding for reference-ordered data' tool selected. The 'Binding Type' dropdown is set to 'OMNI' (labeled j). The 'ROD file' input field contains '4: 1000G\_omni2.5.hg19.vcf' (labeled k). The 'Set training/truth/known sites' dropdown is open, showing 'Is Known Site' (unchecked), 'Is Training Site' (checked), and 'Is Truth Site' (checked) (labeled l). The 'prior probability of being true' input field is set to '12.0' (labeled m). The 'Annotations which should used for calculations' section has 'Select All' and 'Unselect All' buttons.

- o) Binding Type: -> 1000G
- p) ROD file: -> 1000G\_phase1.snps.high\_confidence.hg19.vcf
- q) Use as training/truth/known sites: -> Set training/truth/known sites  
Is Training Site: -> ✓
- r) prior probability of being true: -> 10.0
- s) Click on "Add new Binding for reference-ordered data"
- t) Binding Type: -> dbSNP
- u) ROD file: -> dbsnp\_137.hg19.vcf
- v) Use as training/truth/known sites: -> Set training/truth/known sites  
Is Known Site:-> ✓
- w) prior probability of being true: -> 2.0

**Galaxy / UMN**

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

Tools

- Variant Filtration on VCF files
- Select Variants from VCF files
- VARIANT QUALITY SCORE RECALIBRATION
- Variant Recalibrator
- Apply Variant Recalibration
- VARIANT UTILITIES
- Validate Variants
- Eval Variants
- Combine Variants
- NGS: Variant Detection
- NGS: Peak Calling
- NGS: Simulation
- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models
- SnpEff tools
- Phenotype Association
- VCF Tools
- IGVTools
- MSI

Binding for reference-ordered data 3

Binding Type: 0

ROD file: p 5: 1000G\_phase1.snps.high\_confidence.hg19.vcf

Use as training/truth/known sites: Set training/truth/known sites

Is Known Site: q

Is Training Site:

Is Truth Site:

Is Bad Site:

prior probability of being true: r 10.0

Remove Binding for reference-ordered data 3

Add new Binding for reference-ordered data S

annotations which should used for calculations: Select All Unselect All

History

Variant\_Detection\_RISS 782.8 MB

- 38: Select Variants on data 34 (log)
- 37: SNPs
- 36: Unified Genotyper on data 28, data 6, and data 3 (log)
- 35: Unified Genotyper on data 28, data 6, and data 3 (metrics)
- 34: Unified Genotyper on data 28, data 6, and data 3 (VCF)
- 33: Analyze Covariates on data 30 (log)
- 32: Analyze Covariates on data 30 (HTML)
- 31: Count Covariates on data 6 and data 28 (log)
- 30: Count Covariates on data 6 and data 28 (Covariate File)
- 29: Table Recalibration on data 24

**Galaxy / UMN**

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

Tools

- Variant Filtration on VCF files
- Select Variants from VCF files
- VARIANT QUALITY SCORE RECALIBRATION
- Variant Recalibrator
- Apply Variant Recalibration
- VARIANT UTILITIES
- Validate Variants
- Eval Variants
- Combine Variants
- NGS: Variant Detection
- NGS: Peak Calling
- NGS: Simulation
- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models
- SnpEff tools
- Phenotype Association
- VCF Tools
- IGVTools
- MSI

Binding for reference-ordered data 4

Binding Type: t dbSNP

ROD file: u 6: dbsnp\_137.hg19.vcf

Use as training/truth/known sites: Set training/truth/known sites

Is Known Site: v

Is Training Site:

Is Truth Site:

Is Bad Site:

prior probability of being true: w 2.0

Remove Binding for reference-ordered data 4

Add new Binding for reference-ordered data

annotations which should used for calculations: Select All Unselect All AlleleBalance

History

Variant\_Detection\_RISS 782.8 MB

- 38: Select Variants on data 34 (log)
- 37: SNPs
- 36: Unified Genotyper on data 28, data 6, and data 3 (log)
- 35: Unified Genotyper on data 28, data 6, and data 3 (metrics)
- 34: Unified Genotyper on data 28, data 6, and data 3 (VCF)
- 33: Analyze Covariates on data 30 (log)
- 32: Analyze Covariates on data 30 (HTML)
- 31: Count Covariates on data 6 and data 28 (log)
- 30: Count Covariates on data 6 and data 28 (Covariate File)
- 29: Table Recalibration on data 24

x) Click on “Add new Addition Annotations” **five times**

y) Add the annotations below

Annotation name: -> “FS”

Annotation name: -> “HRun”

Annotation name: -> “MQRankSum”

Annotation name: -> “QD”

Annotation name: -> “ReadPosRankSum”

The screenshot shows the Galaxy interface with the 'Additional annotations' section highlighted. A red circle surrounds the 'Add new Additional annotation' button, which has a red 'X' over it. The 'Recalibration mode:' dropdown is set to 'SNP'. The 'Basic or Advanced GATK options:' dropdown is set to 'Basic'. The 'Basic or Advanced Analysis options:' dropdown is set to 'Basic'. The 'Execute' button is visible at the bottom.

The screenshot shows the Galaxy interface with the 'Additional annotations' section highlighted by a red box. Each entry in the list has a red 'X' next to its name. The annotations listed are: FS, Remove Additional annotation 1; Additional annotation 2, Annotation name: HRun, Remove Additional annotation 2; Additional annotation 3, Annotation name: MQRankSum, Remove Additional annotation 3; Additional annotation 4, Annotation name: QD, Remove Additional annotation 4; Additional annotation 5, Annotation name: ReadPosRan, Remove Additional annotation 5.

- z) Recalibration mode: -> SNP  
 aa) Basic or Advanced GATK options: -> Advanced  
 bb) Click the “Add new Operate on Genomic intervals” button  
 cc) Genomic intervals: -> “tutorial\_exons.bed”

The screenshot shows the Galaxy interface with the following configuration:

- Recalibration mode:** Set to **SNP** (highlighted with a red circle).
- Basic or Advanced GATK options:** Set to **Advanced** (highlighted with a red circle).
- Add new Operate on Genomic intervals** button: This button is highlighted with a red circle.
- Exclude Genomic intervals** button: Located below the operate button.

