

Canadian Bioinformatics Workshops

www.bioinformatics.ca

Creative Commons

This page is available in the following languages:

Afrikaans বাংলা Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Español Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Malayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски (latinica) Sotho svenska
中文 華語 (台灣) isiZulu



Attribution-Share Alike 2.5 Canada

You are free:



to Share — to copy, distribute and transmit the work



to Remix — to adapt the work



Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

[Disclaimer](#)

Your fair dealing and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
[English](#) [French](#)

[Learn how to distribute your work using this licence](#)

Module 7

Galaxy -Lab

Sorana Morrissy & Francis Ouellette
Informatics on High-throughput Sequencing Data
June 10-11, 2015

A screenshot of the Galaxy web interface. The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Cloud", "Help", and "User". The left sidebar contains a "Tools" section with a search bar and a list of tool categories: Get Data, Lift-Over, Text Manipulation, Convert Formats, Filter and Sort, Join, Subtract and Group, NGS: QC and manipulation, NGS: Mapping, NGS: BAM Tools, NGS: Picard, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data, Phenotype Association, snpEff, BEDTools, Genome Diversity, EMBOSS, Regional Variation, FASTA manipulation, Evolution, Multiple Alignments, and Metagenomic analyses. The main content area displays a message: "Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#)." Below this is a large graphic for the "Galaxy Workshop Tokyo 2015" featuring a stylized tower and cherry blossoms. At the bottom, there is a "Tweets" section showing a tweet from the Galaxy Project (@galaxyproject) about GalaxyAdmins Meetup slides and video availability. The right sidebar contains a "History" section with a search bar and a message: "This history is empty. You can [load your own data](#) or [get data from an external source](#)." The top right corner indicates "Using 0 bytes".

SickKids
THE HOSPITAL FOR
SICK CHILDREN

Objectives

- Become more familiar with Galaxy
- Be able to run experiments on Galaxy

Lab exercise

- Lab exercises using the tools from Galaxy

Experiment workflow

- Load the raw data
- Align the raw data to the human reference genome (BWA)
- Sort the reads and perform duplicate removal (Samtools)
- Perform indel cleaning (GATK RealignerTargetCreator and IndelRealigner)
- Call variants (GATK UnifiedGenotyper)
- Filter variants (GATK VariantFiltration)
- Annotate variants (GATK VariantAnnotator with dbSNPs)

NOTE: The best practice when using GATK is to use the VariantRecalibrator. In our data set, we had too few variants to accurately use the variant recalibrator and therefore we used the VariantFiltration tool instead.

Let's try using the real thing
<https://usegalaxy.org>

Getting data

- Most of time, you will get from a file on your computer, or from a URL.

The image shows two screenshots of the Galaxy web interface. The left screenshot shows the 'Tools' sidebar with the 'Get Data' section expanded, listing various data sources. An arrow points from this section to the right screenshot. The right screenshot shows the 'Download data directly from web or upload files from your disk' dialog box. This dialog has a table with columns: Name, Size, Type, Genome, Settings, and Status. A new file is listed with the name 'New File', size '0.2 KB', and type 'Auto-det...'. Below the table, there is a text box containing three URLs: http://cbwmain.entrydns.org/module2_3/NA12878_CBW_chr1_R1.fastq.gz, http://cbwmain.entrydns.org/module2_3/NA12878_CBW_chr1_R2.fastq.gz, and http://cbwmain.entrydns.org/module7/hg19_chr1.fa. A callout box with an arrow pointing to the text box says 'Use URL to upload input files'. At the bottom of the dialog are buttons for 'Choose local file', 'Paste/Fetch data', 'Start', 'Pause', 'Reset', and 'Close'.

Galaxy

Tools

search tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- BX table browser
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- Flymine test server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- WormBase test server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server

Download data directly from web or upload files from your disk

You can Drag & Drop files into this box.

Name	Size	Type	Genome	Settings	Status
New File	0.2 KB	Auto-det...	----- Additional S...	100%	✓

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

http://cbwmain.entrydns.org/module2_3/NA12878_CBW_chr1_R1.fastq.gz
http://cbwmain.entrydns.org/module2_3/NA12878_CBW_chr1_R2.fastq.gz
http://cbwmain.entrydns.org/module7/hg19_chr1.fa

Type (set all): Auto-detect

Genome (set all): ----- Additional Species ...

Choose local file Paste/Fetch data Start Pause Reset Close

Use URL to upload input files

- We will now use different tools available on galaxy
- Check/modify the value(s) for all cases indicated by \longrightarrow
- If not indicated by \longrightarrow ; keep the default parameter(s)

