

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 take, 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 tex-based format for storing a biological sequence and its corresponding quality scores. See 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.

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).
- 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

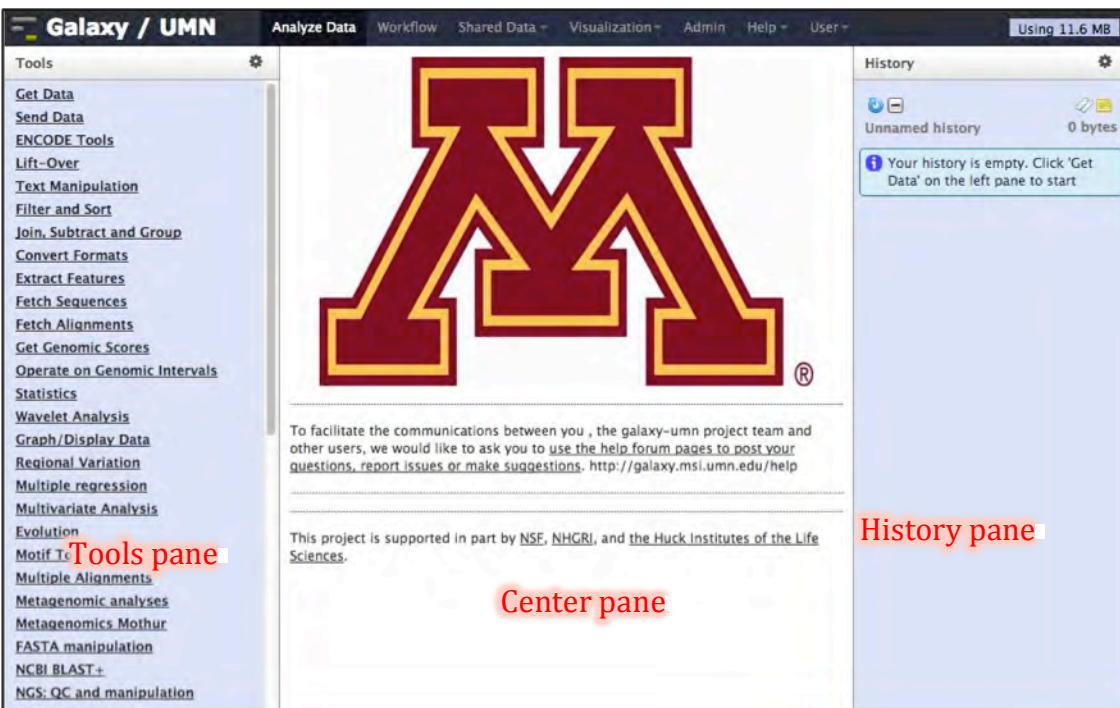
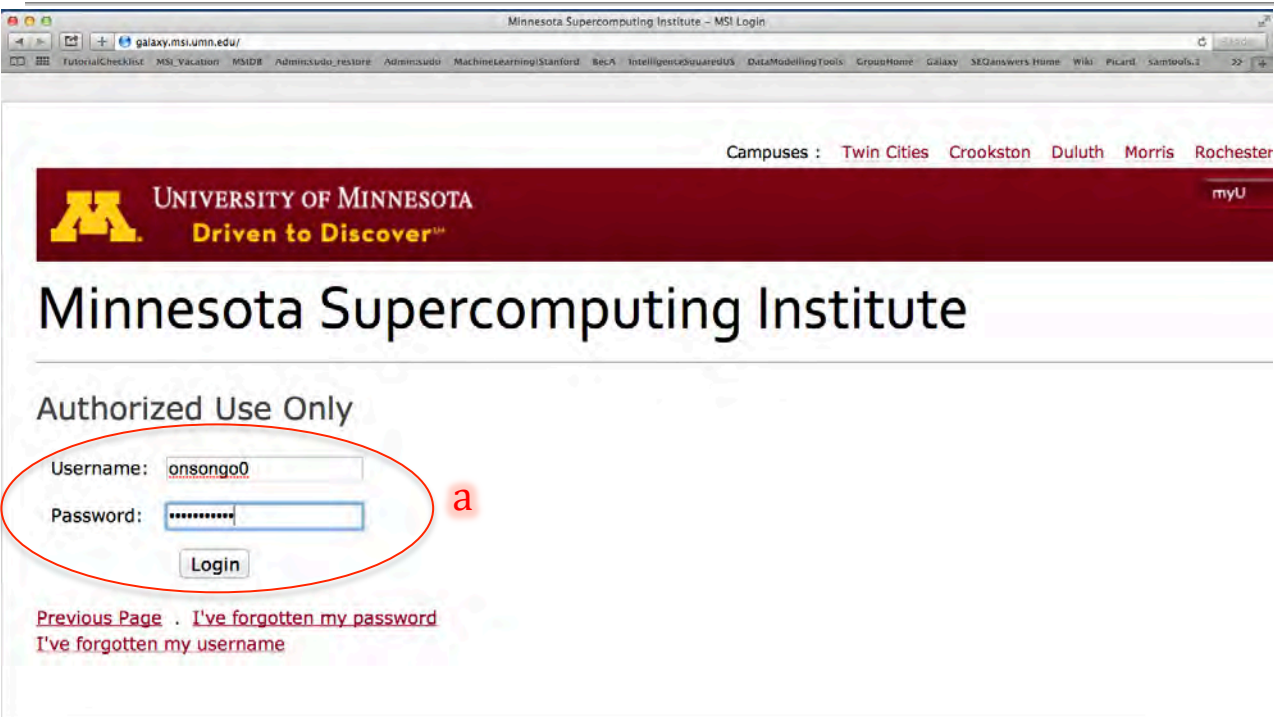
1.2 Outline of tutorial

- 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

- a) Log in with your username and password



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

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

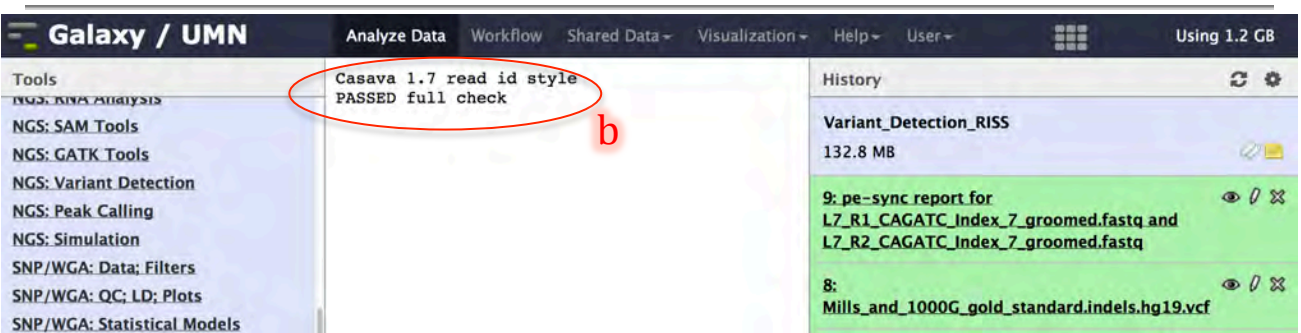
- Load the *pe-sync* tool from the tool pane: “MSI -> *pe-sync*: Paired-end synchronization check”
- Set input files: From the dropdown menu under
Input 1: select “L7_R1_CAGATC_Index_7_groomed.fastq”
Input 2: select “L7_R2_CAGATC_Index_7_groomed.fastq”
- Click “Execute”

The screenshot shows the Galaxy/UMN interface. On the left, the 'Tools' pane is visible, with the 'MSI' category selected and the 'pe-sync: Paired-end synchronization check' tool highlighted by a red box labeled 'a'. The main workspace displays a large 'M' logo and the text 'If you have questions or concerns, please e-mail help@msi.umn.edu.' and 'MSI Galaxy Updates'.

The screenshot shows the Galaxy/UMN interface with the 'pe-sync: Paired-end synchronization check (version 0.1.0)' tool selected. The 'Input 1' dropdown is set to 'L7_R1_CAGATC_Index_7_groomed.fastq' and the 'Input 2' dropdown is set to 'L7_R2_CAGATC_Index_7_groomed.fastq', both highlighted by a red box labeled 'b'. The 'Execute' button is circled in red and labeled 'c'. The right-hand 'History' pane shows a list of previous jobs, including 'Variant_Detection_RISS' and 'Mills_and_1000G_gold_standard.indels.hg19.vcf'.

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



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/BAM
- 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.f...

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

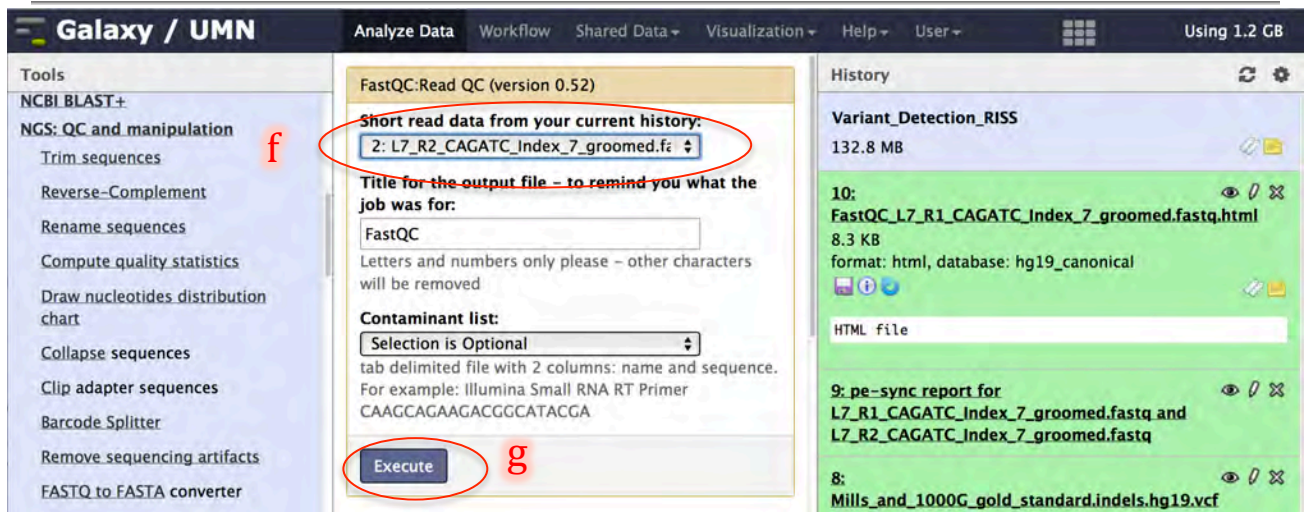
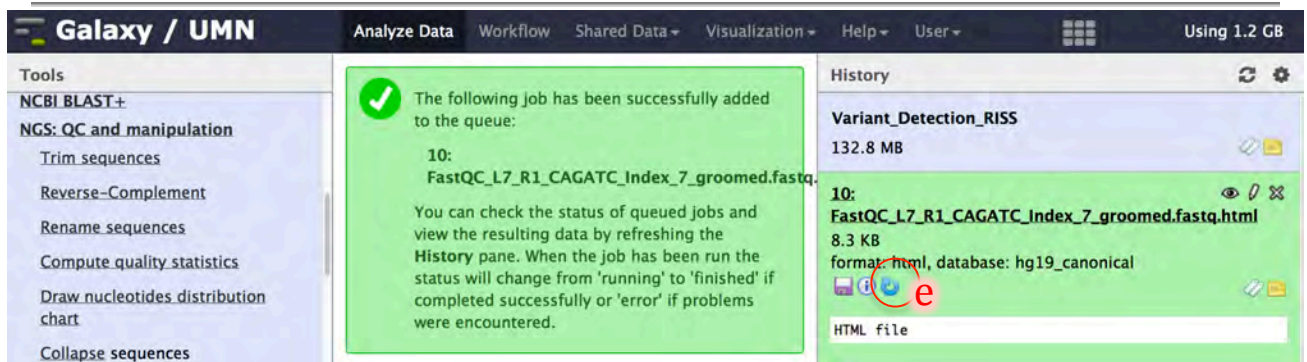
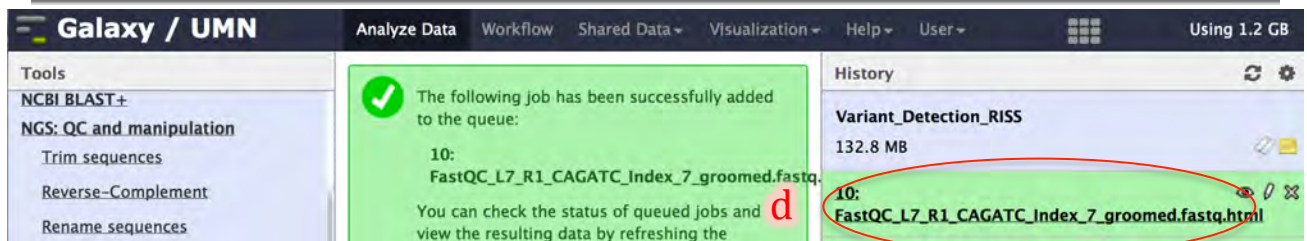
tab delimited file with 2 columns: name and sequence.
For example: Illumina Small RNA RT Primer
CAAGCACAAGACGGCATACGA

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 expand its box and show more options
- 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”



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>.

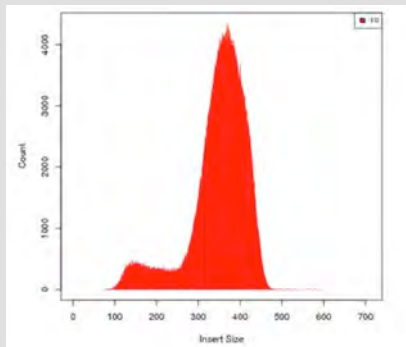
Other aligners besides BWA can be used, provided that the BAM files satisfy the GATK's formatting requirements (see the GATK's website for more details). This section illustrates how to specify the required fields to ensure that your BAM files adhere to the GATK's formatting requirements.

★ 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

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

L7_R2_CAGATC_Index_7_groomed.fastq FastQC Report

FastQC Report

Sat 16 Nov 2013

L7_R2_CAGATC_Index_7_groomed.fastq

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores

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

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 PerM for SOLiD and

Map with BWA for Illumina (version 1.2.3)

Will you select a reference genome from your history or use a built-in index?:

Use a built-in index

Select a reference genome:

hg19_canonical

Is this library mate-paired?:

Paired-end

Forward FASTQ file:

1: L7_R1_CAGATC_Index_7_groomed.f...

FASTQ with either Sanger-scaled quality values (fastqsanger) or illumina-scaled quality values (fastqillumina)

Reverse FASTQ file:

2: L7_R2_CAGATC_Index_7_groomed.f...

FASTQ with either Sanger-scaled quality values (fastqsanger) or illumina-scaled quality values (fastqillumina)

BWA settings to use:

Full Parameter List

For most mapping needs use Commonly Used settings. If you want full control use Full Parameter List

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.snps.high_confidence.hg19.vcf
- 4: 1000G_omni2.5.hg19.vcf
- 3: tutorial_exons.bed

- g) Maximum insert size for a read pair 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

The screenshot displays the Galaxy/UMN interface for configuring the 'NGS: Mapping' tool. The configuration parameters are as follows:

- Maximum insert size for a read pair to be considered as being mapped properly (sampe -a):** 1000 (circled in red, labeled 'g')
- Maximum occurrences of a read for pairing (sampe -o):** 100000
- Specify the read group for this file? (samse/sampe -r):** Yes (circled in red, labeled 'h')
- Read group identifier (ID):** NA_10858_400 (circled in red, labeled 'i')
- Sequencing center that produced the read (CN):** UMGC (circled in red, labeled 'j')

The right-hand panel shows the 'History' section with a list of files, including 'Variant_Detection_RISS' (133.5 MB) and several FASTQ and VCF files.

- 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 displays the Galaxy/UMN interface for configuring a tool. The 'Tools' panel on the left lists various NGS tools. The main configuration area shows the following fields:

- Description (DS):** Coriell_HapMap_400bpIn (marked with a red 'k')
- Date that run was produced (DT):** 2013-11-20 (marked with a red 'l')
- Library name (LB):** NA_10858 (marked with a red 'm')

The 'History' panel on the right shows a list of files generated by the tool, including:

- 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.snps.high_confidence.hg19.vcf
- 4: 1000G_omni2.5.hg19.vcf
- 3: tutorial_exons.bed
- 2: L7_R2_CAGATC_Index_7_groomed.fastq

- 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 displays the Galaxy / UMN interface for the 'Variant Detection - RISS' tool. The configuration panel on the left has several fields circled in red and labeled with letters:

- n**: Predicted median insert size (PI) is set to 400.
- o**: Platform/technology used to produce the reads (PL) is set to ILLUMINA.
- p**: Platform unit (PU) is set to HWUSI-EAS1737:7.
- q**: Sample (SM) is set to NA_10858.
- r**: The 'Execute' button is highlighted.

The 'History' panel on the right shows a list of files generated by the tool, including:

- 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.snps.high_confidence.hg19.vcf
- 4: 1000G_omni2.5.hg19.vcf
- 3: tutorial_exons.bed
- 2: L7_R2_CAGATC_Index_7_groomed.fastq

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 additional 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

Reorder SAM/BAM

Replace SAM/BAM Header

Paired Read Mate Fixer for paired data

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

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 UMN interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'Tools' panel on the left lists various tools like 'NCBI BLAST+', 'NGS: QC and manipulation', and 'NGS: Picard (beta)'. The main workspace displays a green notification box: 'The following job has been successfully added to the queue: 13: InsertSize_Insertion size metrics.html'. Below the notification, it says: '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 'History' panel on the right shows a list of jobs, with job 13 highlighted and its 'eye' icon circled in red, labeled with a red 'a'.

The screenshot shows the Galaxy UMN interface displaying the results of a tool run. The main workspace shows a plot titled 'Insert Size Histogram for All Reads in file dataset_362380.dat'. The plot has 'Count' on the y-axis (0 to 3000) and 'Insert Size' on the x-axis (0 to 600). A red histogram shows a sharp peak around 350-400. A context menu is open over the plot, with 'Save Image to the Desktop' circled in red and labeled with a red 'b'. Below the plot, it says: 'The following output files were created (click the filename to view/download a copy):' followed by two links: 'CollectInsertSizeMetrics.log' and 'CollectInsertSizeMetrics.metrics.txt'. The 'History' panel on the right shows job 13 as 'finished' with its 'eye' icon visible.