The right panel shows the history of workflows, including:

- Variant\_Detection\_RISS
- 38: Select Variants on data 34 (log)
- 37: SNPs
- 36: Unified Genotyper on data 28, data 6, and data 3 (log)
- 35: Unified Genotyper on data 28, data 6, and data 3 (metrics)
- 34: Unified Genotyper on data 28, data 6, and data 3 (VCF)
- 33: Analyze Covariates on data 30 (log)
- 32: Analyze Covariates on data 30 (HTML)
- 31: Count Covariates on data 6 and data 28 (log)
- 30: Count Covariates on data 6 and data 28 (Covariate File)
- 29: Table Recalibration on data 24 (log)

The screenshot shows the Galaxy interface with the following configuration:

- How strict should we be in validating the pedigree information:** Set to **STRICT**.
- Read Filters**: -rf,--read\_filter <read\_filter>
- Operate on Genomic intervals**: -L,--intervals <intervals>
- Operate on Genomic intervals 1**: This step is selected and highlighted with a red circle.
- Genomic intervals:** A dropdown menu is open, showing the option **3: tutorial\_exons.bed** (highlighted with a red circle).
- Add new Operate on Genomic intervals** button: This button is highlighted with a red circle.
- Exclude Genomic intervals**: -XL,--excludeIntervals <excludeIntervals>
- Interval set rule:** UNION
- Type of reads downsampling to employ at a given locus:** NONE

The right panel shows the history of workflows, including:

- Variant\_Detection\_RISS
- 38: Select Variants on data 34 (log)
- 37: SNPs
- 36: Unified Genotyper on data 28, data 6, and data 3 (log)
- 35: Unified Genotyper on data 28, data 6, and data 3 (metrics)
- 34: Unified Genotyper on data 28, data 6, and data 3 (VCF)
- 33: Analyze Covariates on data 30 (log)
- 32: Analyze Covariates on data 30 (HTML)
- 31: Count Covariates on data 6 and data 28 (log)
- 30: Count Covariates on data 6 and data 28 (Covariate File)
- 29: Table Recalibration on data 24 (log)

- dd) Basic or Advanced Analysis options: -> Advanced
- ee) maximum number of Gaussians to try during variational Bayes Algorithm <maxGaussians>: -> 1
- ff) How to specify bad variants: -> Number
- gg) minimum amount of worst scoring variants to use when building the Gaussian mixture model of bad variants. Will override -percentBad argument if necessary <minNumBadVariants>: -> 50
- hh) Click "Execute"

**Variant Detection\_RISS**

782.8 MB

38: Select Variants on data 34 (log) ① 0 ×

37: SNPs ① 0 ×

36: Unified Genotyper on data 28, data 6, and data 3 (log)

35: Unified Genotyper on data 28, data 6, and data 3 (metrics)

34: Unified Genotyper on data 28, data 6, and data 3 (VCF)

33: Analyze Covariates on data 30 (log)

32: Analyze Covariates on data 30 (HTML)

31: Count Covariates on data 6 and data 28 (log)

30: Count Covariates on data 6 and data 28 (Covariate File)

29: Table Recalibration on data 24 (log)

Using 1.5 GB

**Variant Detection\_RISS**

782.8 MB

38: Select Variants on data 34 (log) ① 0 ×

37: SNPs ① 0 ×

36: Unified Genotyper on data 28, data 6, and data 3 (log)

35: Unified Genotyper on data 28, data 6, and data 3 (metrics)

34: Unified Genotyper on data 28, data 6, and data 3 (VCF)

33: Analyze Covariates on data 30 (log)

32: Analyze Covariates on data 30 (HTML)

31: Count Covariates on data 6 and data 28 (log)

30: Count Covariates on data 6 and data 28 (Covariate File)

29: Table Recalibration on data 24 (log)

Using 1.5 GB

### 7.3 Apply recalibration

- Load Apply Variant Recalibration tool from the tool pane: "NGS: GATK Tools -> Apply Variant Recalibration"
- Variant file to annotate: -> "SNPs"
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Recalibration mode: -> SNP
- Click "Execute"

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

**Tools**

- Variant Filtration on VCF files
- Select Variants from VCF files
- VARIANT QUALITY SCORE RECALIBRATION
- Variant Recalibrator
- Apply Variant Recalibration** (circled in red)
- VARIANT UTILITIES
- Validate Variants
- Eval Variants
- Combine Variants
- NGS: Variant Detection
- NGS: Peak Calling
- NGS: Simulation
- SNP/WGA: Data; Filters

**Apply Variant Recalibration (version 0.0.4)**

Choose the source for the reference list: Locally cached

**Variants**

-input,--input <input>

**Variant 1**

**Variant file to annotate:** 37: SNPs (circled in red)

Add new Variant

**Variant Recalibration file:** 39: Variant Recalibrator on data 6, data 37, and others (Recalibration File)

-recalFile,--recal\_file <recal\_file>

**Variant Tranches file:** 40: Variant Recalibrator on data 6, data 37, and others (Tranches File)

-tranchesFile,--tranches\_file <tranches\_file>

**History**

Variant\_Detection\_RISS 782.8 MB

43: Variant Recalibrator on data 6, data 37, and others (log)

42: Variant Recalibrator on data 6, data 37, and others (PDF File)

41: Variant Recalibrator on data 6, data 37, and others (RScript File)

40: Variant Recalibrator on data 6, data 37, and others (Tranches File)

39: Variant Recalibrator on data 6, data 37, and others (Recalibration File)

38: Select Variants on data 34 (log)

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

**Tools**

- Variant Filtration on VCF files
- Select Variants from VCF files
- VARIANT QUALITY SCORE RECALIBRATION
- Variant Recalibrator
- Apply Variant Recalibration
- VARIANT UTILITIES
- Validate Variants
- Eval Variants
- Combine Variants
- NGS: Variant Detection
- NGS: Peak Calling
- NGS: Simulation
- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models
- SnpEff tools
- Phenotype Association

**Variant Recalibrator**

**Variant Tranches file:** 40: Variant Recalibrator on data 6, data 37, and others (Tranches File)

-tranchesFile,--tranches\_file <tranches\_file>

**Using reference genome:** Homo sapiens hg19\_canonical (GATK) (circled in red)

-R,--reference\_sequence <reference\_sequence>

**Basic or Advanced GATK options:** Basic

**Recalibration mode:** SNP (circled in red)

-mode,--mode <mode>

**Ignore Filters**

-ignoreFilter,--ignore\_filter <ignore\_filter>

Add new Ignore Filter

**truth sensitivity level at which to start filtering, used here to indicate filtered variants in plots:** 99.0

-ts\_filter\_level,--ts\_filter\_level <ts\_filter\_level>

**Execute** (circled in red)

**History**

Variant\_Detection\_RISS 782.8 MB

43: Variant Recalibrator on data 6, data 37, and others (log)

42: Variant Recalibrator on data 6, data 37, and others (PDF File)

41: Variant Recalibrator on data 6, data 37, and others (RScript File)

40: Variant Recalibrator on data 6, data 37, and others (Tranches File)

39: Variant Recalibrator on data 6, data 37, and others (Recalibration File)

38: Select Variants on data 34 (log)