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. Accomodation 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 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”

Galaxy / UMN

Tools	chr1	2341670	2341920	0.0	0.0	0.0	0.0
NGS: Mapping	chr1	2342039	2342342	0.0	0.0	0.0	0.0
NGS: Indel Analysis	chr1	2343800	2344040	0.0	0.0	0.0	0.0
NGS: RNA Analysis	chr1	2345006	2345266	0.0	0.0	0.0	0.0
NGS: SAM Tools	chr1	6484986	6485339	0.0	0.0	0.0	0.0
NGS: GATK Tools	chr1	6488256	6488509	1.0	0.913043478261	0.584980237154	0.446640

History

- History Lists
 - Saved Histories
 - Histories Shared with Me
- Current History
 - Create New

Galaxy / UMN

Saved Histories

search history names and tags

Advanced Search

Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated	Status
imported: BamExonCoverage1X10X20X30X	15	0 Tags		724.6 MB	~ 1 hour ago	~ 1 hour ago	current history
imported: TutorialPercOffFragOnTarget	13	0 Tags		648.1 MB	~ 1 hour ago	~ 1 hour ago	
Variant_Detection_RISS	13	0 Tags		387.5 MB	2 days ago	~ 2 hours ago	

For 0 selected: View, Delete, Delete Permanently, Undelete, Share or Publish

History

- imported: BamExonCoverage1X10X20X30X (724.6 MB)
- 15: Coverage1X10X20X30X
- 14: Coverage on data 13 and data 3
- 13: Filter on data 11
- 12: Coverage on data

Galaxy / UMN

Saved Histories

search history names and tags

Advanced Search

Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated	Status
imported: BamExonCoverage1X10X20X30X	15	0 Tags		724.6 MB	~ 1 hour ago	~ 1 hour ago	
imported: TutorialPercOffFragOnTarget	13	0 Tags		648.1 MB	~ 1 hour ago	~ 1 hour ago	
Variant_Detection_RISS	13	0 Tags		387.5 MB	2 days ago	~ 2 hours ago	

For 0 selected histories: Rename, Delete, Delete Permanently, Undelete

Histories that have been deleted for more than a time period specified by the Galaxy administrator(s) may be permanently deleted.

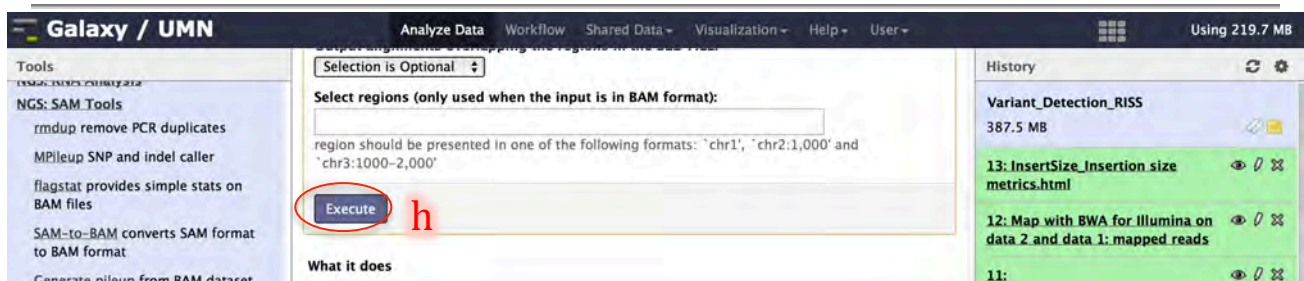
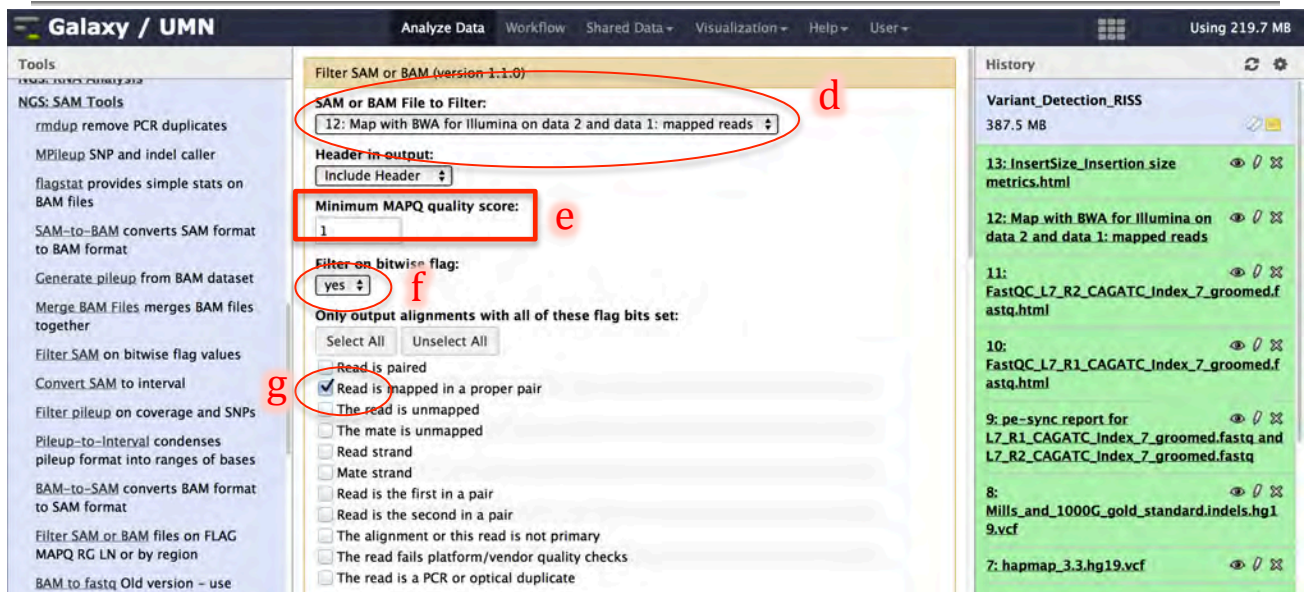
History

- 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.fastq and L7_R2_CAGATC_Index_7_groomed.fastq
- 8: Mills_and_1000G_gold_standard.indels.hg19.vcf

Tools

- NGS: SAM Tools
 - rmDup_remove PCR duplicates
 - Pileup 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 SAM to Fastq tool in wizard section

- 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”



5.2 Sorting Reads and updating mate-pair information

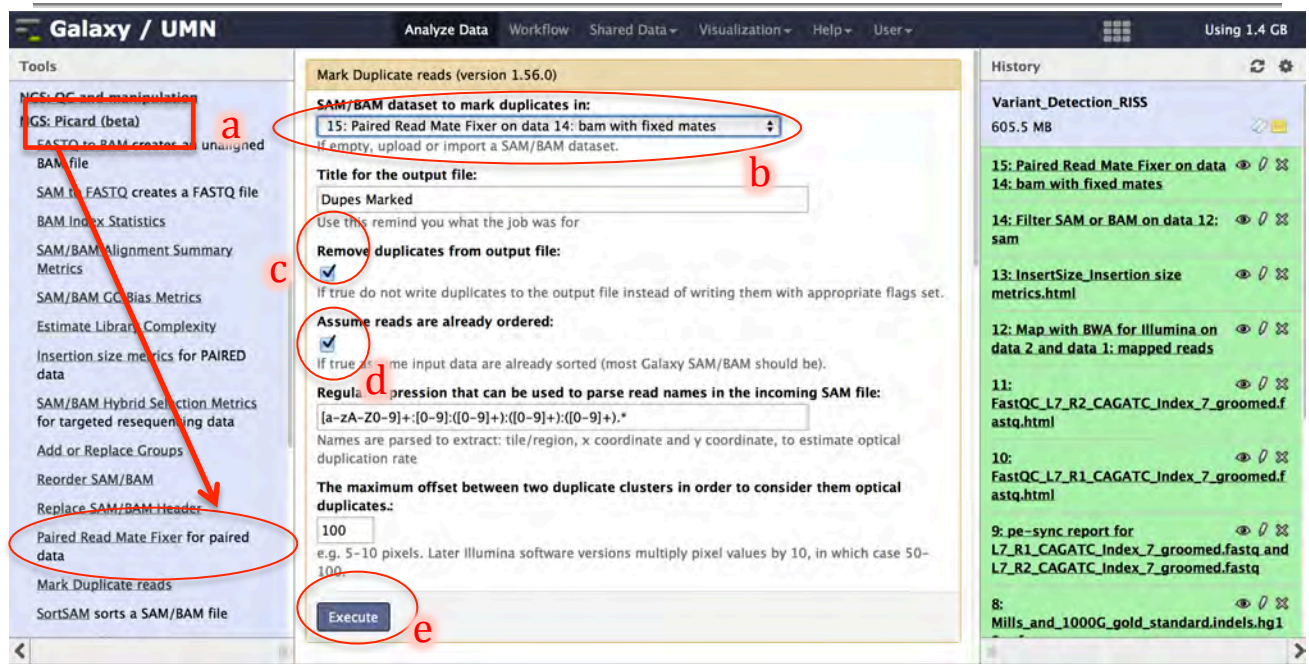
- Load *paired-read mate fixer* tool from the tool pane: “NGS: Picard (beta) -> Paired Read Mate Fixer for paired data”
- SAM/BAM dataset to fix: -> “...Filter SAM or BAM ...”
- Sort order: -> Coordinate sort
- Output BAM instead of SAM: -> check (✓)
- Click “Execute”

The screenshot shows the Galaxy/UMN interface. In the left-hand 'Tools' pane, the 'Paired Read Mate Fixer for paired data' tool is circled in red and labeled with a red 'a'. A red arrow points from this tool to the center of the interface. In the center, a green success message box states: 'The following job has been successfully added to the queue: 14: Filter SAM or BAM on data 12: sam'. The right-hand 'History' pane shows a list of jobs, with job 14 'Filter SAM or BAM on data 12: sam' at the top.

The screenshot shows the configuration interface for the 'Paired Read-Mate Fixer (version 1.56.0)' tool. The 'SAM/BAM dataset to fix' dropdown is set to '14: Filter SAM or BAM on data 12: sam' and is circled in red with a red 'b'. The 'Sort order' dropdown is set to 'Coordinate sort' and is circled in red with a red 'c'. The 'Output BAM instead of SAM' checkbox is checked and circled in red with a red 'd'. The 'Execute' button is circled in red with a red 'e'. The right-hand 'History' pane shows the same list of jobs as in the previous screenshot.

5.3 Removal of duplicates

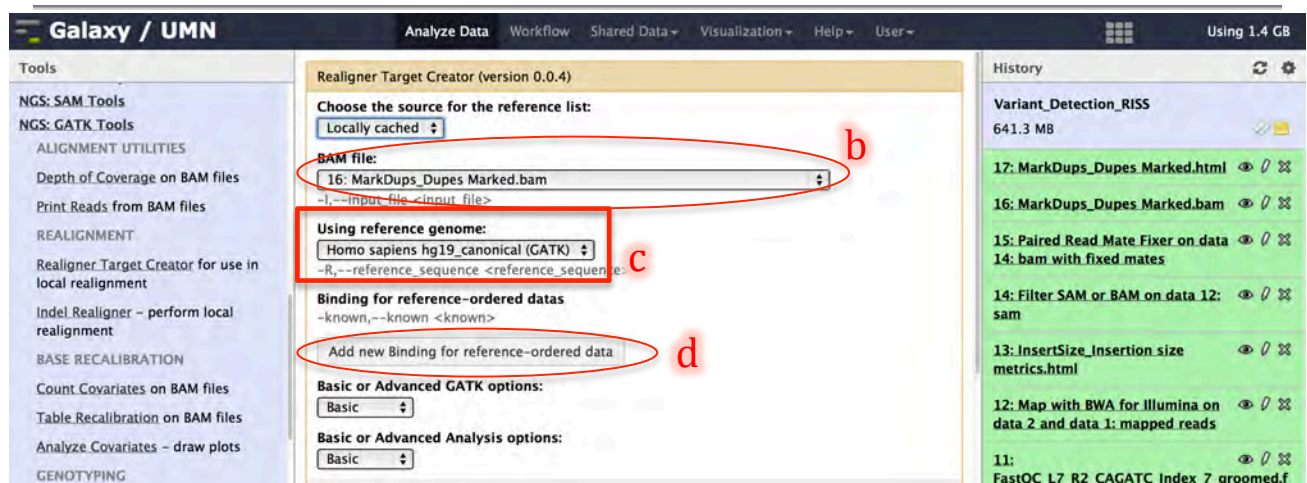
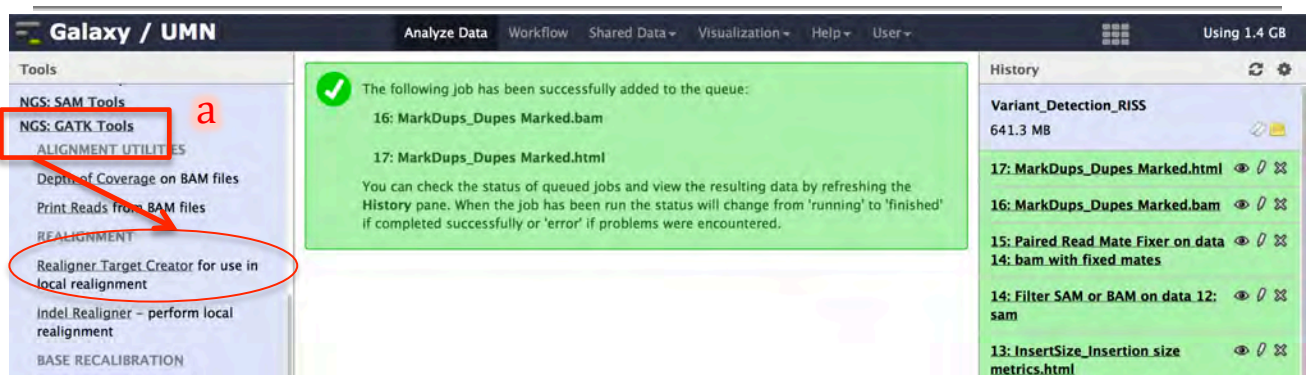
- Load *mark duplicates* tool from the tool pane: “NGS: Picard (beta) -> Mark Duplicate reads”
- SAM/BAM dataset to mark duplicates in: -> “.....Paired Read Mate Fixer on data.....”
- Remove duplicates from output file: -> check (✓)
- Assume reads are already ordered: -> check (✓)
- Click “Execute”



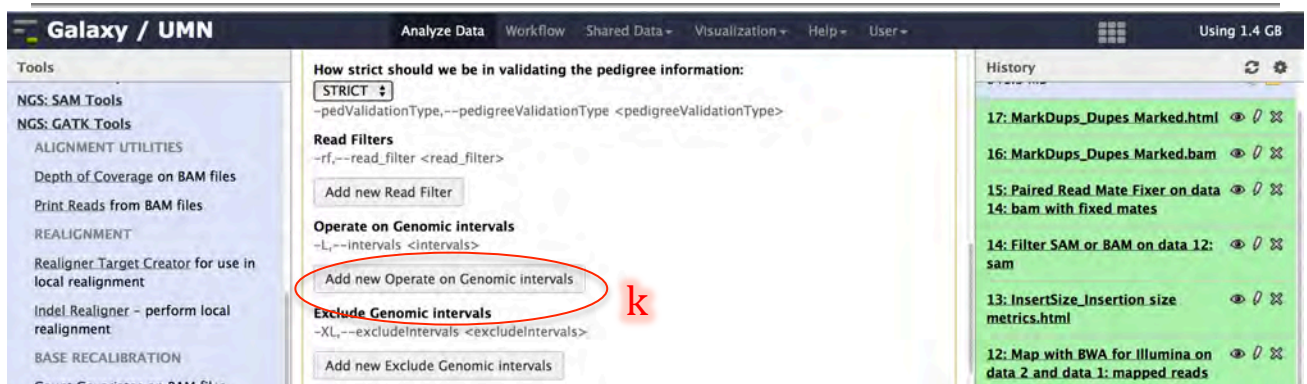
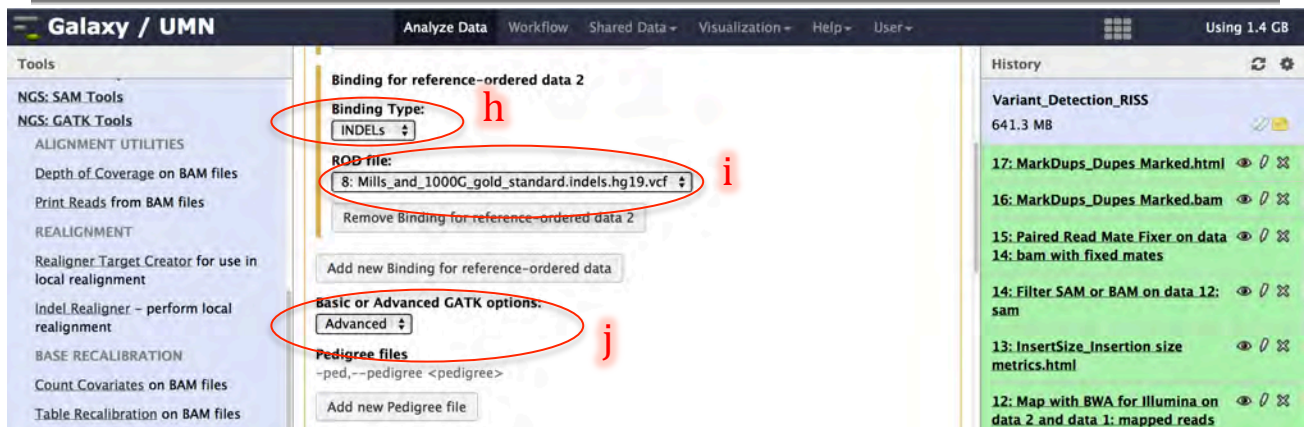
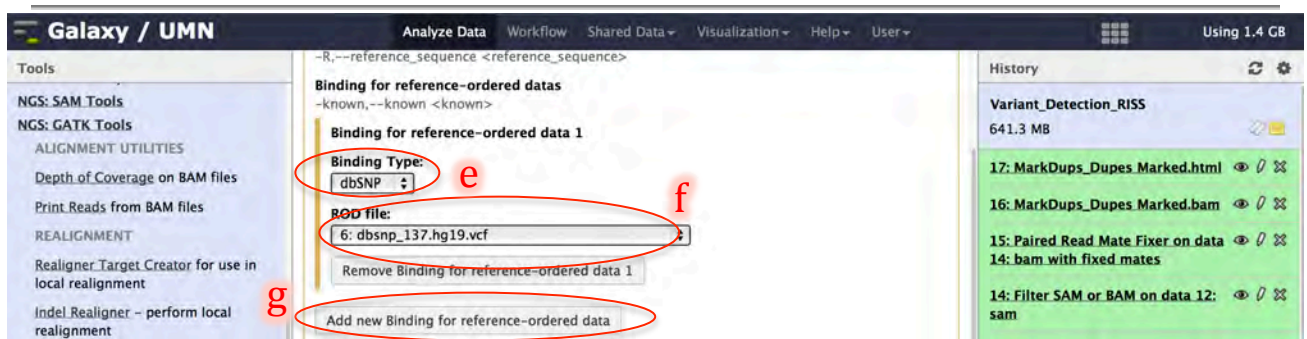
INDEL Realignment

5.4 Create Targets for Realignment

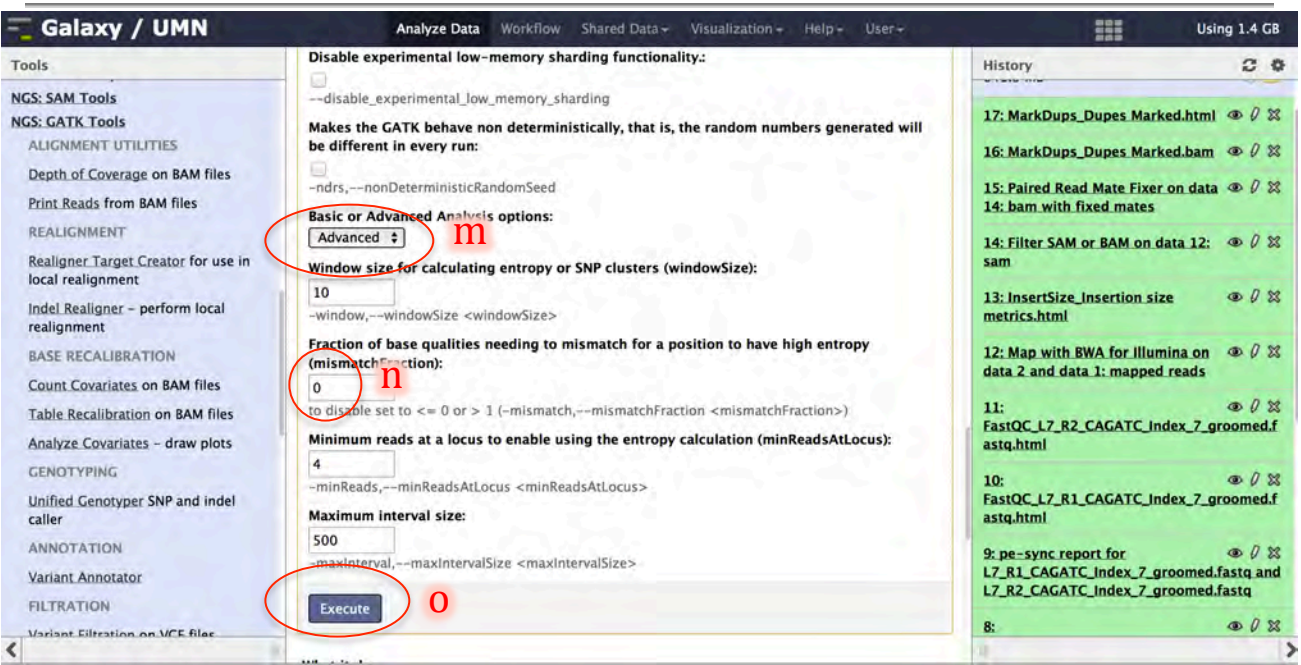
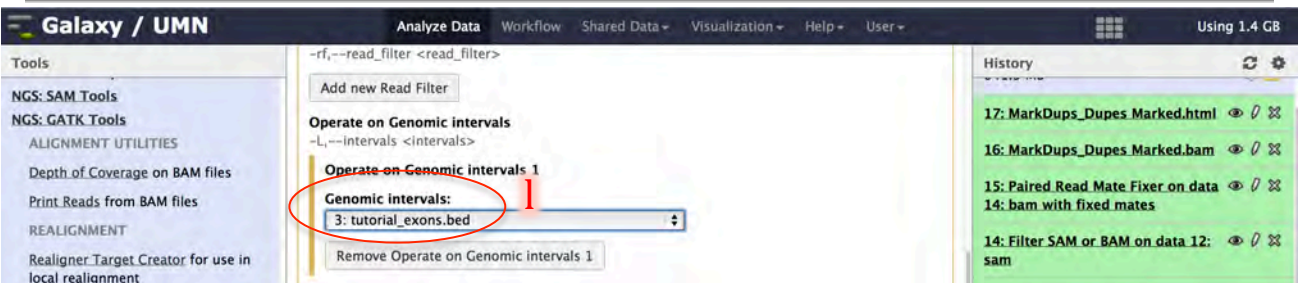
- a) Load *realigner target creator* tool from the tool pane: “NGS: GATK Tools -> Realigner Target Creator for use in local realignment”
- b) BAM file: -> “...MarkDups_Dupes Marked.bam”
- c) Using reference genome: -> Homo sapiens hg19_canonical (GATK)
- d) Click the “Add new Binding for reference-ordered data” button



- 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



- l) 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”



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”

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.4 GB

Tools

NGS: SAM Tools

NGS: GATK Tools **a**

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

The following job has been successfully added to the queue:

18: Realigner Target Creator on data 6, data 16, and others (GATK intervals)

19: Realigner Target Creator on data 6, data 16, 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.

History

Variant_Detection_RISS 668.9 MB

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

15: Paired Read Mate Fixer on data

14: bam with fixed mates

Galaxy / UMN

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 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

Locally cached

BAM file: **b**
16: MarkDups_Dupes Marked.bam

Using reference genome: **c**
Homo sapiens hg19_canonical (GATK)

Restrict realignment to provided intervals: **d**
18: Realigner Target Creator on data 6, data 16, and others (GATK interval)

Binding for reference-ordered data

LOD threshold above which the realigner will proceed to realign:
5.0

Use only known indels provided as RODs:

Basic or Advanced GATK options:
Basic

Basic or Advanced Analysis options:
Basic

e Execute

History

Variant_Detection_RISS 668.9 MB

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

15: Paired Read Mate Fixer on data

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

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”

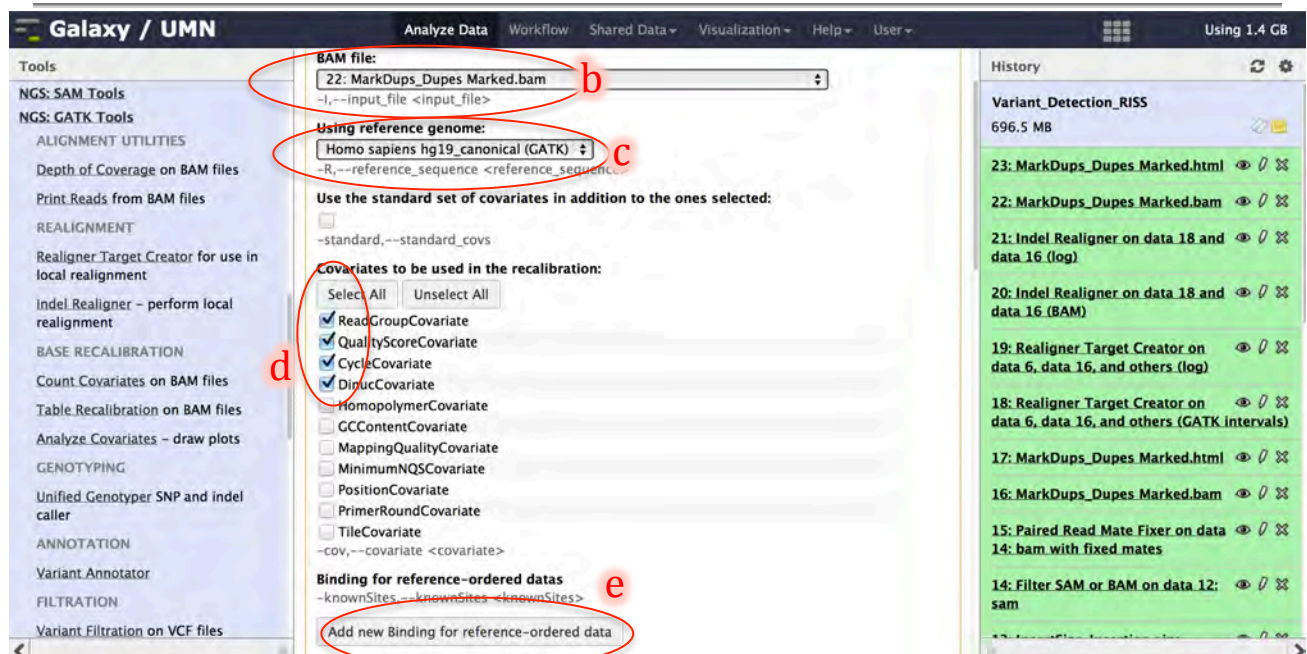
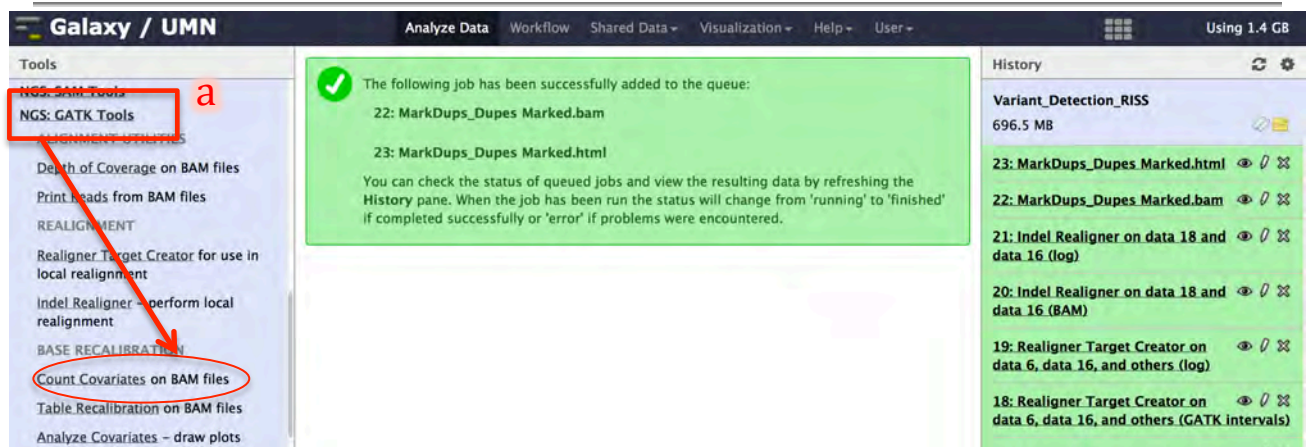
The screenshot shows the Galaxy/UMN interface. On the left, the 'Tools' pane is open, and 'Mark Duplicate reads' is selected under the 'NGS: Picard (beta)' category. A red box highlights this tool, with a red arrow pointing to it from the letter 'a'. In the center, a green notification box states: 'The following job has been successfully added to the queue: 20: Indel Realigner on data 18 and data 16 (BAM)'. On the right, the 'History' pane shows a list of jobs, including '21: Indel Realigner on data 18 and data 16 (log)' and '20: Indel Realigner on data 18 and data 16 (BAM)'.

The screenshot shows the configuration page for the 'Mark Duplicate reads (version 1.56.0)' tool. The 'SAM/BAM dataset to mark duplicates in:' field is set to '20: Indel Realigner on data 18 and data 16 (BAM)', with a red box and the letter 'b' around it. The 'Remove duplicates from output file:' checkbox is checked, with a red box and the letter 'c' around it. The 'Assume reads are already ordered:' checkbox is also checked, with a red box and the letter 'd' around it. The 'Execute' button is highlighted with a red box and the letter 'e'. The 'Title for the output file:' field is set to 'Dups Marked'. The 'Regular expression that can be used to parse read names in the incoming SAM file:' field contains the expression '[a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+).*'. The 'The maximum offset between two duplicate clusters in order to consider them optical duplicates:' field is set to '100'.

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



- f) Binding Type: -> dbSNP
- g) ROD file: -> dbsnp_137.hg19.vcf
- h) Click "Execute"

The screenshot displays the Galaxy/UMN interface for configuring a tool. 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 shows the configuration for a tool with the following settings:

- Binding Type:** dbSNP (labeled 'f')
- ROD file:** 6: dbsnp_137.hg19.vcf (labeled 'g')
- Basic or Advanced GATK options:** Basic
- Basic or Advanced Analysis options:** Basic
- Execute** button (labeled 'h')

A warning message is displayed at the bottom: "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." However, if you do not provide this file, the '--run_without_dbsnp_potentially_ruining_quality' flag will be automatically used, and the command will be allowed to run.

The right sidebar shows the 'History' section with a list of jobs, including 'Variant_Detection_RISS' (696.5 MB) and several 'MarkDups_Dupes Marked' jobs.

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”

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.4 GB

Tools

NGS: SAM Tools

NGS: GATK Tools **a**

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

✓ The following job has been successfully added to the queue:

24: Count Covariates on data 6 and data 22 (Covariate File)

25: Count Covariates on data 6 and data 22 (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
724.1 MB

25: Count Covariates on data 6 and data 22 (log)

24: Count Covariates on data 6 and data 22 (Covariate File)

23: MarkDups_Dupes_Marked.html

22: MarkDups_Dupes_Marked.bam

21: Indel Realigner on data 18 and data 16 (log)

20: Indel Realigner on data 18 and data 16 (BAM)

19: Realigner.Target.Creator on data 6, data 16, and others (log)

Galaxy / UMN

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

Analyze Covariates (version 0.0.5)

Covariates table recalibration file:

24: Count Covariates on data 6 and data 22 (Covariate File) **b**

-recal file --recal file <recal file>

Basic or Advanced options:

Basic

Execute **c**

History

Variant_Detection_RISS
724.1 MB

25: Count Covariates on data 6 and data 22 (log)

24: Count Covariates on data 6 and data 22 (Covariate File)

23: MarkDups_Dupes_Marked.html

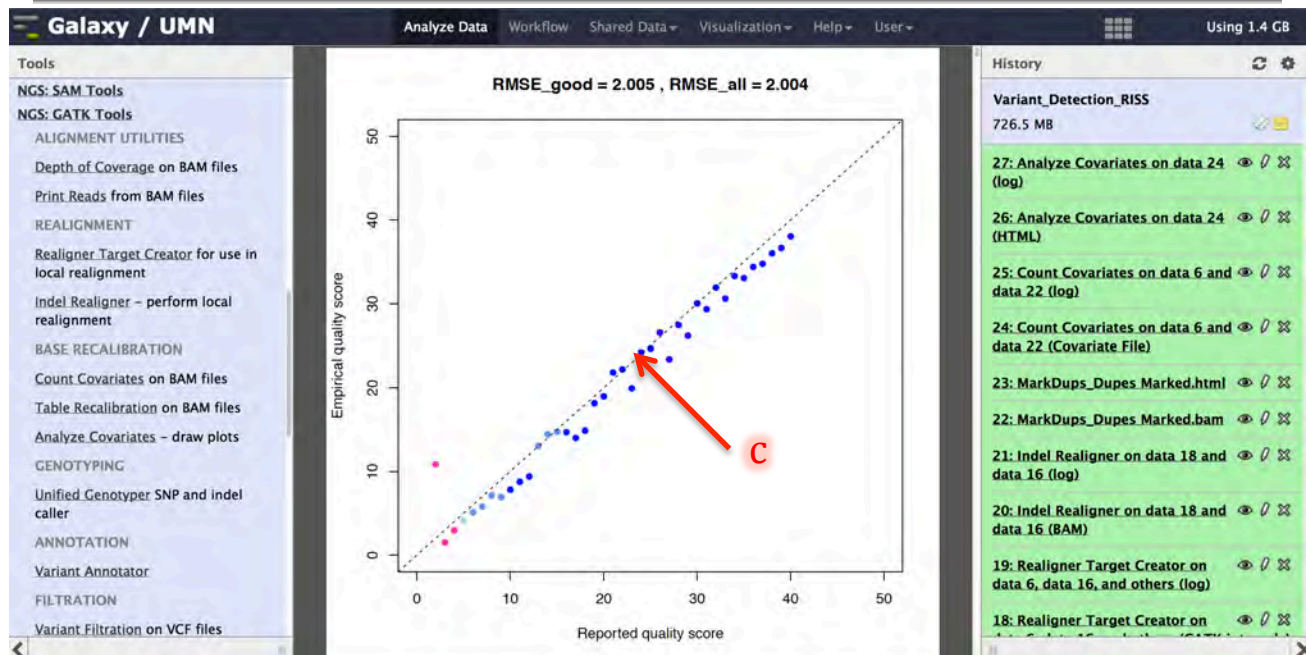
5.9 Review Covariate plots (before base quality recalibration)

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

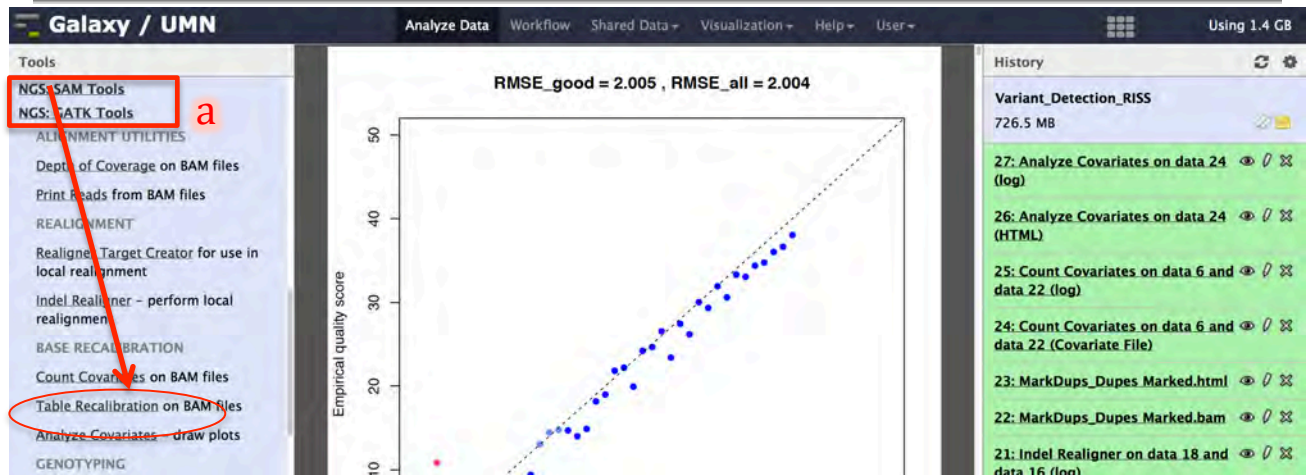
The screenshot shows the Galaxy / UMN interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'Tools' panel on the left lists various NGS tools. The main area displays a green notification box: 'The following job has been successfully added to the queue: 26: Analyze Covariates on data 24 (HTML)'. Below this, it says '27: Analyze Covariates on data 24 (log)'. A 'View data' button is visible next to job 26 in the History pane on the right, which is circled in red and labeled 'a'.