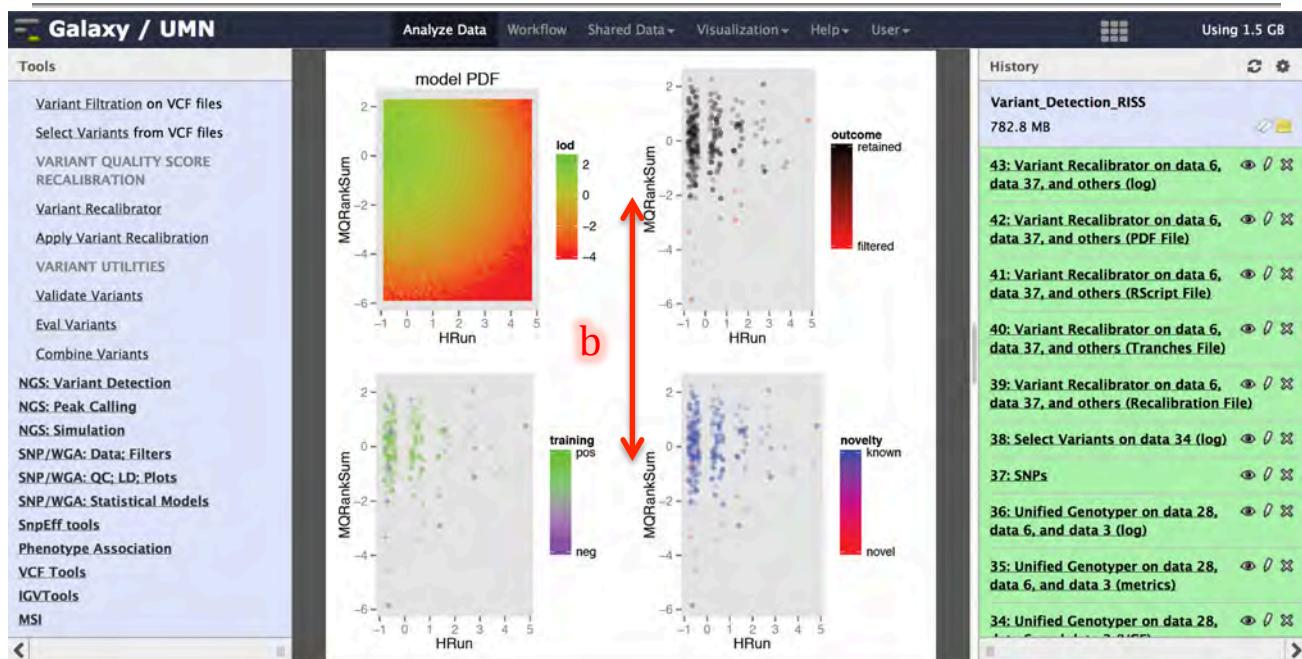
37: SNPs

36: Unified Genotyper on data 28, data 6, and data 3 (log)

## 7.4 Review Variant Recalibration Models

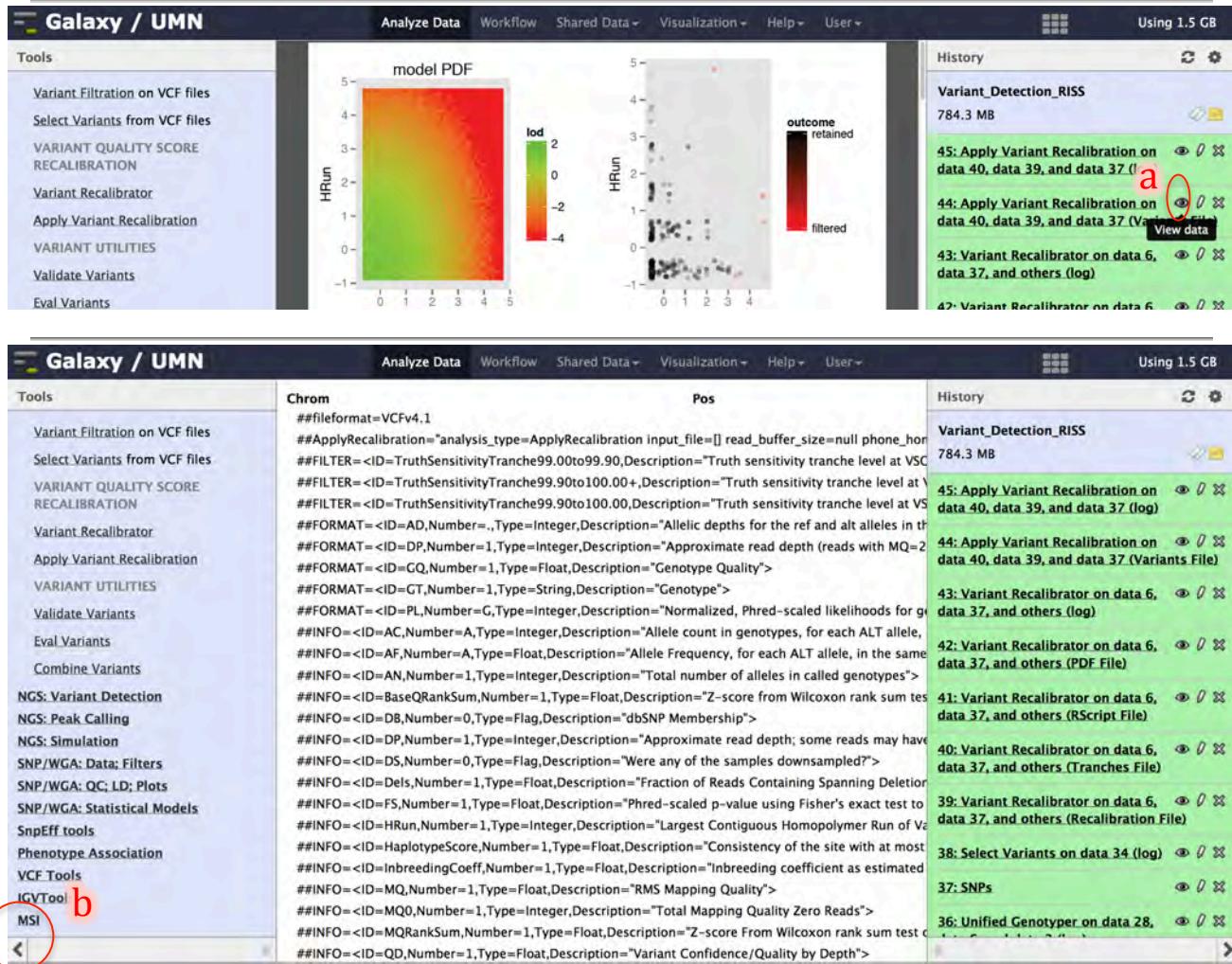
- In the history pane click the eye icon next to the name of the *Variant Recalibrator* pdf, "Variant Recalibrator... (PDF File)" file to display the file in the center pane
- Move the *center pane* up and down to examine how well the models are discriminating between positive (known variants/SNPs) and negative (bad variants) training data.

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". The right side of the screen displays the "History" pane titled "Using 1.5 GB". The history list contains several entries, with the fourth entry (42: Variant Recalibrator...) circled in red and labeled 'a'. A green message box in the center states: "The following job has been successfully added to the queue: 39: Variant Recalibrator on data 6, data 37, and others (Recalibration File) 40: Variant Recalibrator on data 6, data 37, and others (Tranches File) 41: Variant Recalibrator on data 6, data 37, and others (RScript File) 42: Variant Recalibrator on data 6, data 37, and others (PDF File) 43: Variant Recalibrator on data 6, data 37, and others (log) You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered."



## 7.5 Review Recalibrated Variants (SNPs)

- In the history pane click the eye icon next to the variant file, "Apply Variant .... (Variant File)", produced by the "NGS: GATK Tools -> Apply Variant Recalibration" tool
- Click the arrow at the bottom of the tools pane to the left of the browser minimize it



- c) Click the arrow at the bottom of the *history pane* to the right of the browser minimize it
- d) Scroll to the right and look at the column labeled “FILTER”. Compare raw variant file produced by the Genotyper
- e) Click the arrow at the bottom-left corner of your browser to bring the *tools pane* back to view

**Galaxy / UMN**

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

Chrom	Pos
#fileformat=VCFv4.1	
##ApplyRecalibration="analysis_type=ApplyRecalibration input_file=[] read_buffer_size=null phone_home=NO_ET read_filter=[] intervals=null exclude=false"	
##FILTER=<ID=TruthSensitivityTranche99.00to99.90,Description="Truth sensitivity tranche level at VSQ Lod: -5.9187 <= x < -1.1549">	
##FILTER=<ID=TruthSensitivityTranche99.90to100.00+,Description="Truth sensitivity tranche level at VQS Lod < -113.3625">	
##FORMAT=<ID=DP,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">	
##FORMAT=<ID=DP,Q,Number=1,Type=Float,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">	
##FORMAT=<ID=CQ,Number=1,Type=Float,Description="Genotype Quality">	
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">	
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">	
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">	
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">	
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">	
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">	
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">	
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">	
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">	
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">	
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">	
##INFO=<ID=HRun,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of Variant Allele In Either Direction">	
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">	
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample within a family">	
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">	
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">	
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">	
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">	

History

- Variant\_Detection\_RISS 784.3 MB
- 45: Apply Variant Recalibration on data 40, data 39, and data 37 (log)
- 44: Apply Variant Recalibration on data 40, data 39, and data 37 (Variants File)
- 43: Variant Recalibrator on data 6, data 37, and others (log)
- 42: Variant Recalibrator on data 6, data 37, and others (PDF File)
- 41: Variant Recalibrator on data 6, data 37, and others (RScript File)
- 40: Variant Recalibrator on data 6, data 37, and others (Tranches File)
- 39: Variant Recalibrator on data 6, data 37, and others (Recalibration File)
- 38: Select Variants on data 34 (log)
- 37: SNPs
- 36: Unified Genotyper on data 28 C

**Galaxy / UMN**

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

QUAL	FILTER
39.88	PASS
135.79	PASS
494.43	PASS
294.10	PASS
1069.60	PASS
1831.14	PASS
112.50	TruthSensitivityTranche99.90to100.00
708.53	PASS
709.61	PASS
1586.61	PASS
1809.61	PASS
191.57	PASS
170.14	PASS
100.75	PASS

d

e

- f) Click the arrow at the bottom-right corner of your browser to bring the *tools pane* back to view

The screenshot shows the Galaxy / UMN web application. At the top, there is a navigation bar with links for Analyze Data, Workflow, Shared Data, Visualization, Help, User, and a grid icon labeled "Using 1.5 GB". On the left, a sidebar titled "Tools" lists various bioinformatics tools categorized under "Variant Filtration on VCF files", "VARIANT QUALITY SCORE RECALIBRATION", "Variant Recalibrator", "Apply Variant Recalibration", "VARIANT UTILITIES", "Validate Variants", "Eval Variants", "Combine Variants", "NGS: Variant Detection", "NGS: Peak Calling", "NGS: Simulation", "SNP/WGA: Data; Filters", "SNP/WGA: QC; LD; Plots", "SNP/WGA: Statistical Models", "SnpEff tools", "Phenotype Association", "VCF Tools", "IGVTools", and "MSI". A red letter "f" is overlaid on the sidebar area. To the right of the sidebar, a table displays variant quality scores and filters. A red circle highlights the bottom-right corner of the main content area, where a small arrow icon points towards the sidebar. The table data is as follows:

	QUAL	FILTER
	39.88	PASS
	135.79	PASS
	494.43	PASS
	294.10	PASS
	1069.60	PASS
	1831.14	PASS
	112.50	TruthSensitivity/Tranche99.90to100.00
	708.53	PASS
	709.61	PASS
	1586.61	PASS
	1809.61	PASS
	191.57	PASS
	170.14	PASS

## 7.6 Select INDELS

- Load *Select variants* tool from the tool pane: "NGS: GATK Tools -> Select Variants from VCF files"
- Variant file to select: -> "...Unified Genotyper....(VCF)"
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Basic or Advanced Analysis options: -> Advanced

The screenshot shows the Galaxy interface with the title bar "Galaxy / UMN". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". On the left, a sidebar titled "Tools" lists several categories: ANNOTATION, FILTRATION, VARIANT QUALITY SCORE, and RECALIBRATION. Under the FILTRATION category, the "Select Variants from VCF files" tool is highlighted with a red oval and labeled 'a'. To its right is the main workspace and a history panel on the right.

This screenshot shows the "Select Variants (version 0.0.2)" configuration dialog. The "Choose the source for the reference list:" dropdown is set to "Locally cached" (labeled 'b'). The "Variant file to select:" dropdown contains "34: Unified Genotyper on data 28, data 6, and data 3 (VCF)" (labeled 'b'). The "Using reference genome:" dropdown contains "Homo sapiens hg19\_canonical (GATK)" (labeled 'c'). The history panel on the right shows three entries related to variant recalibration.

This screenshot shows the configuration dialog with the "Basic or Advanced GATK options:" dropdown set to "Advanced" (labeled 'd'). The "Basic or Advanced Analysis options:" dropdown is also set to "Advanced". The "Exclude Samples by files" section is visible, containing the command "-xl\_sf,--exclude\_sample\_file <exclude\_sample\_file>". The history panel on the right shows three entries related to variant recalibration.