The screenshot shows the Galaxy / UMN interface with the 'GATK Output' pane. It lists several files, including 'NA_10858_400.QualityScoreCovariate.dat.quality_emp_v_stated.pdf', which is highlighted with a red box and labeled 'b'. The History pane on the right shows the job '26: Analyze Covariates on data 24 (HTML)' selected.



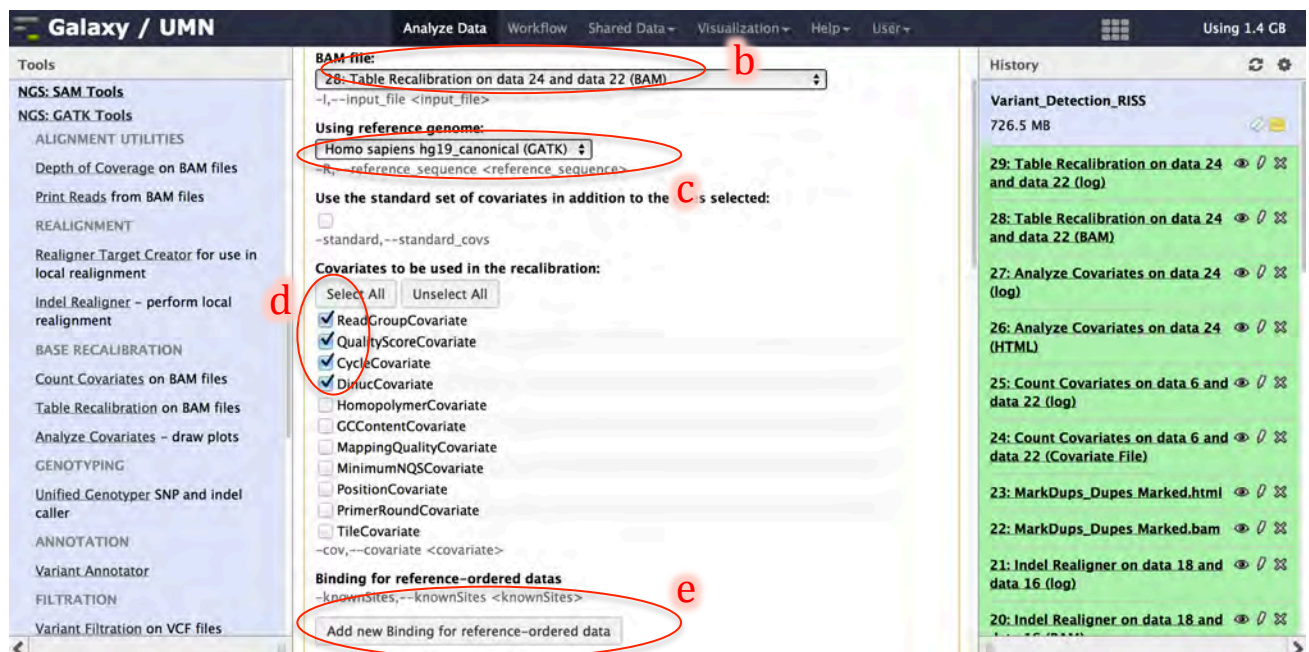
5.10 Recalibrate base quality scores

- 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”



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



- f) Binding Type: -> dbSNP
- g) ROD file: -> dbsnp_137.hg19.vcf
- h) Click "Execute"

The screenshot displays the Galaxy/UMN interface for configuring a tool. The central panel is titled "Binding for reference-ordered data 1" and contains the following elements:

- Binding Type:** A dropdown menu set to "dbSNP", circled in red and labeled with a red "f".
- ROD file:** A text input field containing "6: dbsnp_137.hg19.vcf", circled in red and labeled with a red "g".
- Execute button:** A blue button labeled "Execute", circled in red and labeled with a red "h".

The left sidebar lists various tool categories under "NGS: GATK Tools", including "ALIGNMENT UTILITIES", "REALIGNMENT", "BASE RECALIBRATION", and "GENOTYPING". The right sidebar shows a "History" panel with a list of jobs, including "Variant_Detection_RISS" and several "Table Recalibration" and "Analyze Covariates" jobs.

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”

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.4 GB

Tools

NGS: SAM Tools

NGS: GATK Tools **a**

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

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)

Galaxy / UMN

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 RECALIBRATION

Count Covariates on BAM files

Table Recalibration on BAM files

Analyze Covariates - draw plots

GENOTYPING

Analyze Covariates (version 0.0.5)

Covariates table recalibration file: **b**

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

recalFile,--recal_file <recal_file>

Basic or Advanced options:

Basic

Execute **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)

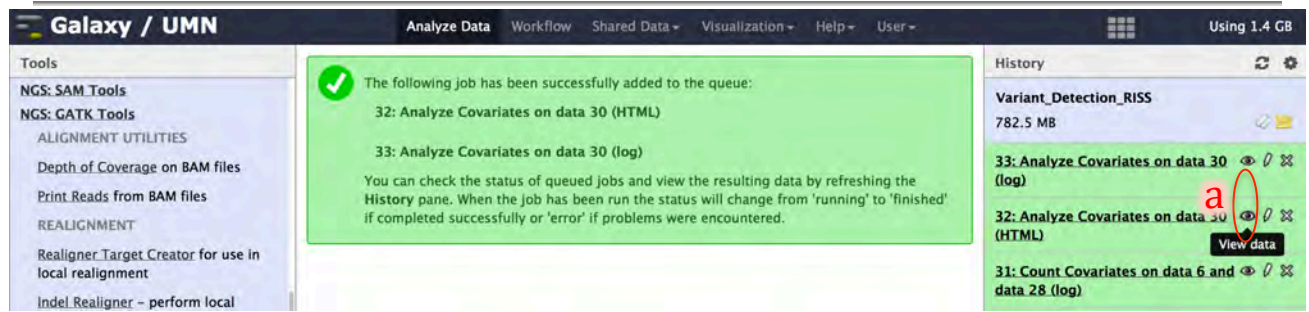
28: Table Recalibration on data 24 and data 22 (BAM)

27: Analyze Covariates on data 24 (log)

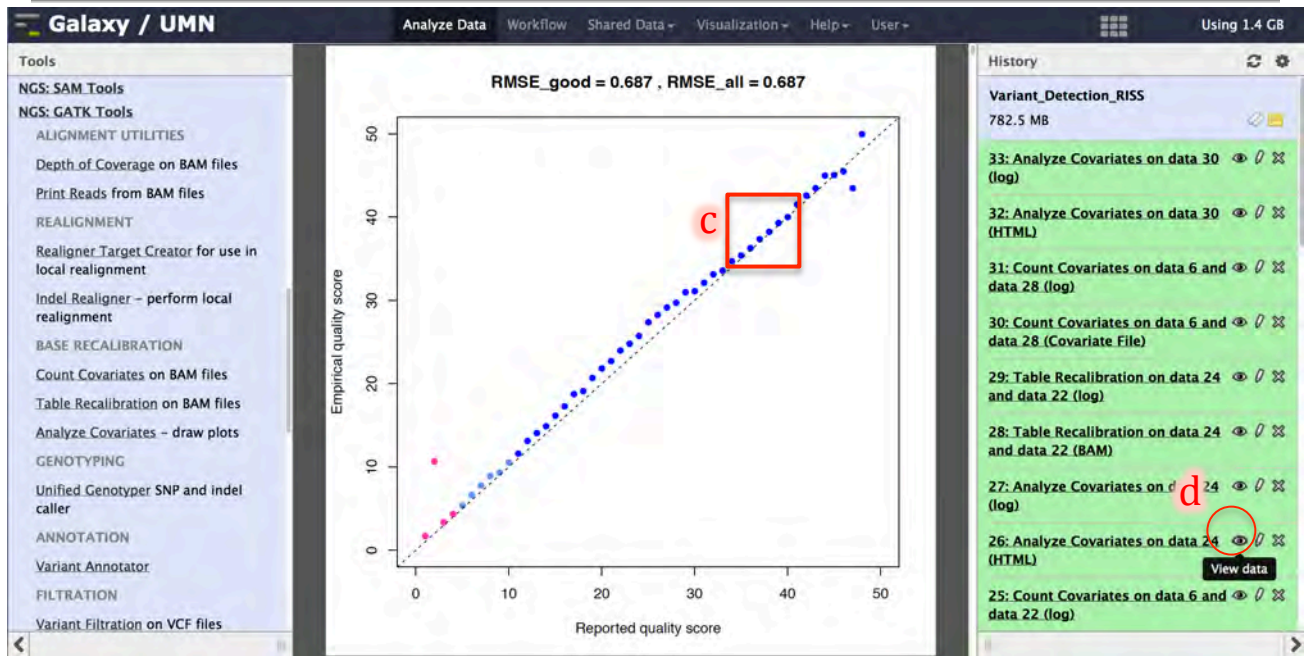
26: Analyze Covariates on data 24 (HTML)

5.13 Review Covariate plots (after base quality recalibration)

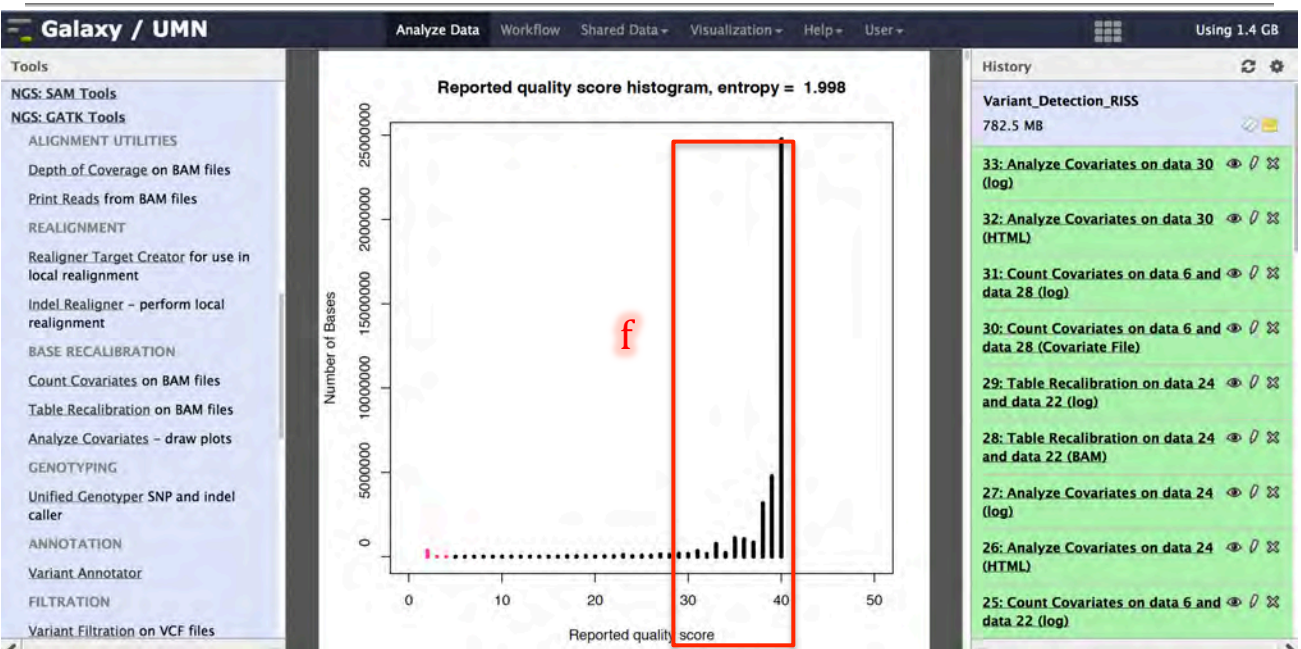
- In the history pane click the eye icon next to the name of the *analyze covariates* HTML output file to display the file in the center pane
- Click on the pdf output file "*NA_10858_400.QualityScoreCovariate.dat.quality_emp_v_stated.pdf*" showing the difference between reported and empirical base calls as a function of the different covariate



- 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

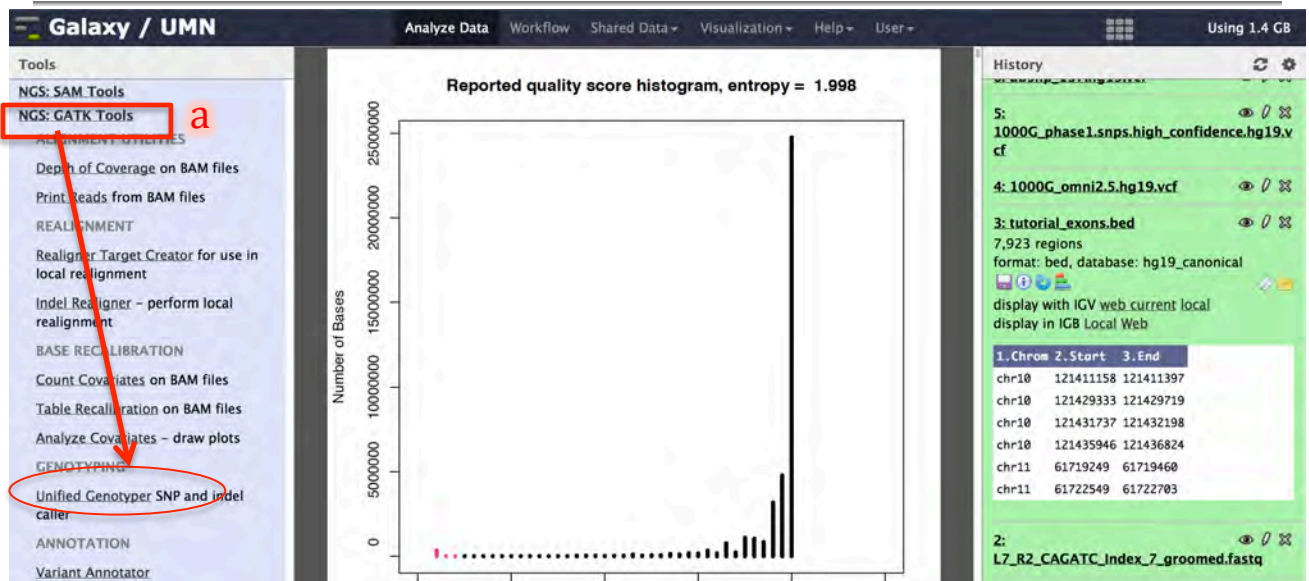
The GATK's Unified Genotyper employs a Bayesian model to compute the likelihood for each of the 10 possible bi-allelic diploid genotypes (AA, AC, AG, AT, CC,CG,CT,GG,GT,TT), as described in the equation below. The likelihood is computed across the entire pileup of bases at a position, taking into consideration 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 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



The screenshot shows the Galaxy / UMN interface with the 'Unified Genotyper (version 0.0.6)' tool configuration form. The 'Tools' pane on the left shows 'Unified Genotyper SNP and indel caller' selected. The main panel contains the following fields and options:

- Choose the source for the reference list:**
- BAM files:**
BAM file 1: (circled in red and labeled 'b')
- Using reference genome:** (circled in red and labeled 'c')
- Binding for reference-ordered data:**
 (circled in red and labeled 'd')
- Genotype likelihoods calculation model to employ:**

The 'History' pane on the right shows a list of jobs, including '1000G_phase1.snps.high_confidence.hg19.vcf' and 'L7_R2_CAGATC_Index_7_groomed.fastq'.

- 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

The screenshot shows the Galaxy/UMN interface for configuring a GATK tool. The left sidebar lists various tools under 'NGS: GATK Tools'. The main panel is titled 'Binding for reference-ordered data 1'. It contains several configuration fields:

- Binding Type:** A dropdown menu set to 'dbSNP' (circled in red with letter 'e').
- ROD file:** A text input field containing '6: dbsnp_137.hg19.vcf' (circled in red with letter 'f').
- Genotype likelihoods calculation model to employ:** A dropdown menu set to 'BOTH' (circled in red with letter 'g').
- The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be called:** A text input field set to '20.0' (circled in red with letter 'h').
- 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:** A text input field set to '20.0' (circled in red with letter 'i').
- Basic or Advanced GATK options:** A dropdown menu set to 'Advanced' (circled in red with letter 'j').
- Pedigree files:** A text input field with the value '-ped,--pedigree <pedigree>'.