- e) Select only a certain type of variants from the input file: -> check box next to ✓ INDEL
- f) Click "Execute"
- g) Click the pencil icon next to the output file to edit attributes
- h) Enter "INDELS" under Name:
- i) Click "Save"

Galaxy / UMN

Select a random subset of variants:  
Use all variants

Don't include loci found to be non-variant after the subsetting procedure:  
-env,--excludeNonVariants

Select only a certain type of variants from the input file:  
Select All Unselect All  
 INDEL  
 SNP  
 MIXED  
 MNP  
 SYMBOLIC  
 NO\_VARIATION  
-selectType,--selectTypeToInclude <selectTypeToInclude>

**Execute**

History

- Variant\_Detection\_RISS 784.3 MB
- 45: Apply Variant Recalibration on data 40, data 39, and data 37 (log)
- 44: Apply Variant Recalibration on data 40, data 39, and data 37 (Variants File)
- 43: Variant Recalibrator on data 6, data 37, and others (log)
- 42: Variant Recalibrator on data 6, data 37, and others (PDF File)
- 41: Variant Recalibrator on data 6, data 37, and others (RScript File)

Galaxy / UMN

The following job has been successfully added to the queue:

46: Select Variants on data 34 (Variant File)  
47: Select Variants on data 34 (log)

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

- Variant\_Detection\_RISS 784.4 MB
- 47: Select Variants on data 34 (log)
- 46: Select Variants on data 34 (Variant File)
- 45: Apply Variant Recalibration on data 40, data 39, and data 37 (log)

Galaxy / UMN

Attributes Convert Format Datatype Permissions

Edit Attributes

Name: **INDELS**

Info:

Annotation / Notes:

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build:

Human hg19 in GATK canonical chr...

Number of comment lines:  
 53

**Save**

History

- Variant\_Detection\_RISS 784.4 MB
- 47: Select Variants on data 34 (log)
- 46: Select Variants on data 34 (Variant File)
- 45: Apply Variant Recalibration on data 40, data 39, and data 37 (log)
- 44: Apply Variant Recalibration on data 40, data 39, and data 37 (Variants File)
- 43: Variant Recalibrator on data 6, data 37, and others (log)
- 42: Variant Recalibrator on data 6, data 37, and others (PDF File)
- 41: Variant Recalibrator on data 6, data 37, and others (RScript File)

## 7.7 Recalibrate INDELs

- Load *variant recalibration* tool from the tool pane: "NGS: GATK Tools -> Variant Recalibrator"
- Variant file to recalibrate: -> "...INDELs"
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Click on "Add new Binding for reference-ordered data"
- Binding Type: -> Custom
- ROD Name: -> mills
- ROD file: -> Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf
- Use as training/truth/known sites: -> Set training/truth/known sites
  - Is Training Site: -> ✓
  - Is Truth Site: -> ✓
- prior probability of being true: -> 12.0
- Click on "Add new Binding for reference-ordered data"

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

Tools

- ANNOTATION
- Variant Annotator
- FILTRATION
- Variant Filtration on VCF files
- Select Variants from VCF files
- VARIANT QUALITY SCORE RECALIBRATION
- Variant Recalibrator**
- Apply Variant Recalibration
- VARIANT UTILITIES
- Validate Variants
- Eval Variants
- Combine Variants
- NGS: Variant Detection
- NGS: Peak Calling
- NGS: Simulation
- SNP/WGA: Data; Filters

Variant Recalibrator (version 0.0.4)

Choose the source for the reference list: Locally cached

**b**

Variant 1

Variant file to recalibrate: 46: INDELs

Add new Variant

Using reference genome: Homo sapiens hg19\_canonical (GATK)

c

Binding for reference-ordered data

Annotations which should used for calculations:

d

History

- Variant\_Detection\_RISS
- 784.4 MB
- 47: Select Variants on data 34 (log) 0/0
- 46: INDELs 0/0
- 45: Apply Variant Recalibration on data 40, data 39, and data 37 (log) 0/0
- 44: Apply Variant Recalibration on data 40, data 39, and data 37 (Variants File) 0/0
- 43: Variant Recalibrator on data 6, data 37, and others (log) 0/0
- 42: Variant Recalibrator on data 6, data 37, and others (PDF File) 0/0
- 41: Variant Recalibrator on data 6, data 37, and others (RScript File) 0/0
- 40: Variant Recalibrator on data 6, data 37, and others (Tranches File) 0/0

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

Tools

- ANNOTATION
- Variant Annotator
- FILTRATION
- Variant Filtration on VCF files
- Select Variants from VCF files
- VARIANT QUALITY SCORE RECALIBRATION
- Variant Recalibrator
- Apply Variant Recalibration
- VARIANT UTILITIES
- Validate Variants
- Eval Variants
- Combine Variants
- NGS: Variant Detection
- NGS: Peak Calling
- NGS: Simulation
- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models
- SnpEff tools
- Phenotype Association

Binding for reference-ordered data 1

Binding Type: Custom

e

ROD Name: mills

f

ROD file: 8: Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf

g

Use as training/truth/known sites: Set training/truth/known sites

Is Known Site:

Is Training Site:  h

Is Truth Site:

Is Bad Site:

prior probability of being true: 12.0

i

Remove Binding for reference-ordered data 1

j

Annotations which should used for calculations:

History

- Variant\_Detection\_RISS
- 784.4 MB
- 47: Select Variants on data 34 (log) 0/0
- 46: INDELs 0/0
- 45: Apply Variant Recalibration on data 40, data 39, and data 37 (log) 0/0
- 44: Apply Variant Recalibration on data 40, data 39, and data 37 (Variants File) 0/0
- 43: Variant Recalibrator on data 6, data 37, and others (log) 0/0
- 42: Variant Recalibrator on data 6, data 37, and others (PDF File) 0/0
- 41: Variant Recalibrator on data 6, data 37, and others (RScript File) 0/0
- 40: Variant Recalibrator on data 6, data 37, and others (Tranches File) 0/0
- 39: Variant Recalibrator on data 6, data 37, and others (Recalibration File) 0/0
- 38: Select Variants on data 34 (log) 0/0

- k) Binding Type: -> dbSNP
- l) ROD file: -> dbsnp\_137.hg19.vcf
- m) Use as training/truth/known sites: -> Set training/truth/known sites  
Is Known Site:-> ✓
- n) prior probability of being true: -> 2.0
- o) Click on “Add new Addition Annotations” three times

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

**Tools**

- ANNOTATION
  - Variant Annotator
- FILTRATION
  - Variant Filtration on VCF files
  - Select Variants from VCF files
- VARIANT QUALITY SCORE RECALIBRATION
  - Variant Recalibrator
  - Apply Variant Recalibration
- VARIANT UTILITIES
  - Validate Variants
  - Evaluate Variants
  - Combine Variants
- NGS: Variant Detection
  - NGS: Peak Calling
  - NGS: Simulation
- SNP/WGA: Data; Filters
  - SNP/WGA: QC; LD; Plots
  - SNP/WGA: Statistical Models
- SnpEff tools
- Phenotype Association