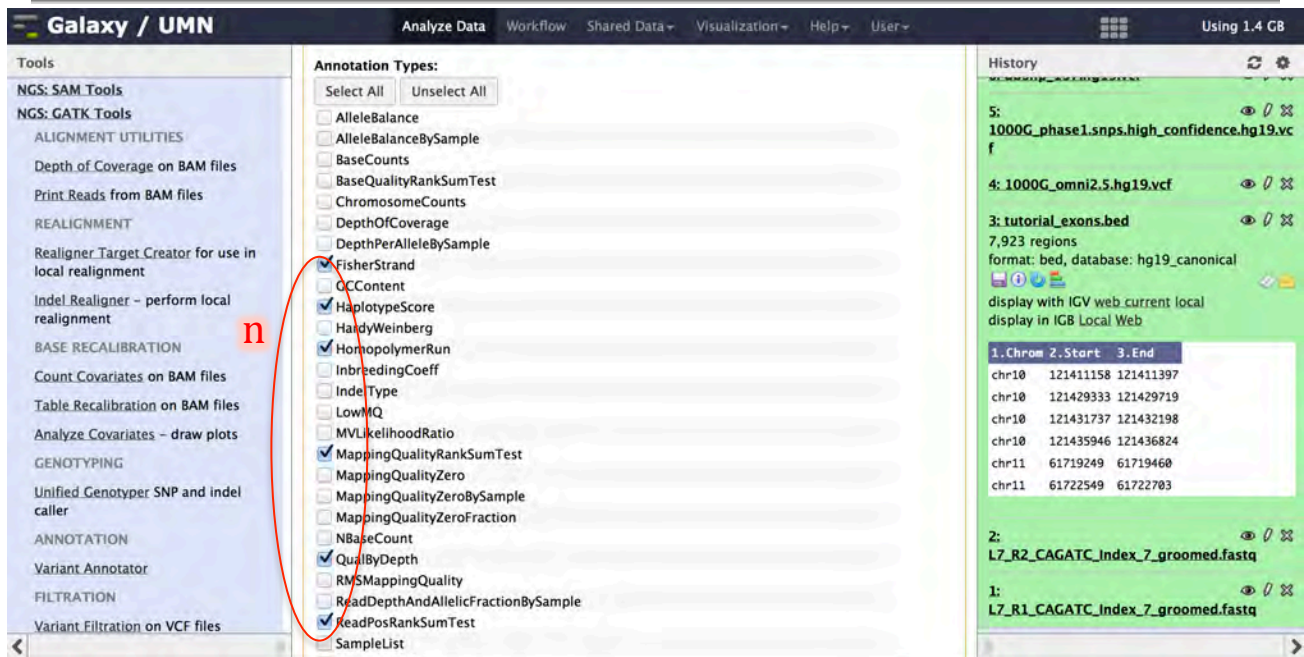
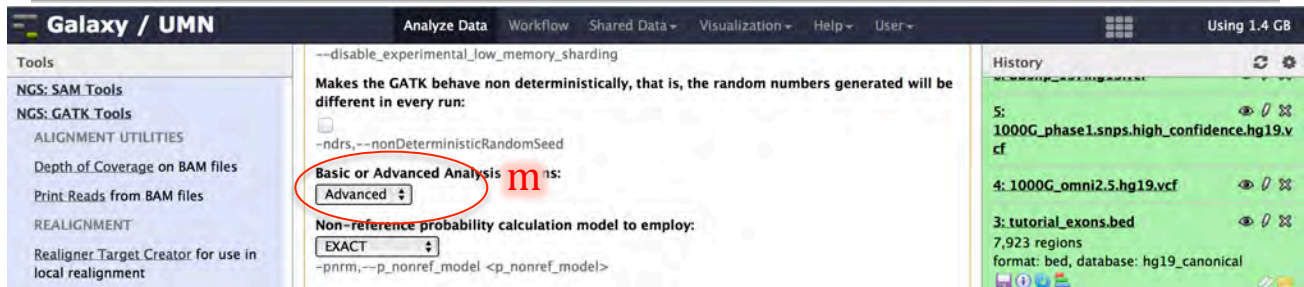
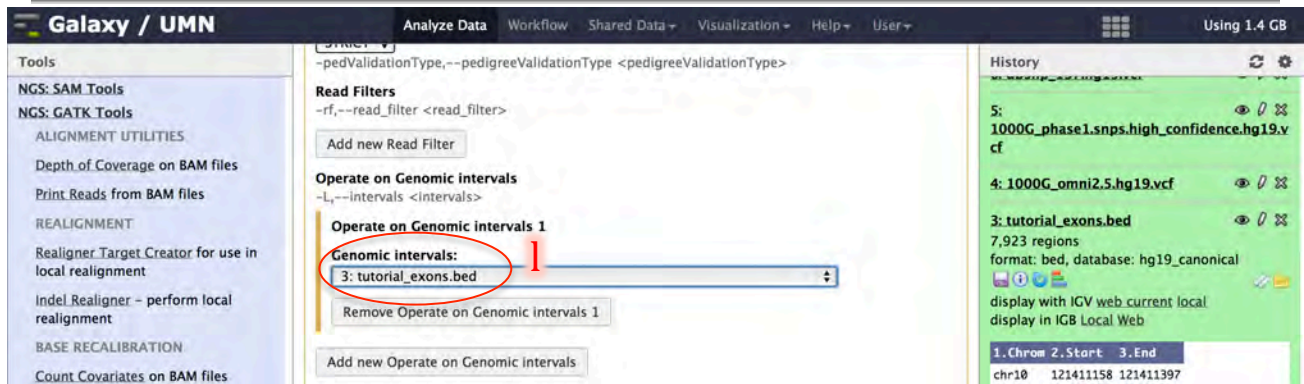
The right sidebar shows a 'History' panel with a list of tracks, including '1000G_phase1.snps.high_confidence.hg19.vcf' and 'tutorial_exons.bed'. A table below the tracks shows genomic coordinates for chromosomes 10 and 11.

This screenshot shows the same Galaxy/UMN interface, but with the configuration panel scrolled down to show additional options. The 'Add new Operate on Genomic intervals' button is circled in red with letter 'k'. Other visible options include:

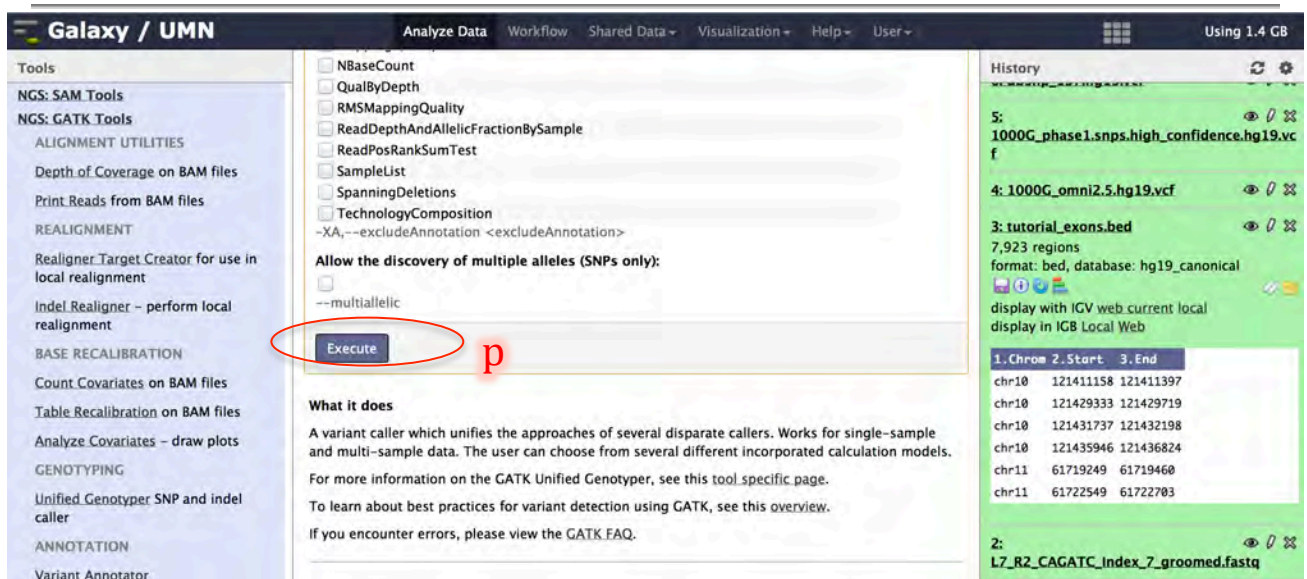
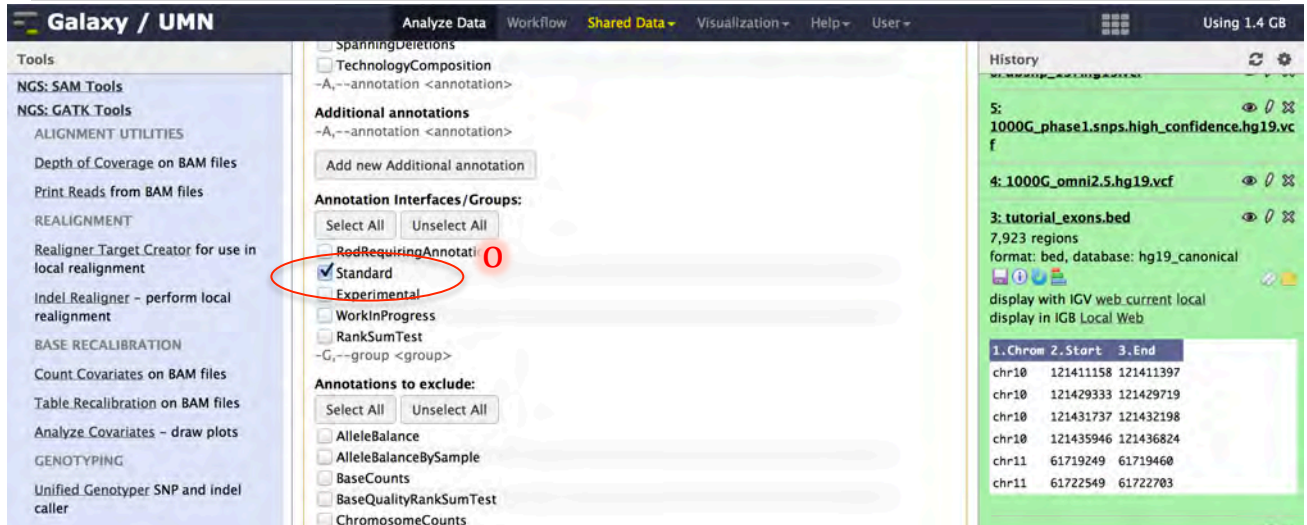
- Read Filters:** A text input field with '-rf,--read_filter <read_filter>' and an 'Add new Read Filter' button.
- Operate on Genomic intervals:** A text input field with '-L,--intervals <intervals>' and an 'Add new Operate on Genomic intervals' button (circled in red with letter 'k').
- Exclude Genomic intervals:** A text input field with '-XL,--excludeIntervals <excludeIntervals>' and an 'Add new Exclude Genomic intervals' button.

The right sidebar remains the same, showing the 'History' panel and track list.

- 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



- o) Annotation Interfaces/Groups: -> check box next to
 - ✓ Standard
- p) Click “Execute”



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 interface. The top navigation bar includes 'Galaxy / UMN', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The right side of the top bar shows 'Using 1.4 GB'. On the left, a 'Tools' pane is visible with categories like 'NGS: SAM Tools', 'NGS: GATK Tools', 'ALIGNMENT UTILITIES', 'ALIGNMENT', 'BASE RECALIBRATION', 'GENOTYPING', 'ANNOTATION', and 'FILTRATION'. The main content area displays a list of contigs with their IDs, lengths, and assemblies, followed by a reference file path and a list of chromosomes (chr1). A red letter 'f' is positioned above a red circle that highlights a back arrow icon in the bottom right corner of the browser window.

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

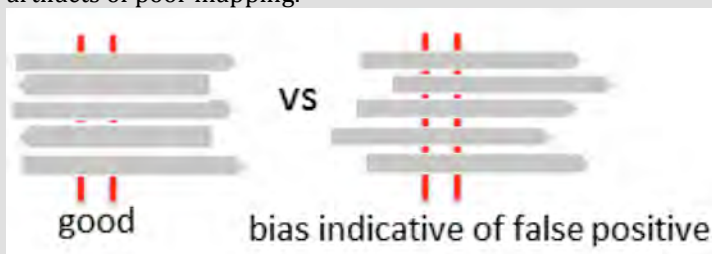
Transition (A<->G, C<->T) and transversion (A<->C, A<->T, C<->G, G<->T) ratios are not random (0.5). Selection pressure works against transversions in coding DNA and in other structurally or functionally-relevant regions. High throughput validation studies have established that Whole genome sequencing typically yields Ti/Tv rates ~2.0-2.1, and exome data around ~3.0-3.3. 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 strand are more likely to be artifacts of poor mapping.



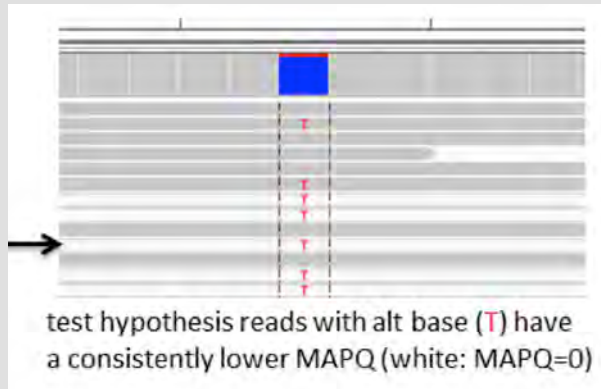
• HomopolymerRun (HRun)

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.

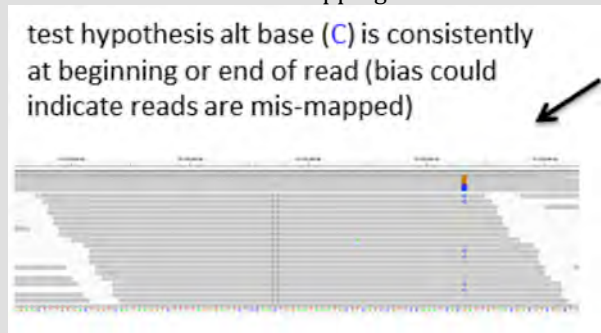


- **QualByDepth (QD)**

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.



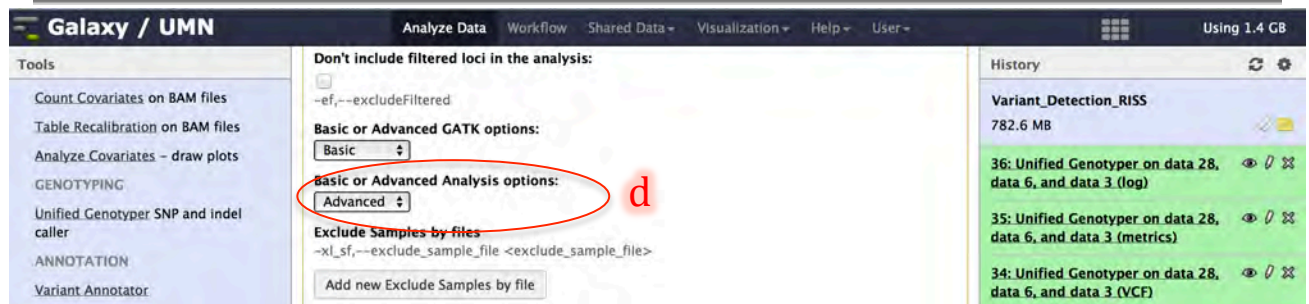
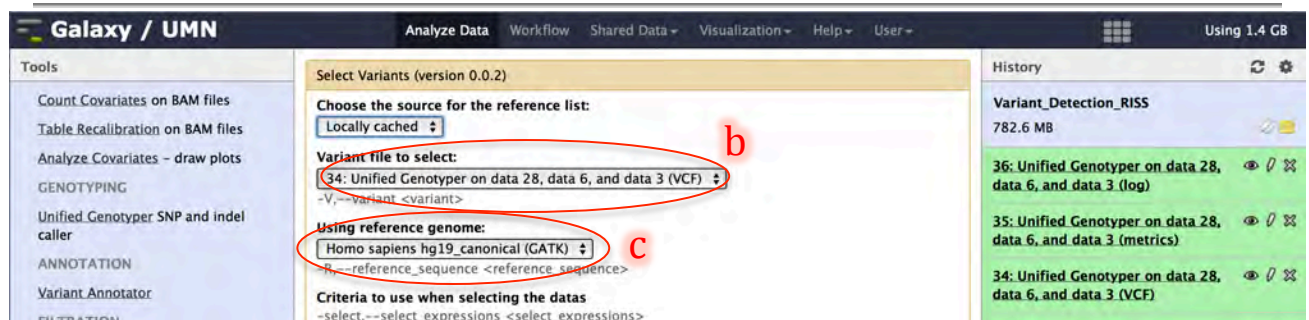
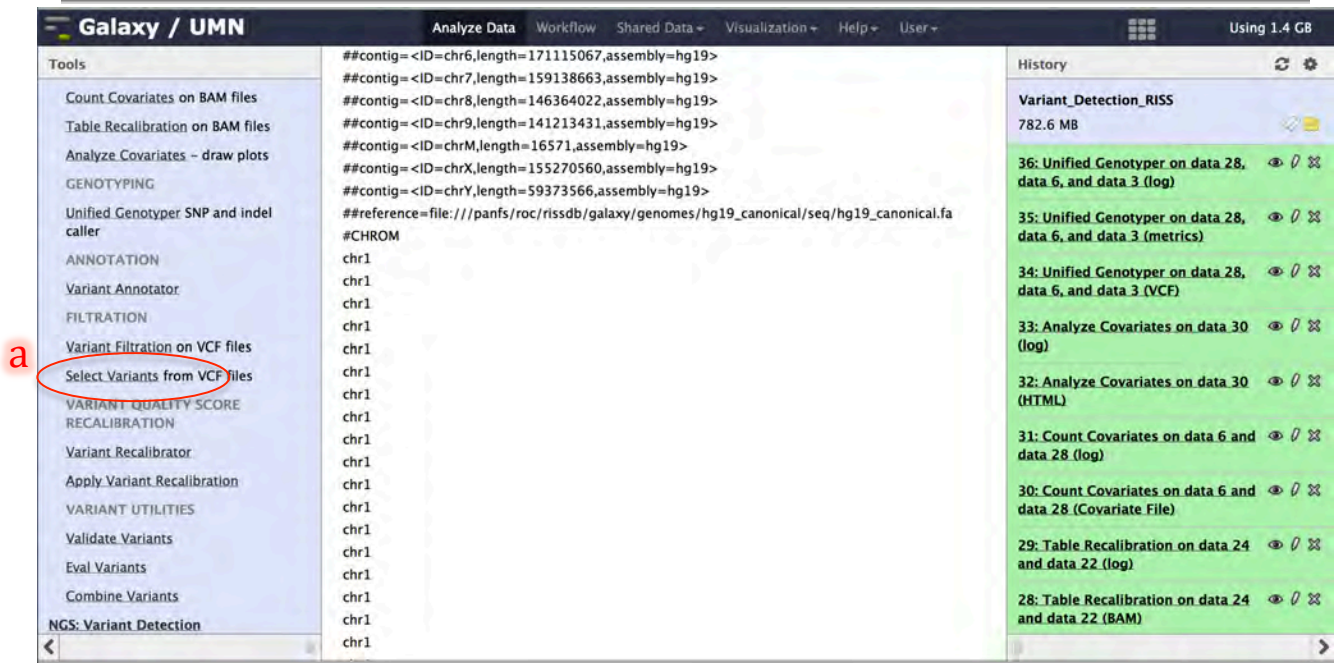
- ★ **Variant Annotation**

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.

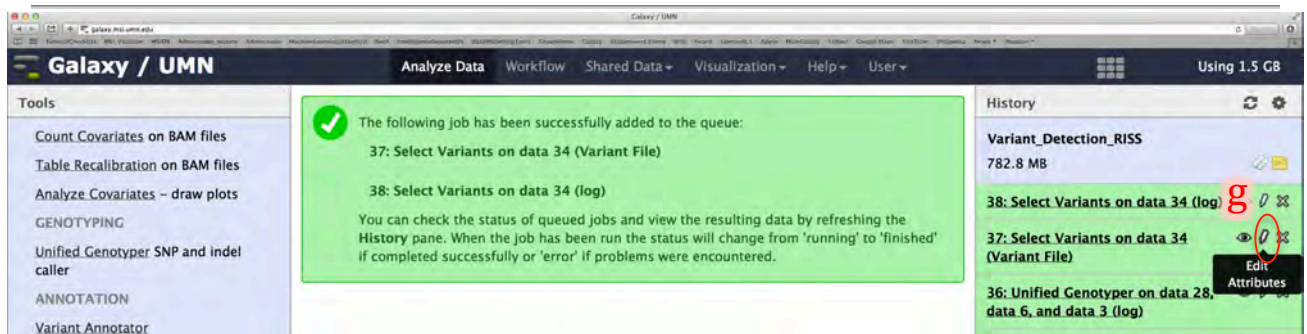
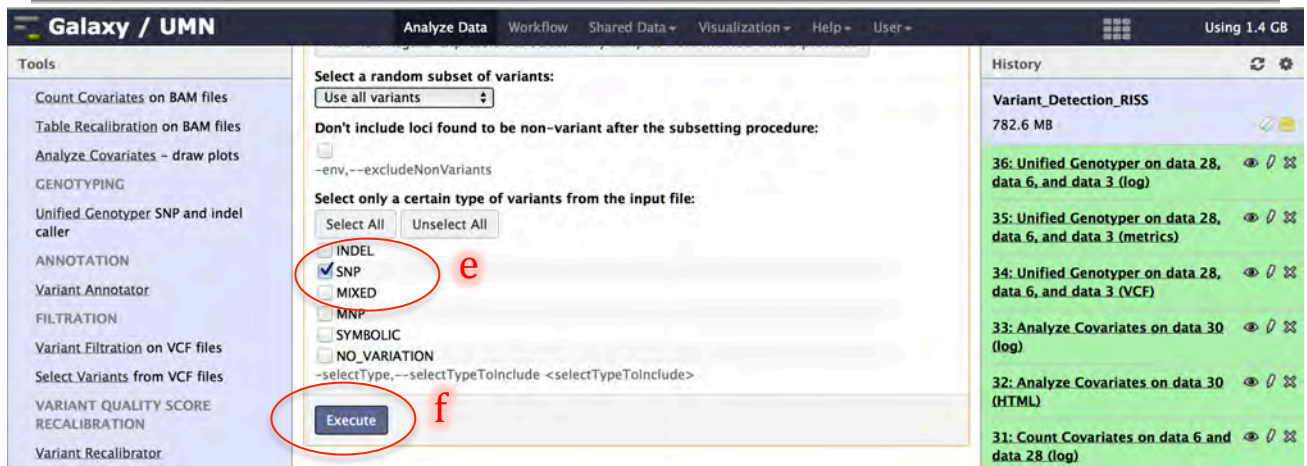
Variant Recalibration

7.1 Select SNPs

- a) Load *Select variants* tool from the tool pane: “NGS: GATK Tools -> Select Variants from VCF files”
- b) Variant file to select: -> “...Unified Genotyper...(VCF)”
- c) Using reference genome: -> Homo sapiens hg19_canonical (GATK)
- d) Basic or Advanced Analysis options: -> Advanced



- 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



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

The screenshot shows the Galaxy / UMN interface. The main panel is titled "Edit Attributes" and contains the following fields and options:

- Name:** A text input field containing "SNPs". A red circle highlights this field, and a red letter "h" is placed to its right.
- Info:** An empty text input field.
- Annotation / Notes:** An empty text input field.
- Database/Build:** A dropdown menu showing "Human hg19 in GATK canonical chr...".
- Number of comment lines:** A checkbox labeled "S3" is checked.
- Save:** A button highlighted with a red circle. A red letter "i" is placed to its right.
- Auto-detect:** A button.

Below the "Auto-detect" button, there is a note: "This will inspect the dataset and attempt to correct the above column values if they are not accurate."

The left sidebar shows a list of tools under "NGS: Variant Detection", including "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", and "Combine Variants".

The right sidebar shows the "History" panel for the dataset "Variant_Detection_RISS" (782.8 MB). It lists several operations, including "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)", and "30: Count Covariates on data 6 and data 28 (Covariate File)".

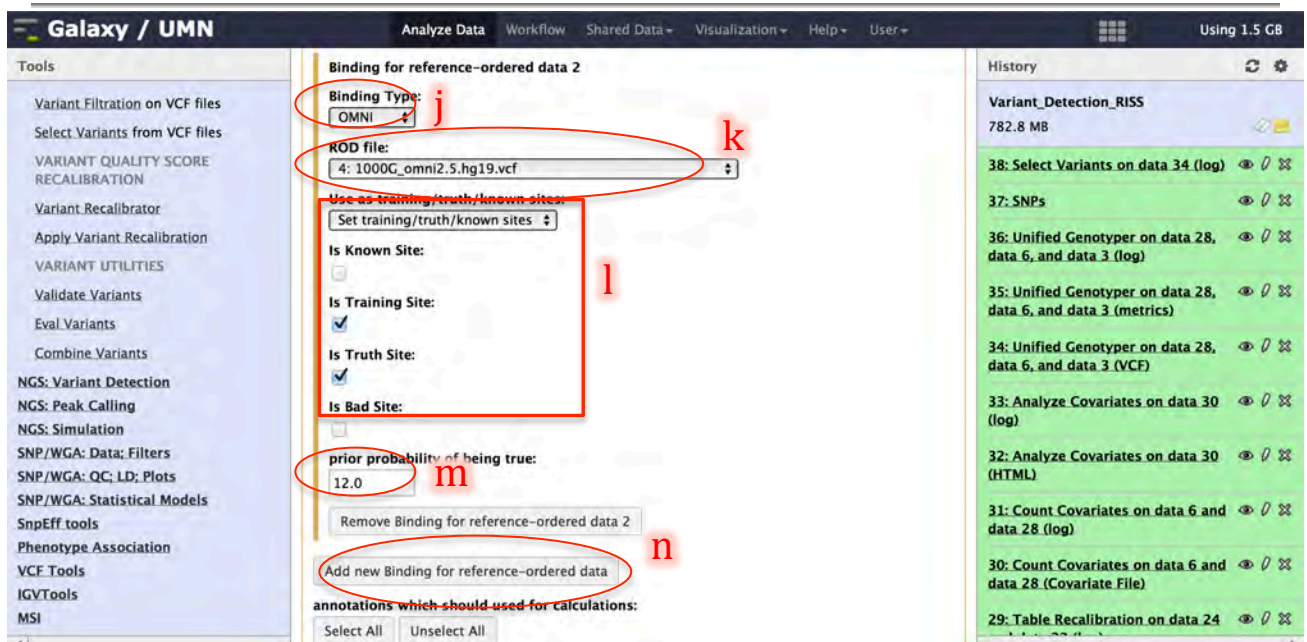
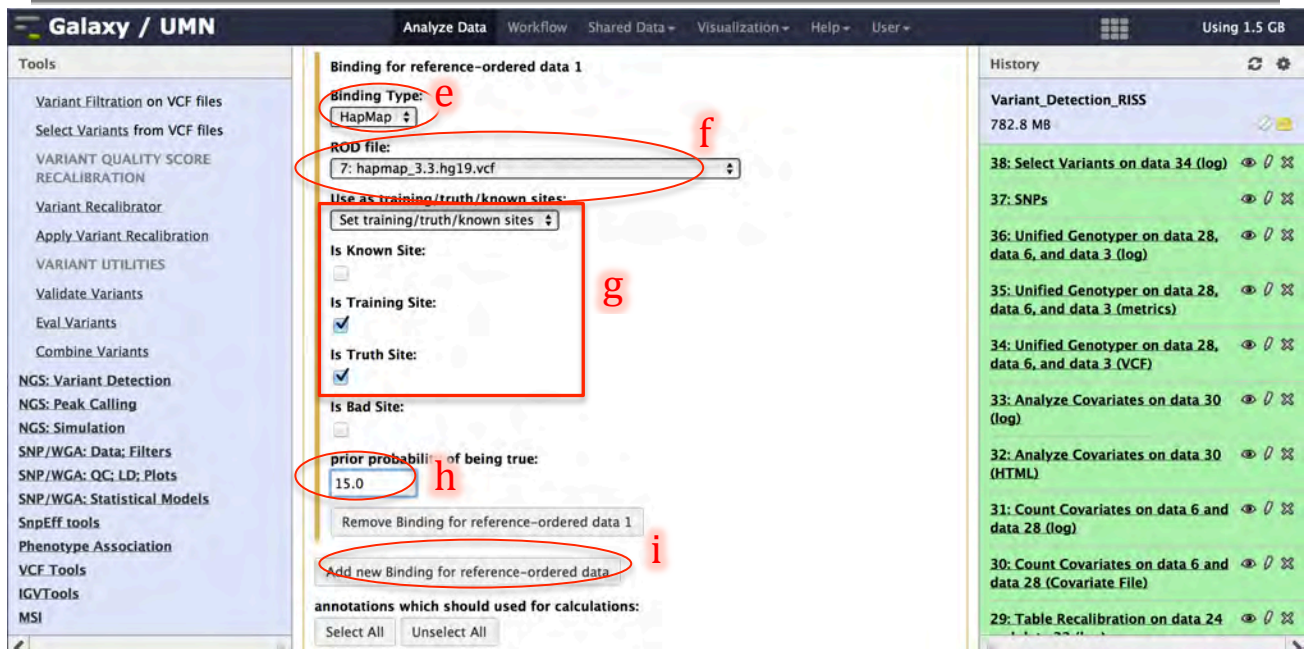
7.2 Recalibrate SNPs

- 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”

The screenshot shows the Galaxy/UMN interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'Tools' pane on the left lists various tools, with 'Variant Recalibrator' circled in red and labeled 'a'. The main panel shows the 'Edit Attributes' section for the tool, with fields for 'Name' (SNPs), 'Info', and 'Annotation / Notes'. A green notification bar at the top indicates 'Attributes updated'. The right-hand 'History' pane shows a list of previous jobs, including 'Variant_Detection_RISS' and several 'Unified Genotyper' jobs.

The screenshot shows the configuration screen for the 'Variant Recalibrator (version 0.0.4)' tool. The 'Choose the source for the reference list' dropdown is set to 'Locally cached'. The 'Variants' section shows a list of variants, with '37: SNPs' selected and circled in red, labeled 'b'. The 'Using reference genome' dropdown is set to 'Homo sapiens hg19_canonical (GATK)' and circled in red, labeled 'c'. The 'Binding for reference-ordered datas' section has the 'Add new Binding for reference-ordered data' button circled in red, labeled 'd'. The 'Annotations which should be used for calculations' section has 'Select All' and 'Unselect All' buttons. The right-hand 'History' pane shows a list of previous jobs, including 'Variant_Detection_RISS' and several 'Analyze Covariates' jobs.

- 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"



- 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
IGV.Tools
MSI

Binding for reference-ordered data 3

Binding Type: 1000G **o**

ROD file: 5: 1000G_phase1.snps.high_confidence.hg19.vcf **p**

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

Is Known Site:

Is Training Site: **q**

Is Truth Site:

Is Bad Site:

prior probability of being true:
10.0 **r**

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
IGV.Tools
MSI

Binding for reference-ordered data 4

Binding Type: dbSNP **t**

ROD file: 6: dbsnp_137.hg19.vcf **u**

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:
2.0 **w**

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

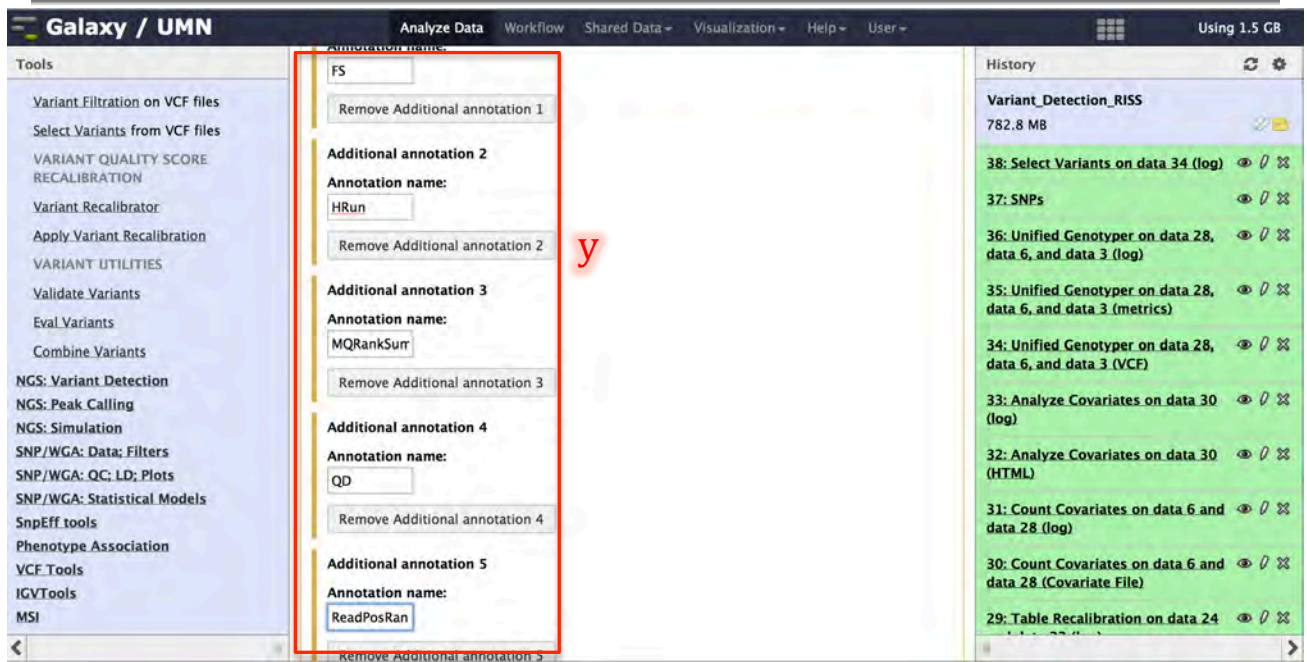
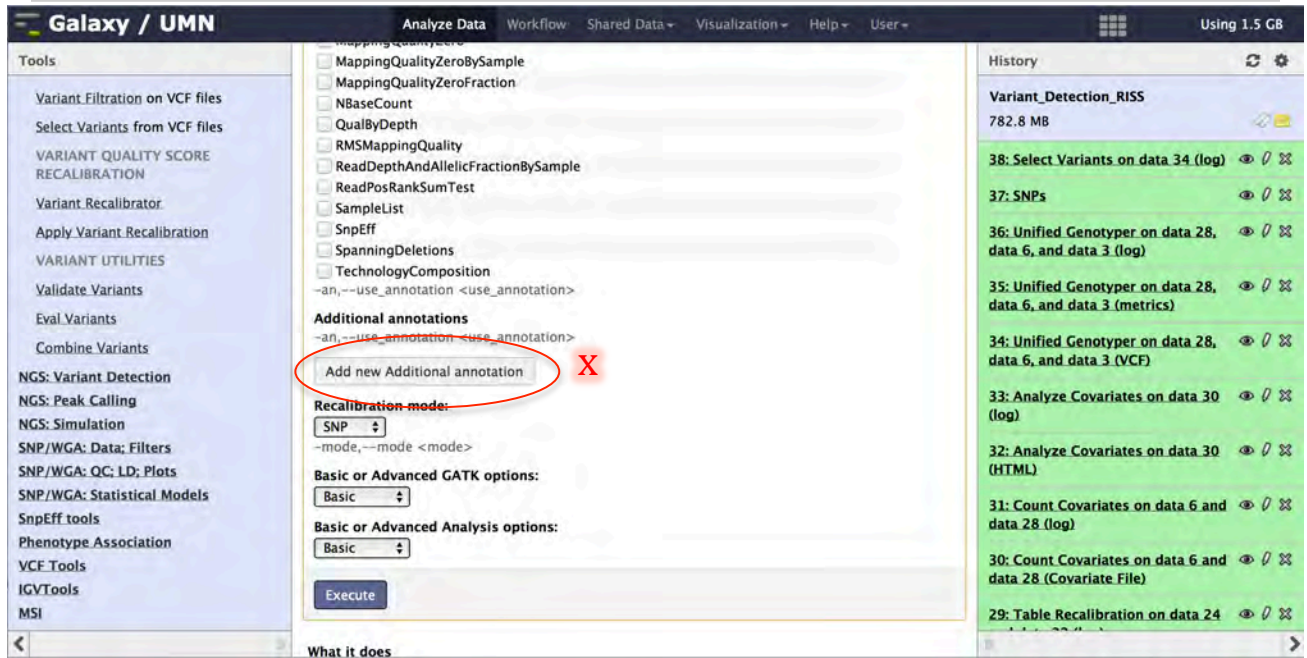
Annotation name: -> “FS”

Annotation name: -> “HRun”

Annotation name: -> “MQRankSum”

Annotation name: -> “QD”

Annotation name: -> “ReadPosRankSum”



- 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/UMN interface for the GATK Recalibration tool. The left sidebar lists various tools, including Variant Filtration, Variant Recalibrator, and NGS tools. The main panel shows the tool configuration with several options circled in red and labeled with letters:

- z**: The "Recalibration mode" dropdown is set to "SNP".
- aa**: The "Basic or Advanced GATK options" dropdown is set to "Advanced".
- bb**: The "Add new Operate on Genomic intervals" button is highlighted.

Other visible options include Pedigree files, Pedigree strings, How strict should we be in validating the pedigree information (set to STRICT), Read Filters, and Exclude Genomic intervals.