**Binding for reference-ordered data 2**

**Binding Type:** dbSNP (k)

**ROD file:** 6: dbsnp\_137.hg19.vcf (l)

**Use as training/truth/known sites:** Set training/truth/known sites (m)

**Is Known Site:**

**Is Training Site:**

**Is Truth Site:**

**Is Bad Site:**

**prior probability of being true:** 2.0 (n)

**Add new Binding for reference-ordered data**

**annotations which should used for calculations:** Select All Unselect All

**History**

- Variant\_Detection\_RISS 784.4 MB
- 47: Select Variants on data 34 (log) ④ / ④
- 46: INDELS ④ / ④
- 45: Apply Variant Recalibration on data 40, data 39, and data 37 (log)
- 44: Apply Variant Recalibration on data 40, data 39, and data 37 (Variants File)
- 43: Variant Recalibrator on data 6, data 37, and others (log)
- 42: Variant Recalibrator on data 6, data 37, and others (PDF File)
- 41: Variant Recalibrator on data 6, data 37, and others (RScript File)
- 40: Variant Recalibrator on data 6, data 37, and others (Tranches File)
- 39: Variant Recalibrator on data 6, data 37, and others (Recalibration File)
- 38: Select Variants on data 34 (log) ④ / ④

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

**Tools**

- ANNOTATION
  - Variant Annotator
- FILTRATION
  - Variant Filtration on VCF files
  - Select Variants from VCF files
- VARIANT QUALITY SCORE RECALIBRATION
  - Variant Recalibrator
  - Apply Variant Recalibration
- VARIANT UTILITIES
  - Validate Variants

**Additional annotations**

-an,--use\_annotation <use\_annotation>

**Add new Additional annotation** (o)

**Recalibration mode:** SNP

-mode,--mode <mode>

**Basic or Advanced GATK options:** Basic

**Basic or Advanced Analysis options:** Basic

**Execute**

**History**

- Variant\_Detection\_RISS 784.4 MB
- 47: Select Variants on data 34 (log) ④ / ④
- 46: INDELS ④ / ④
- 45: Apply Variant Recalibration on data 40, data 39, and data 37 (log)
- 44: Apply Variant Recalibration on data 40, data 39, and data 37 (Variants File)
- 43: Variant Recalibrator on data 6, data 37, and others (log)

p) Add the annotations below

Annotation name: -> "FS"

Annotation name: -> "MQRankSum"

Annotation name: -> "ReadPosRankSum"

q) Recalibration mode: -> INDEL

r) Basic or Advanced GATK options: -> Advanced

s) Click the "Add new Operate on Genomic intervals" button

The screenshot shows the Galaxy interface with a tool panel on the left and a history panel on the right. A red box labeled 'p' highlights the 'Additional annotation' section where three annotations (FS, MQRankSum, ReadPosRan) are listed. Below this, a red circle labeled 'q' surrounds the 'Recalibration mode' dropdown set to 'INDEL'. Another red circle labeled 'r' surrounds the 'Basic or Advanced GATK options' dropdown set to 'Advanced'. The history panel on the right lists several workflow steps, including variant detection, recalibration, and pedigree analysis.

The screenshot shows the Galaxy interface with a tool panel on the left and a history panel on the right. A red circle labeled 's' highlights the 'Add new Operate on Genomic intervals' button in the 'Operate on Genomic intervals' section. The history panel on the right lists several workflow steps, including variant detection, recalibration, and pedigree analysis.

- t) Genomic intervals: -> "tutorial\_exons.bed"
- u) Basic or Advanced Analysis options: -> Advanced
- v) maximum number of Gaussians to try during variational Bayes Algorithm <maxGaussians>: -> 1
- w) How to specify bad variants: -> Number
- x) minimum amount of worst scoring variants to use when building the Gaussian mixture model of bad variants. Will override -percentBad argument if necessary <minNumBadVariants>: -> 5
- y) Click "Execute"

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

Tools

ANNOTATION Variant Annotator FILTRATION Variant Filtration on VCF files Select Variants from VCF files VARIANT QUALITY SCORE RECALIBRATION Variant Recalibrator Apply Variant Recalibration VARIANT UTILITIES

Read Filters  
-rf,--read\_filter <read\_filter>  
Add new Read Filter

Operate on Genomic Intervals  
-L,--intervals <intervals>

**Operate on Genomic intervals 1**

Genomic intervals:  
**3: tutorial\_exons.bed** t

Remove Operate on Genomic intervals 1  
Add new Operate on Genomic intervals

History

Variant\_Detection\_RISS 784.4 MB

47: Select Variants on data 34 (log) ① 0 ②

46: INDELS ① 0 ②

45: Apply Variant Recalibration on data 40, data 39, and data 37 (log)

44: Apply Variant Recalibration on data 40, data 39, and data 37 (Variants File)

43: Variant Recalibrator on data 6, data 37, and others (log) ① 0 ②

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

Tools

ANNOTATION Variant Annotator FILTRATION Variant Filtration on VCF files Select Variants from VCF files VARIANT QUALITY SCORE RECALIBRATION

Makes the GATK behave non deterministically, that is, the random numbers generated will be different in every run:  
-ndrs,--nonDeterministicRandomSeed

**Basic or Advanced Analysis options:** u

maximum number of Gaussians to try during variational Bayes Algorithm:  
**1** v

-mG,--maxGaussians <maxGaussians>

History

Variant\_Detection\_RISS 784.4 MB

47: Select Variants on data 34 (log) ① 0 ②

46: INDELS ① 0 ②

45: Apply Variant Recalibration on data 40, data 39, and data 37 (log)

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

Tools

ANNOTATION Variant Annotator FILTRATION Variant Filtration on VCF files Select Variants from VCF files VARIANT QUALITY SCORE RECALIBRATION Variant Recalibrator Apply Variant Recalibration VARIANT UTILITIES Validate Variants Eval Variants Combine Variants NGS: Variant Detection NGS: Peak Calling NGS: Simulation SNP/WGA: Data; Filters SNP/WGA: QC; LD; Plots SNP/WGA: Statistical Models SnpEff tools Phenotype Association

How to specify bad variants:  
Number w

minimum amount of worst scoring variants to use when building the Gaussian mixture model of bad variants. Will override -percentBad argument if necessary:  
**5** x

-minNumBad,--minBadVariants <minNumBadVariants>

expected novel Ti/Tv ratio to use when calculating FDR tranches and for display on optimization curve output figures. (approx 2.15 for whole genome experiments). ONLY USED FOR PLOTTING PURPOSES:  
2.15

-titv,--target\_titv <target\_titv>

levels of novel false discovery rate (FDR, implied by ti/tv) at which to slice the data. (in percent, that is 1.0 for 1 percent):  
100.0, 99.9,

-tstranche,--TStranchise <TStranchise>

Ignore Filters  
-ignoreFilter,--ignore\_filter <ignore\_filter>  
Add new Ignore Filter

truth sensitivity level at which to start filtering, used here to indicate filtered variants in plots:  
99.0

-ts\_Filter\_Level,--ts\_filter\_level <ts\_filter\_level>

**Execute** y

History

Variant\_Detection\_RISS 784.4 MB

47: Select Variants on data 34 (log) ① 0 ②

46: INDELS ① 0 ②

45: Apply Variant Recalibration on data 40, data 39, and data 37 (log)

44: Apply Variant Recalibration on data 40, data 39, and data 37 (Variants File)

43: Variant Recalibrator on data 6, data 37, and others (log) ① 0 ②

42: Variant Recalibrator on data 6, data 37, and others (PDF File)