This screenshot shows the "Operate on Genomic intervals" section of the tool configuration. A red circle highlights the "Add new Operate on Genomic intervals" button, which is labeled with the letter **cc**. Below it, the "Genomic intervals" dropdown menu is open, showing the selected option "3: tutorial_exons.bed".

Other options in this section include "Exclude Genomic intervals" and "Interval set rule" (set to UNION).

- 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"

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

ALL

-lm,--interval_merging <interval_merging>

Read group black lists
 --rgbl,--read_group_black_list <read_group_black_list>

Add new Read group black list

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

maximum number of Gaussians to try during variational Bayes Algorithm:
 1
 -mG,--maxGaussians <maxGaussians>

maximum number of maximum number of VBEM iterations to be performed in variational Bayes Algorithm:
 100
 -ml,--maxIterations <maxIterations>

number of k-means iterations to perform in order to initialize the means of the Gaussians in the Gaussian mixture model:
 30
 -nKM,--numKMeans <numKMeans>

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

How to specify bad variants:
 Number

minimum amount of worst scoring variants to use when building the Gaussian mixture model of bad variants. Will override -percentBad argument if necessary:
 50
 --minNumBad,--minNumBadVariants <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
 --tranche,--TStranche <TStranche>

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

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

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 a
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: b
37: SNPs

Add new Variant

Variant Recalibration file:
39: Variant Recalibrator on data 6, data 37, and others (Recalibration File) d
-recalFile,--recal_file <recal_file>

Variant Tranches file:
40: Variant Recalibrator on data 6, data 37, and others (Tranches File) d
-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 Tranches file:
40: Variant Recalibrator on data 6, data 37, and others (Tranches File) d
-tranchesFile,--tranches_file <tranches_file>

Using reference genome: c
Homo sapiens hg19_canonical (GATK)

Basic or Advanced GATK options:
Basic

Recalibration mode: d
SNP

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 e

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/UMN interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'Tools' pane on the left lists various variant analysis tools. A central green notification box 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)'. Below this, it says: '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 'History' pane on the right shows a list of jobs, with job 42 'Variant Recalibrator on data 6, data 37, and others (PDF File)' highlighted. A red circle 'a' is drawn around the eye icon next to this job, and a 'View data' button is visible below it.

The screenshot shows the Galaxy/UMN interface with four plots displayed in the center pane. The top-left plot is a heatmap titled 'model PDF' with 'MCRankSum' on the y-axis and 'HRun' on the x-axis. The top-right plot is a scatter plot of 'MCRankSum' vs 'HRun' with a legend for 'outcome retained' (red) and 'filtered' (grey). The bottom-left plot is a scatter plot of 'MCRankSum' vs 'HRun' with a legend for 'training pos' (green) and 'neg' (purple). The bottom-right plot is a scatter plot of 'MCRankSum' vs 'HRun' with a legend for 'novelty known' (blue) and 'novel' (red). A red arrow 'b' points from the top-right plot down to the bottom-right plot. The 'History' pane on the right shows a list of jobs, with job 42 'Variant Recalibrator on data 6, data 37, and others (PDF File)' highlighted.

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

The screenshot shows the Galaxy/UMN interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'Tools' pane on the left lists various tools, with 'Apply Variant Recalibration' selected. The main workspace displays two plots: 'model PDF' (a heatmap of HRun vs. lod) and a scatter plot of HRun vs. Pos. The 'History' pane on the right shows a list of jobs, with job 44 'Apply Variant Recalibration on data 40, data 39, and data 37 (Variant File)' highlighted. A red circle 'a' is around the eye icon for job 44, and a red circle 'b' is around the arrow icon at the bottom of the tools pane.

The screenshot shows the Galaxy/UMN interface with a detailed view of a variant file. The 'Tools' pane on the left is expanded to show 'MSI' (Marked as Selected in Image). The main workspace displays a list of variant information, including Chrom, Pos, and various INFO fields. The 'History' pane on the right shows a list of jobs, with job 44 'Apply Variant Recalibration on data 40, data 39, and data 37 (log)' highlighted. A red circle 'a' is around the eye icon for job 44, and a red circle 'b' is around the arrow icon at the bottom of the tools pane.