41: Variant Recalibrator on data 6, data 37, and others (RScript File)

40: Variant Recalibrator on data 6, data 37, and others (Tranches File)

39: Variant Recalibrator on data 6, data 37, and others (Recalibration File)

38: Select Variants on data 34 (log) ① 0 ②

## 7.8 Apply recalibration

- Load Apply Variant Recalibration tool from the tool pane: "NGS: GATK Tools -> Apply Variant Recalibration"
- Variant file to annotate: -> "INDELS"
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Recalibration mode: -> INDEL
- Click "Execute"

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". On the left, the "Tools" sidebar lists various bioinformatics tools. The "Apply Variant Recalibration" tool is highlighted with a red oval and labeled 'a'. A green success message box contains the following text:

```

The following job has been successfully added to the queue:
48: Variant Recalibrator on data 3, data 8, and others (Recalibration File)
49: Variant Recalibrator on data 3, data 8, and others (Tranches File)
50: Variant Recalibrator on data 3, data 8, and others (RScript File)
51: Variant Recalibrator on data 3, data 8, and others (PDF File)
52: Variant Recalibrator on data 3, data 8, and others (log)

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

```

The right side of the screen shows the "History" pane with several completed jobs, including "Variant\_Detection\_RISS" and "Variant Recalibrator" entries.

This screenshot shows the "Apply Variant Recalibration (version 0.0.4)" tool configuration page. The "Variants" section is expanded, showing a dropdown menu where "46: INDELS" is selected, indicated by a red oval and labeled 'b'. The "Using reference genome" dropdown is also circled in red and labeled 'c', showing "Homo sapiens hg19\_canonical (GATK)". Other sections visible include "Variant Recalibration file" (set to "48: Variant Recalibrator on data 3, data 8, and others (Recalibration File)"), "Variant Tranches file" (set to "49: Variant Recalibrator on data 3, data 8, and others (Tranches File)"), and "Basic or Advanced GATK options" (set to "Basic"). The right side of the screen shows the "History" pane with several completed jobs.

This screenshot shows the final configuration of the "Apply Variant Recalibration" tool. The "Recalibration mode" dropdown is set to "INDEL", indicated by a red oval and labeled 'd'. The "Execute" button at the bottom is circled in red and labeled 'e'. The right side of the screen shows the "History" pane with several completed jobs.

## 7.9 Review Recalibrated Variants (INDELS)

- In the history pane click the eye icon next to the variant file, "Variant Filtration ..... (Variant File)", produced by the "NGS: GATK Tools -> Variant Filtration on VCF files" tool
- Click the arrow at the bottom of the tools pane to the left of the browser minimize it

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". A message box in the center says: "The following job has been successfully added to the queue: 53: Apply Variant Recalibration on data 49, data 48, and data 46 (Variants File) 54: Apply Variant Recalibration on data 49, data 48, and data 46 (log)". Below this, instructions say: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered." To the right is the "History" pane, which lists several jobs: "Variant\_Detection\_RISS 784.6 MB", "54: Apply Variant Recalibration on data 49, data 48, and data 46 (log)" (with a red circle 'a' around the eye icon), "53: Apply Variant Recalibration on data 49, data 48, and data 46 (Variants File)" (with a red circle 'a' around the eye icon), and "52: Variant Recalibrator on data 3, data 8, and others (log)".

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". The "Tools" sidebar on the left lists various bioinformatics tools. In the main area, there is a large block of log output starting with "Chrom" and "Pos", followed by numerous lines of genomic data and statistics. To the right is the "History" pane, which lists several jobs: "Variant\_Detection\_RISS 784.6 MB", "54: Apply Variant Recalibration on data 49, data 48, and data 46 (log)", "53: Apply Variant Recalibration on data 49, data 48, and data 46 (Variants File)", "52: Variant Recalibrator on data 3, data 8, and others (log)", "51: Variant Recalibrator on data 3, data 8, and others (PDF File)", "50: Variant Recalibrator on data 3, data 8, and others (RScript File)", "49: Variant Recalibrator on data 3, data 8, and others (Tranches File)", "48: Variant Recalibrator on data 3, data 8, and others (Recalibration File)", "47: Select Variants on data 34 (log)", "46: INDELS", and "45: Apply Variant Recalibration on data 49, data 48, and data 46 (Variants File)". A red circle 'b' is drawn around the bottom-left corner of the tools sidebar.

- c) Click the arrow at the bottom of the *history pane* to the right of the browser minimize it
- d) Scroll to the left and right to inspect recalibrated INDELS
- e) Click the arrow at the bottom-left corner of your browser to bring the *tools pane* back to view

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

Chrom	Pos	ID	History
	##fileformat=VCFv4.1		Variant_Detection_RISS
	##ApplyRecalibration="analysis_type=ApplyRecalibration input_file=[] read_buffer_size=null phone_home=NO_ET read_filter=[] intervals=null exclude=false"		784.6 MB
	##FILTER=<ID=TruthSensitivityTranche99.00to99.90,Description="Truth sensitivity tranche level at VSQ Lod: 0.426 <= x < 0.426">		S4: Apply Variant Recalibration on data 49, data 48, and data 46 (log)
	##FILTER=<ID=TruthSensitivityTranche99.90to100.00+,Description="Truth sensitivity tranche level at VQS Lod < -5.2111">		S3: Apply Variant Recalibration on data 49, data 48, and data 46 (Variants File)
	##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">		S2: Variant Recalibrator on data 3, data 8, and others (log)
	##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">		S1: Variant Recalibrator on data 3, data 8, and others (PDF File)
	##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">		S0: Variant Recalibrator on data 3, data 8, and others (RScript File)
	##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">		49: Variant Recalibrator on data 3, data 8, and others (Tranches File)
	##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification"		48: Variant Recalibrator on data 3, data 8, and others (Recalibration File)
	##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">		47: Select Variants on data 34 (log)
	##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">		46: INDELS
	##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">		45: Apply Variant Recalibration on C
	##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">		
	##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">		
	##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">		
	##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">		
	##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">		
	##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">		
	##INFO=<ID=HRUN,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of Variant Allele In Either Direction">		
	##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">		
	##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample wise">		
	##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">		
	##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">		
	##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">		
	##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">		

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

REF	ALT
T	TA
CCAT	C
GA	G
GT	G
G	GA
A	AT
AT	A
A	ATCT
CT	C
GT	G
TGCA	T
G	GGGC
AC	A
A	AT
CAA	C
TAGCAGTGAC	T
G	GA
TGA	T
ATATCT	A
C	CT
G	GT

- f) Click the arrow at the bottom-right corner of your browser to bring the *tools pane* back to view

The screenshot shows the Galaxy / UMN web application. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', 'User', and 'Using 1.5 GB'. On the left, a 'Tools' sidebar lists various bioinformatics tools categorized under 'ANNOTATION', 'FILTRATION', 'VARIANT QUALITY SCORE', 'RECALIBRATION', 'VARIANT UTILITIES', 'NGS', 'SNP/WGA', 'SnpEff tools', and 'Phenotype Association'. The main content area displays a table of genomic variants:

	REF	ALT
T	TA	
CCAT	C	
GA	G	
GT	G	
G	GA	
A	AT	
AT	A	
A	ATCT	
CT	C	
GT	G	
TGCA	T	
G	GGGC	
AC	A	
A	AT	
CAA	C	
TAGCACTGAC	T	
G	GA	
TGA	T	
ATATCT	A	
C	CT	
G	GT	

A red circle highlights the bottom-right corner of the browser window, where a small arrow icon points towards the right, indicating the direction to click to bring the tools pane back into view.