- 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 exclu
##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">
##FILTER=<ID=TruthSensitivityTranche99.90to100.00,Description="Truth sensitivity tranche level at VSQ Lod: -113.3625 <= x < -5.9187">
##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">
```

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, data 27, and others (log)

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
122.22	PASS

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 interface. On the left is a 'Tools' pane with a list of bioinformatics tools. On the right is a data table with columns 'QUAL' and 'FILTER'. A red letter 'f' is positioned above a red circle that highlights a back arrow icon in the bottom right corner of the browser window.

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
112.50	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

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** a
- VARIANT QUALITY SCORE
- RECALIBRATION

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)

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

Tools

Select Variants (version 0.0.2)

Choose the source for the reference list:
Locally cached

Variant file to select: b
34: Unified Genotyper on data 28, data 6, and data 3 (VCF)

Using reference genome: c
Homo sapiens hg19_canonical (GATK)

Criteria to use when selecting the datas

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)

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 1.5 GB

Tools

Basic or Advanced Analysis options: d
Advanced

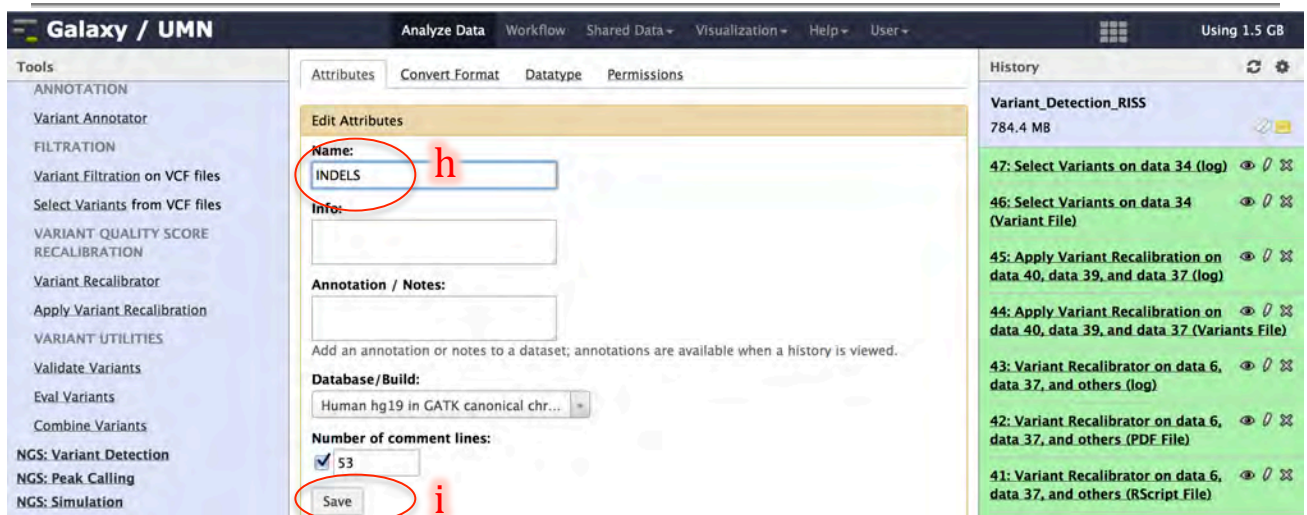
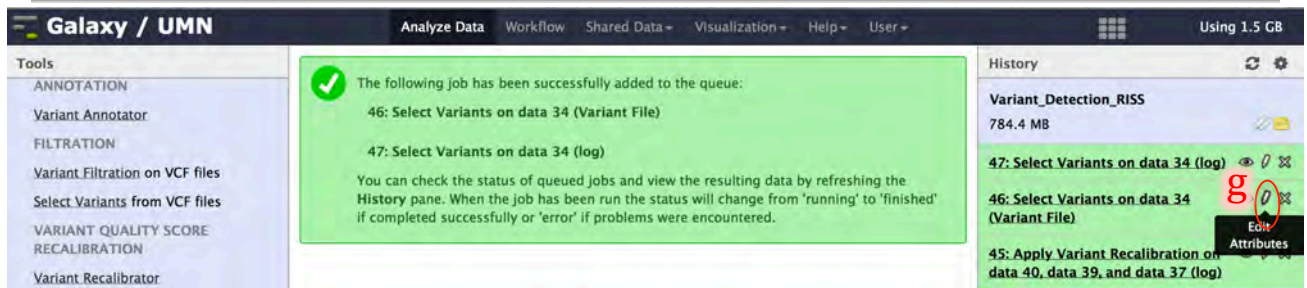
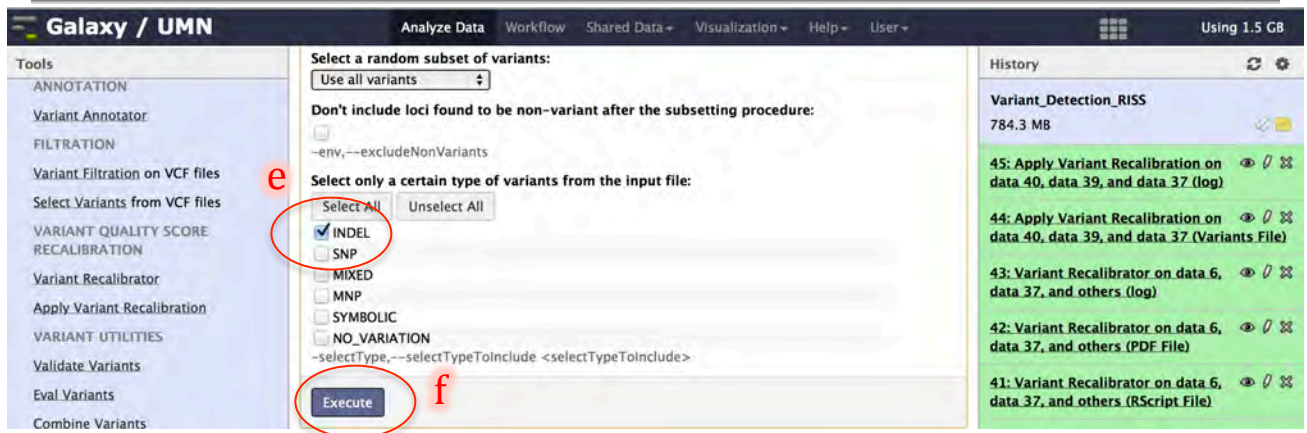
Exclude Samples by files
-xl_sf,--exclude_sample_file <exclude_sample_file>

Samples by files

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)

- 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"



7.7 Recalibrate INDELS

- a) Load *variant recalibration* tool from the tool pane: “NGS: GATK Tools -> Variant Recalibrator”
- b) Variant file to recalibrate: -> “...INDELS”
- c) Using reference genome: -> Homo sapiens hg19_canonical (GATK)
- d) Click on “Add new Binding for reference-ordered data”
- e) Binding Type: -> Custom
- f) ROD Name: -> mills
- g) ROD file: -> Mills_and_1000G_gold_standard.indels.hg19.vcf
- h) Use as training/truth/known sites: -> Set training/truth/known sites
 - Is Training Site: -> ✓
 - Is Truth Site: -> ✓
- i) prior probability of being true: -> 12.0
- j) 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 **a**

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

Variants

--input,--input <input>

Variant 1

Variant file to recalibrate: **b**

46: INDELS

Add new Variant

Using reference genome: **c**

Homo sapiens hg19_canonical (GATK)

--R,--reference_sequence <reference_sequence>

Binding for reference-ordered data

--resource,--resource <resources>

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

annotations which should used for calculations:

Select All Unselect All

History

Variant_Detection_RISS

784.4 MB

47: Select Variants on data 34 (log)

46: INDELS

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)

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

SnEff tools

Phenotype Association

Binding for reference-ordered data 1

Binding Type: **e**

Custom

ROD Name: **f**

mills

ROD file: **g**

8: Mills_and_1000G_gold_standard.indels.hg19.vcf

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: **i**

12.0

Remove Binding for reference-ordered data 1

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

annotations which should used for calculations:

Select All Unselect All

History

Variant_Detection_RISS

784.4 MB

47: Select Variants on data 34 (log)

46: INDELS

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)

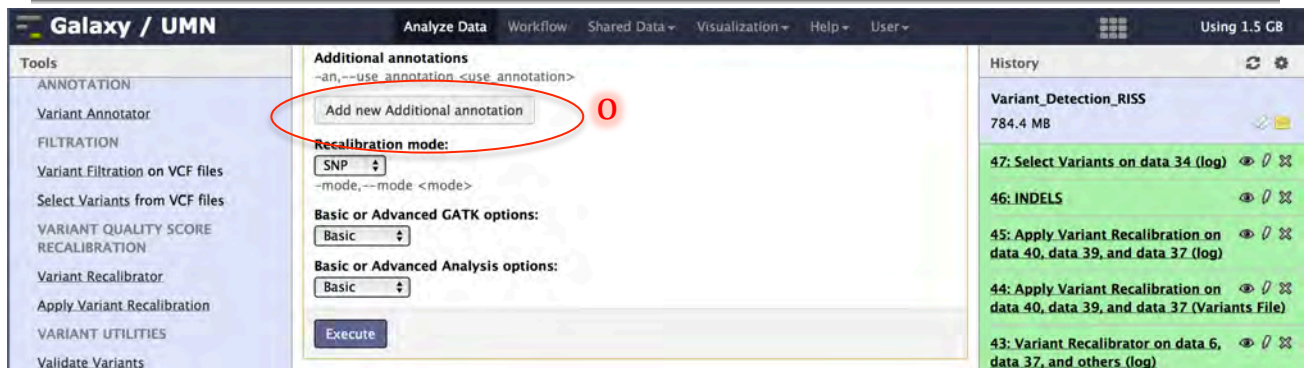
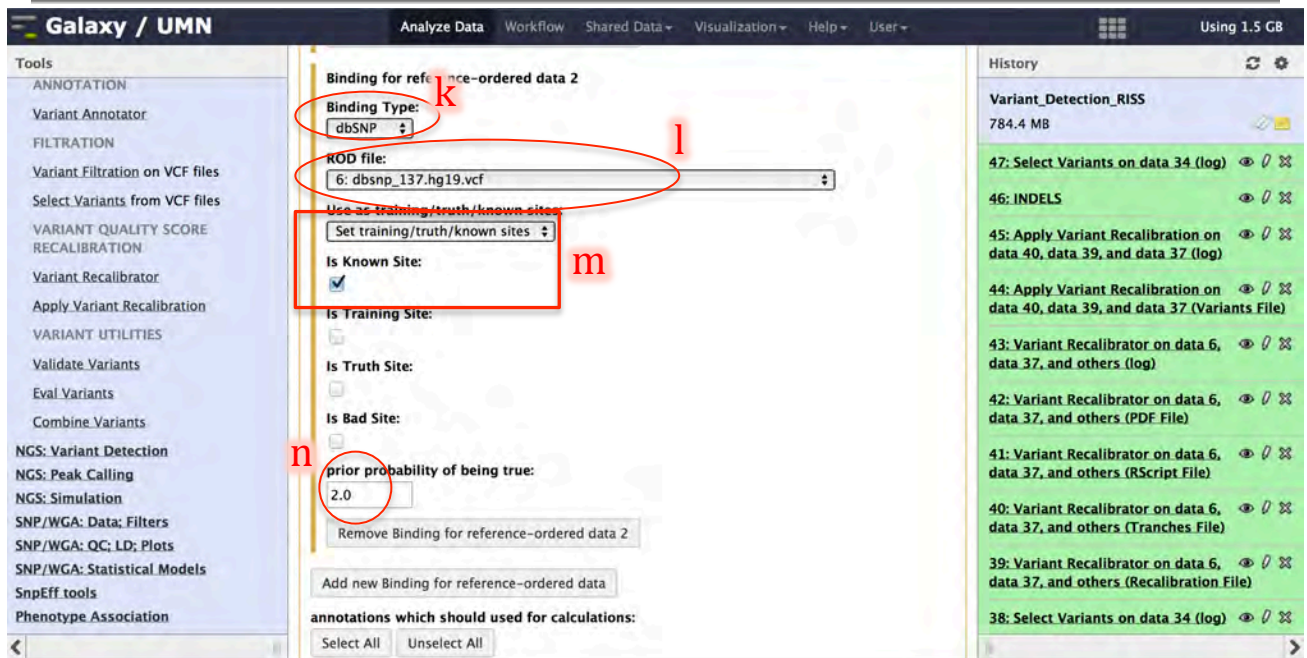
41: Variant Recalibrator on data 6, data 37, and others (RScript File)

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

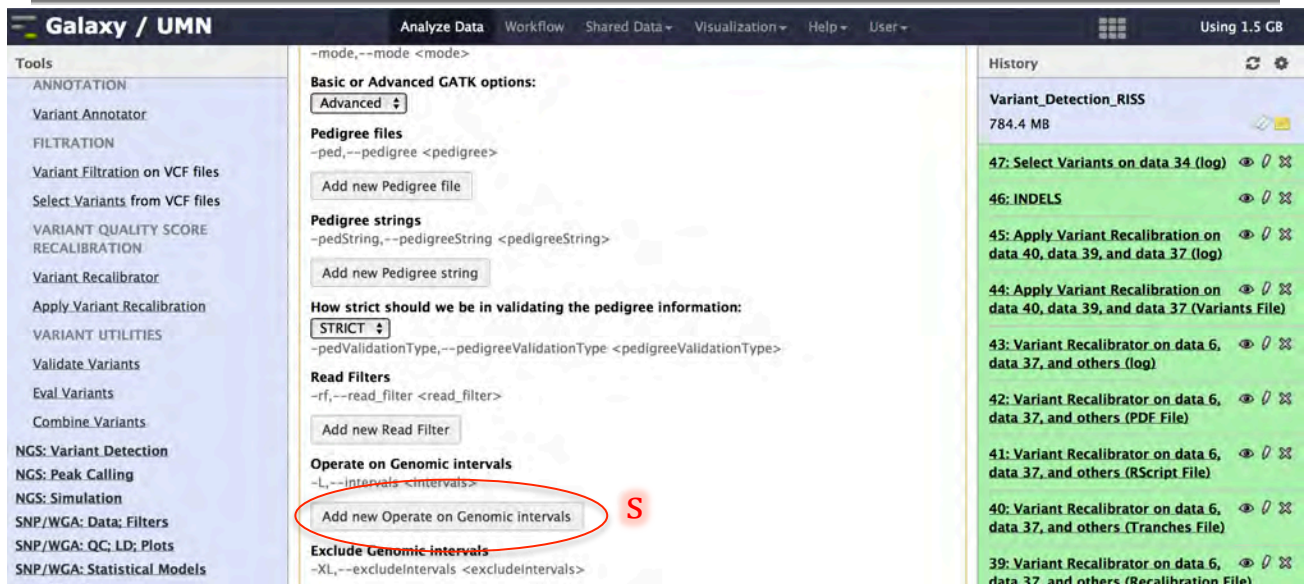
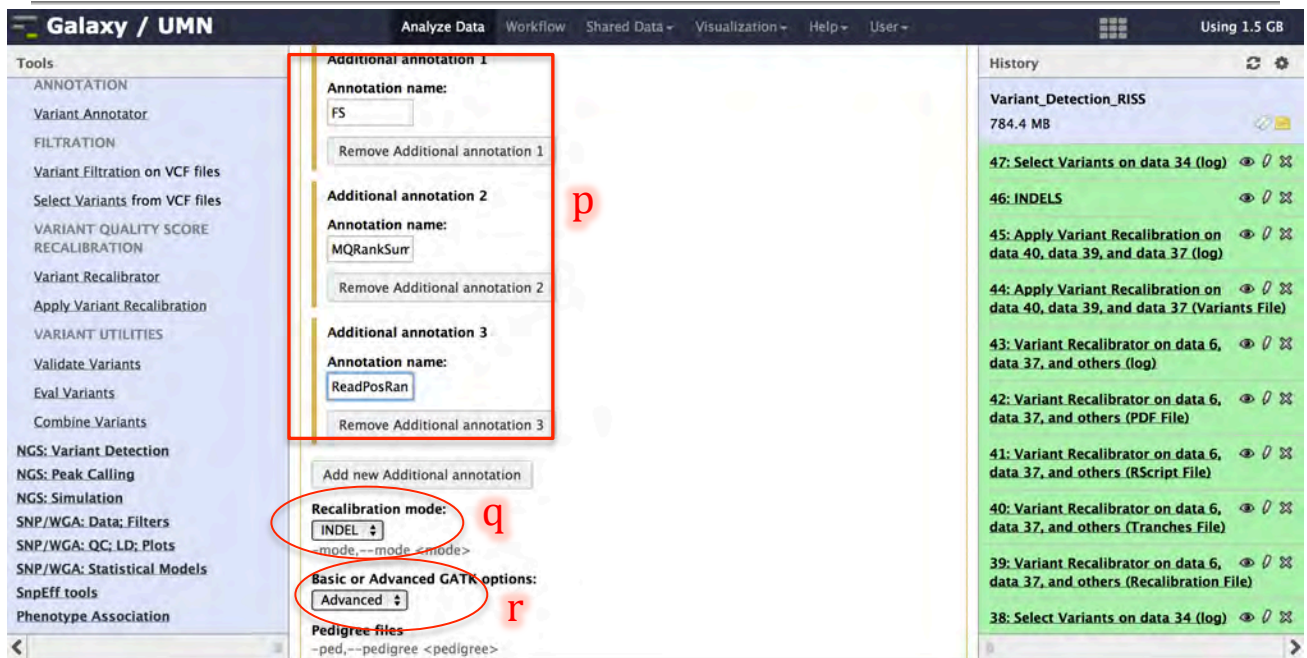
39: Variant Recalibrator on data 6, data 37, and others (Recalibration File)

38: Select Variants on data 34 (log)

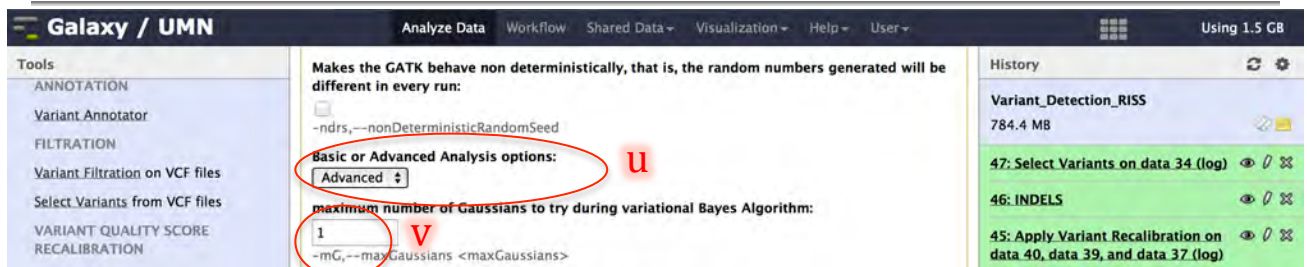
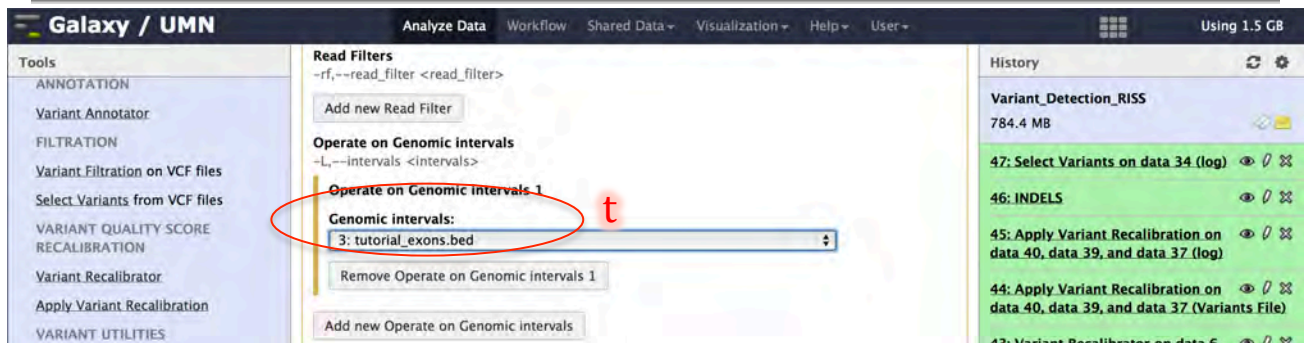
- k) Binding Type: -> dbSNP
- l) ROD file: -> dbsnp_137.hg19.vcf
- m) Use as training/truth/known sites: -> Set training/truth/known sites
Is Known Site:-> ✓
- n) prior probability of being true: -> 2.0
- o) Click on “Add new Addition Annotations” **three times**



- 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



- 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"



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/UMN interface. On the left, the 'Tools' panel is visible, with 'Apply Variant Recalibration' circled in red and labeled 'a'. The main area displays a green notification box with a checkmark, stating: 'The following job has been successfully added to the queue: 48: Variant Recalibrator on data 3, data 8, and others (Recalibration File)'. Below this, a list of jobs is shown, including '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)', and '52: Variant Recalibrator on data 3, data 8, and others (log)'. The 'History' panel on the right shows a list of jobs, including 'Variant_Detection_RISS' and several 'Variant Recalibrator' jobs.

The screenshot shows the configuration interface for the 'Apply Variant Recalibration (version 0.0.4)' tool. The 'Choose the source for the reference list' dropdown is set to 'Locally cached'. The 'Variants' section has 'Variant 1' selected, with 'Variant file to annotate:' set to '46: INDELS' (circled in red and labeled 'b'). The 'Variant Recalibration file:' is set to '48: Variant Recalibrator on data 3, data 8, and others (Recalibration File)'. The 'Variant Tranches file:' is set to '49: Variant Recalibrator on data 3, data 8, and others (Tranches File)'. The 'Using reference genome:' dropdown is set to 'Homo sapiens hg19_canonical (GATK)' (circled in red and labeled 'c'). The 'Basic or Advanced GATK options:' dropdown is set to 'Basic'. The 'History' panel on the right shows a list of jobs, including 'Variant_Detection_RISS' and several 'Variant Recalibrator' jobs.

The screenshot shows the execution interface for the 'Apply Variant Recalibration (version 0.0.4)' tool. The 'Basic or Advanced GATK options:' dropdown is set to 'Basic'. The 'Recalibration mode:' dropdown is set to 'INDEL' (circled in red and labeled 'd'). The 'Ignore Filters' section has 'Add new Ignore Filter' button. The 'truth sensitivity level at which to start filtering, used here to indicate filtered variants in plots:' is set to '99.0'. The 'Execute' button is circled in red and labeled 'e'. The 'History' panel on the right shows a list of jobs, including 'Variant_Detection_RISS' and several 'Variant Recalibrator' 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/UMN interface. At the top, there's a navigation bar with 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. Below this is a 'Tools' pane on the left with categories like ANNOTATION, FILTRATION, and RECALIBRATION. A central green notification box states: 'The following job has been successfully added to the queue: 53: Apply Variant Recalibration on data 49, data 48, and data 46 (Variants File)'. To the right is a 'History' pane showing a list of jobs, including '54: Apply Variant Recalibration on data 49, data 48, and data 46 (log)' and '53: Apply Variant Recalibration on data 49, data 48, and data 46 (Variants File)'. A red circle highlights the eye icon next to job 53, and another red circle highlights the 'View data' link next to it.

The screenshot shows the Galaxy/UMN interface with a VCF file viewer open. The 'Tools' pane on the left is expanded to show 'NGS: Variant Detection' and 'NGS: Peak Calling'. The main area displays VCF header information, including file format, filters, and various INFO fields. A red circle highlights the left arrow icon at the bottom of the tools pane. The 'History' pane on the right shows a list of jobs, including '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)', and '46: INDELS'. A red circle highlights the eye icon next to the '46: INDELS' job.

- 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

The screenshot shows the Galaxy / UMN interface. The main area displays the content of a VCF file, including headers like ##fileformat=VCFv4.1 and various INFO fields. The right-hand side features a 'History' pane with a list of recent jobs, such as 'Variant_Detection_RISS' and several 'Apply Variant Recalibration' and 'Variant Recalibrator' jobs. A red circle highlights a right-pointing arrow at the bottom of the history pane.

This screenshot shows a zoomed-in view of the VCF file content from the previous image. It displays the REF and ALT columns with various nucleotide sequences. A red double-headed arrow labeled 'd' is positioned between the two columns. In the bottom-left corner, a red circle highlights a left-pointing arrow, labeled 'e', which is used to toggle the tools pane.

f) Click the arrow at the bottom-right corner of your browser to bring the *tools pane* back to 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 ...” (recalibrated INDEL vcf))
- Variant name: -> “indels”
- Using reference genome: -> Homo sapiens hg19_canonical (GATK)
- Click “Execute”

The screenshot shows the Galaxy / UMN interface. The tool pane on the left is expanded to 'NGS: Variant Detection', and the 'Combine Variants' tool is highlighted with a red circle and labeled 'a'. The main panel shows the tool's configuration options, including 'REF', 'T', and 'CCAT'.

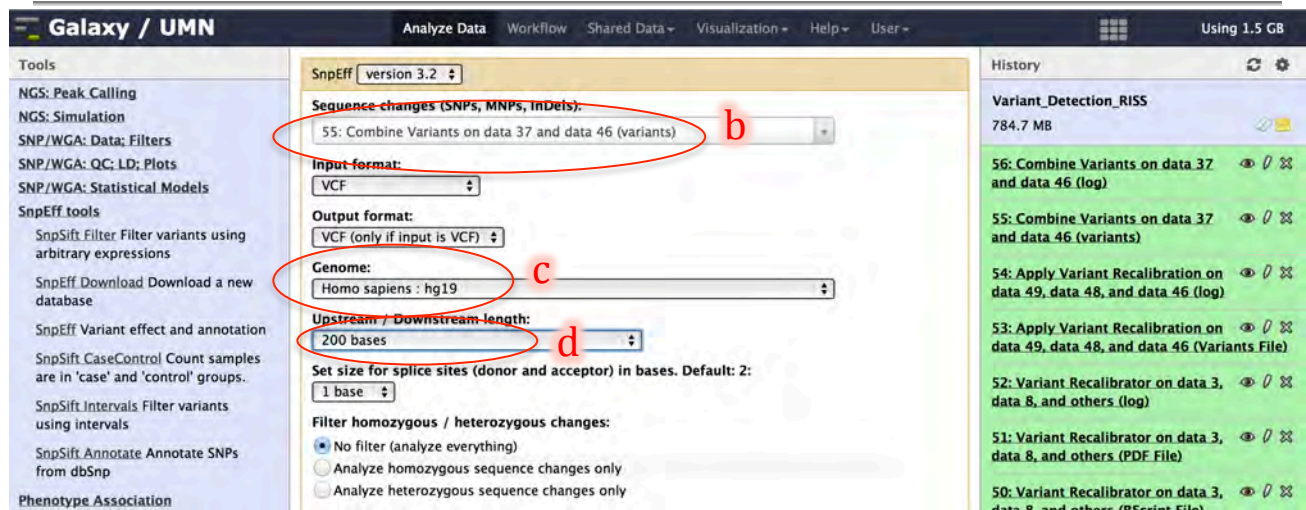
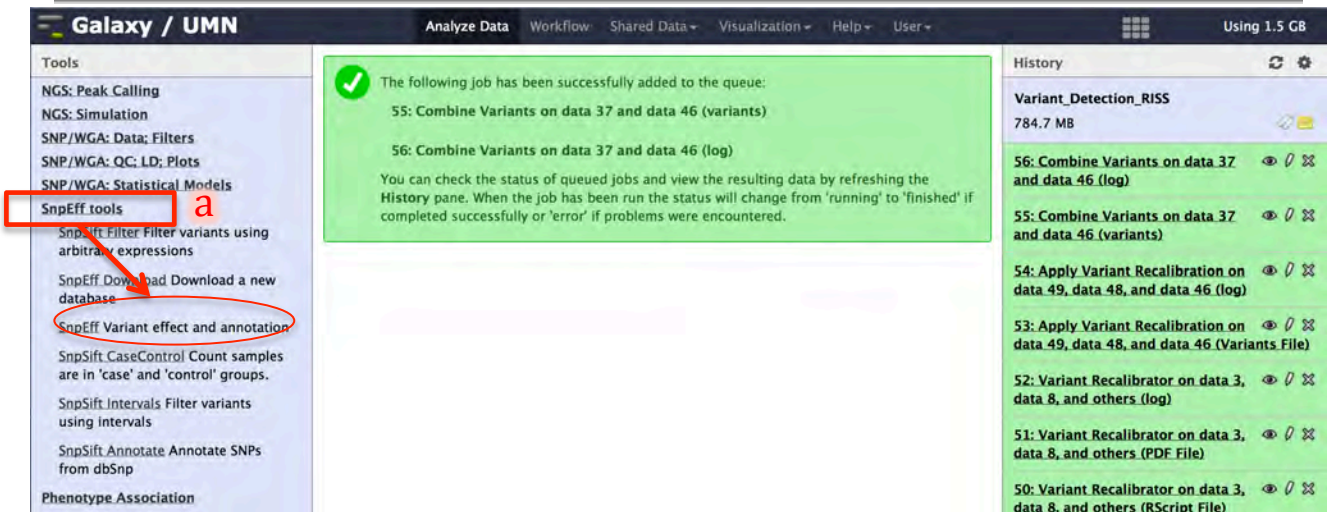
The screenshot shows the 'Combine Variants (version 0.0.4)' tool configuration. The 'Input variant file' field is set to '37: SNPs' and is circled in red with label 'b'. The 'Variant name' field is set to 'snps' and is circled in red with label 'c'. The 'Add new Variants to Merge' button is circled in red with label 'd'. The 'Choose the source for the reference list' dropdown is set to 'Locally cached'.

The screenshot shows the 'Combine Variants' tool configuration for a second merge. The 'Input variant file' field is set to '46: INDELS' and is circled in red with label 'e'. The 'Variant name' field is set to 'indels' and is circled in red with label 'f'. The 'Using reference genome' dropdown is set to 'Homo sapiens hg19_canonical (GATK)' and is circled in red with label 'g'. The 'Execute' button is circled in red with label 'h'. The 'How should we merge genotype records' dropdown is set to 'PRIORITIZE'.

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



e) Click “Execute”

The screenshot displays the Galaxy/UMN web interface for the SnpEff tool. The top navigation bar includes 'Galaxy / UMN', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The left sidebar lists tool categories such as 'NGS: Peak Calling', 'NGS: Simulation', 'SNP/WGA: Data; Filters', 'SNP/WGA: QC; LD; Plots', 'SNP/WGA: Statistical Models', 'SnpEff tools', 'Phenotype Association', 'VCE Tools', 'IGVTools', 'MSI', 'Masonic Cancer Center Tools', and 'EMBOSS'. The main configuration area for SnpEff includes a dropdown menu for 'Only use the transcripts in this file. Format: One transcript ID per line..', a 'Filter output:' section with 'Select All' and 'Unselect All' buttons, and several checkboxes for filtering changes (DOWNSTREAM, INTERGENIC, INTRON, UPSTREAM, and 5_PRIME_UTR or 3_PRIME_UTR). The 'Chromosomal position:' section has radio buttons for 'Use default (based on input type)', 'Force zero-based positions (both input and output)', and 'Force one-based positions (both input and output)'. A text input field is provided for 'Text to prepend to chromosome name:'. Below this, there are checkboxes for 'Produce Summary Stats:' (checked) and 'Do not report usage statistics to server:' (checked). The 'Execute' button is highlighted with a red circle and a red letter 'e'. The bottom of the configuration area contains the text: 'This tool calculate the effect of variants (SNPs/MNPs/Insertions) and deletions.' The right sidebar shows a 'History' section with a job titled 'Variant_Detection_RISS' (784.7 MB) and a list of 8 sub-jobs, each with a refresh and delete icon.

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_SYNONYMOUS CODING

The screenshot displays the Galaxy web interface. On the left, the 'Tools' panel lists various tools, with 'SnpEff Variant effect and annotation' highlighted. The main workspace shows the output of this tool, a VCF file with annotations. A red arrow labeled 'b' points to the annotation 'NON_SYNONYMOUS CODING' in the 'INFO' field. On the right, the 'History' pane shows a list of jobs. A red circle labeled 'a' highlights the 'View data' button next to the job '58: SnpEff on data 55'. Below the history pane, a table shows the coordinates of the variant: 1.Chrom and 2.Pos. The VCF content in the main workspace includes lines like 'snps;EFF=NON_SYNONYMOUS_CODING(MODERATE|MISSENSE|Acc/Ccc|T2:...' and '##fileformat=VCFv4.1'.