## 7.10 Combine SNPs and INDELS

- Load *combine variants* tool from the tool pane: "NGS: GATK Tools -> Combine Variants"
- Input variant file: -> "..SNPs ..." (recalibrated SNP vcf)
- Variant name: -> "snps"
- Click "Add new Variants to Merge"
- Select the INDEL file (Input variant file: -> "..INDELS ..." (recibrated INDEL vcf))
- Variant name: -> "indels"
- Using reference genome: -> Homo sapiens hg19\_canonical (GATK)
- Click "Execute"

The screenshots illustrate the steps to combine SNPs and INDELS using the Galaxy platform:

- Screenshot 1 (Step a):** Shows the Galaxy tool pane with the 'Combine Variants' tool highlighted and circled in red.
- Screenshot 2 (Step b):** The 'Combine Variants' tool configuration screen. The 'Input variant file' field contains '37: SNPs' and the 'Variant name' field contains 'snps'. A red circle highlights the 'Input variant file' field, another highlights the 'Variant name' field, and a third highlights the 'Add new Variants to Merge' button.
- Screenshot 3 (Step c):** The 'Variants to Merge 2' configuration screen. The 'Input variant file' field contains '46: INDELS' and the 'Variant name' field contains 'indels'. A red circle highlights the 'Input variant file' field, another highlights the 'Variant name' field, and a third highlights the 'Using reference genome' dropdown set to 'Homo sapiens hg19\_canonical (GATK)'. A fourth red circle highlights the 'Execute' button at the bottom.

# Variant Annotation

## 7.11 Annotate variants using SnpEff

- a) Load *variant annotation* tool from the tool pane: "SnpEff tools -> SnpEff Variant effect and annotation"
- b) Sequence changes (SNPs, MNPs, InDels): -> "Combine Variants.... (variants)"
- c) Genome: -> hg19
- d) Upstream / Downstream length: -> 200 bases

The screenshot shows the Galaxy interface with the title "Galaxy / UMN". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". The left sidebar under "Tools" lists various bioinformatics tools, with "SnpEff tools" highlighted by a red box and labeled 'a'. A red circle highlights the "SnpEff Variant effect and annotation" tool. The main content area displays a green success message: "The following job has been successfully added to the queue: 55: Combine Variants on data 37 and data 46 (variants)". Below this, it says "56: Combine Variants on data 37 and data 46 (log)". A note states: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered." To the right, the "History" pane shows a list of jobs, with the most recent ones being "55: Combine Variants on data 37 and data 46 (variants)" and "54: Apply Variant Recalibration on data 49, data 48, and data 46 (log)".

This screenshot shows the configuration of the "SnpEff Variant effect and annotation" tool. The "Sequence changes (SNPs, MNPs, InDels)" input field is circled in red and labeled 'b'. The "Genome" dropdown is set to "Homo sapiens : hg19" and is circled in red, labeled 'c'. The "Upstream / Downstream length:" dropdown is set to "200 bases" and is circled in red, labeled 'd'. Other visible parameters include "Input format: VCF", "Output format: VCF (only if input is VCF)", and "Set size for splice sites (donor and acceptor) in bases. Default: 2: 1 base". The right side of the screen shows the same history of jobs as the previous screenshot.

e) Click "Execute"

The screenshot shows the Galaxy / UMN web interface. On the left, a sidebar lists various tools under categories like NGS, SNP/WGA, SnpEff, Phenotype Association, VCF Tools, IGV Tools, MSI, and Masonic Cancer Center Tools. The main workspace displays the SnpEff tool configuration. It includes sections for 'Filter output' (checkboxes for downstream, intergenic, intron, upstream, and 5'/3' UTR changes), 'Chromosomal position' (radio buttons for default, zero-based, or one-based positions), and 'Text to prepend to chromosome name' (a text input field). Below these are options for 'Produce Summary Stats' (checked) and 'Do not report usage statistics to server' (checked). A large red circle highlights the 'Execute' button at the bottom of the configuration panel. To the right, a 'History' panel shows a list of previous workflows, such as 'Variant\_Detection\_RISS' and several 'Combine Variants' and 'Apply Variant Recalibration' steps. The bottom of the workspace has a note: 'This tool calculate the effect of variants (SNPs/MNPs/Insertions) and deletions.'

## 7.12 Review Annotated Variants

- In the history pane click the eye icon next to the VCF file produced by the “SnpEff tools -> SnpEff Variant effect and annotation” tool
- Notice the **INFO** is now appended with annotation information such as effect of the variant e.g., NON\_SYNOMYMOUS CODING

The screenshot shows the Galaxy / UMN web interface. On the left, there's a sidebar with a 'Tools' section containing various bioinformatics tools. The main area shows a command-line log with some code and annotations. A red arrow labeled 'b' points to the word 'NON\_SYNOMYMOUS\_CODING' in the log. To the right, the 'History' pane lists several analysis steps. One step, '58: SnpEff on data 55', has its eye icon circled in red and labeled 'a'. Below it, another step, '56: Combine Variants on data 37 and data 46 (log)', is also listed.

```

Galaxy / UMN
Analyze Data Workflow Shared Data+ Visualization+ Help+ User+ Using 1.5 GB

Tools
NGS: Peak Calling
NGS: Simulation
SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models
SnpEff_tools
  SnpSift_Filter variants using arbitrary expressions
  SnpEff_Download Download a new database
  SnpEff_Variant effect and annotation
  SnpSift_CaseControl Count samples are in 'case' and 'control' groups.
  SnpSift_Interval Filter variants using intervals
  SnpSift_Annotate Annotate SNPs from dbSnp
Phenotype Association
VCF Tools

History
Variant_Detection_RISS 784.8 MB
58: SnpEff on data 55 437 lines, 58 comments
format: vcf, database: hg19_canonical
display with IGV web current local
1. Chrom 2. Pos
##fileformat=VCFv4.1
##CombineVariants="analysis_type=CombineVariants canonical/seq/hg19_canonical.fa rodBind=□ nones=-1 validation_strictness=SILENT unsafe=null logging_level=INFO log_to_file=null help=false s.vcf] out=org.broadinstitute.sting.gatk.io.s

56: Combine Variants on data 37 and data 46 (log)

```