Upload 3 files

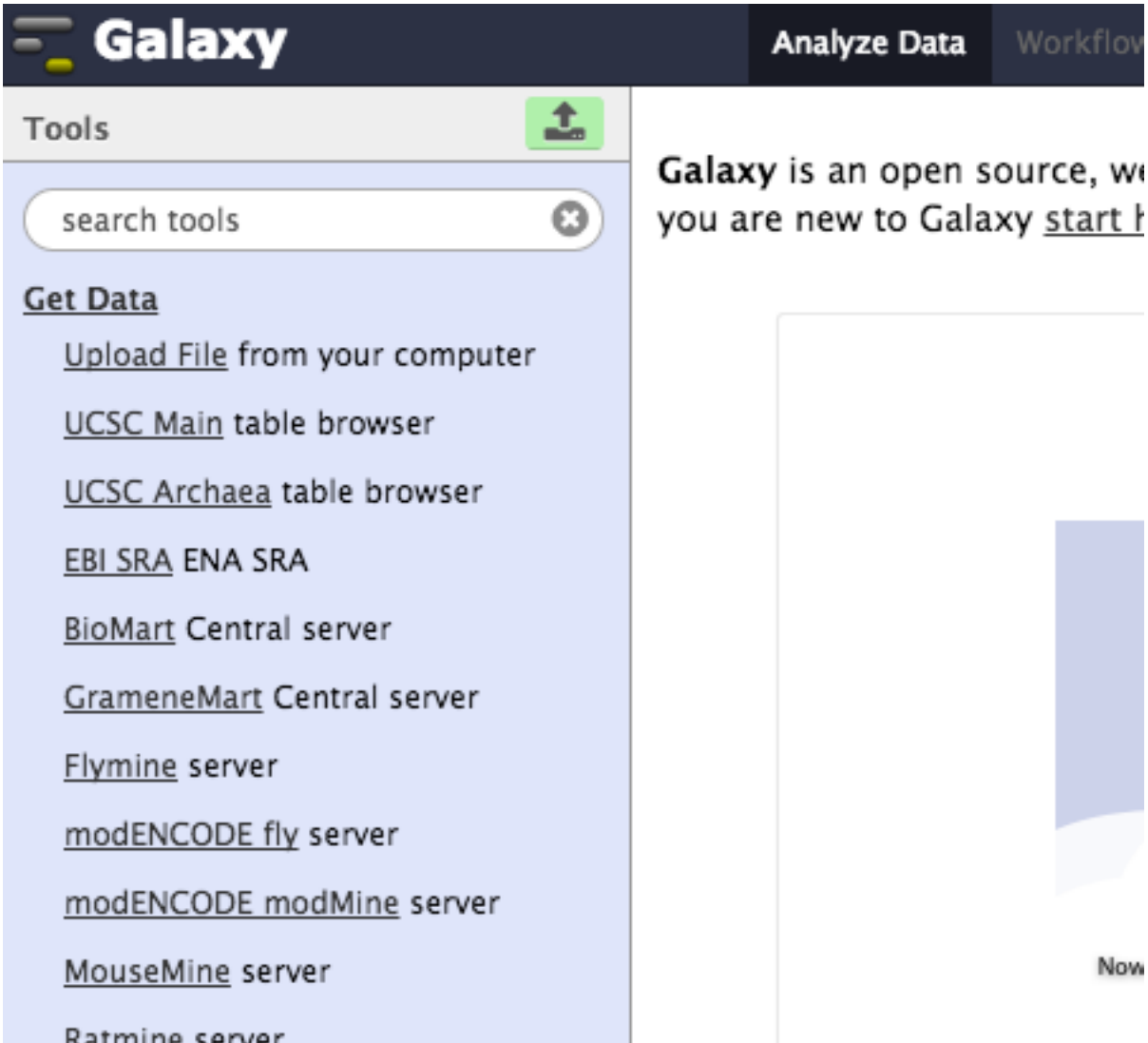
Under **Get Data** and **Upload File** in the “**Paste/Fetch Data**” box:

- NA12878_CBW_chr1_R1.fastq.gz
http://cbw##.dyndns.info/module2/NA12878_CBW_chr1_R1.fastq.gz
- NA12878_CBW_chr1_R2.fastq.gz
http://cbw##.dyndns.info/module2/NA12878_CBW_chr1_R2.fastq.gz
- hg19_chr1.fa
http://cbw##.dyndns.info/module7/hg19_chr1.fa

Alternatively, copy the files to your local machine, then upload using “**Choose Local File**”:


```
scp -i CBWNY.pem ubuntu@cbw##.dyndns.info:~/CourseData/HT_data/Module2/NA*78* .  
scp -i CBWNY.pem ubuntu@cbw##.dyndns.info:~/CourseData/HT_data/Module7/hg* .  
scp -i CBWNY.pem ubuntu@cbw49.dyndns.info:~/CourseData/HT_data/Module2/  
dbSNP_135_chr1.vcf.gz .
```


Load file(s) to Galaxy



The screenshot shows the Galaxy web interface. The top navigation bar includes the Galaxy logo, 'Analyze Data', and 'Workflow'. The left sidebar is titled 'Tools' and contains a search bar and a list of tool categories. Two arrows point to the 'Get Data' section, which is highlighted in light blue. The first arrow points to the 'Get Data' header, and the second arrow points to the 'Upload File from your computer' option. Other options in the 'Get Data' section include 'UCSC Main table browser', 'UCSC Archaea table browser', 'EBI SRA ENA SRA', 'BioMart Central server', 'GrameneMart Central server', 'Flymine server', 'modENCODE fly server', 'modENCODE modMine server', 'MouseMine server', and 'Ratmine server'. The main content area on the right displays a message: 'Galaxy is an open source, we you are new to Galaxy [start here](#)'. Below this message is a large, empty rectangular box with a light blue gradient.

Galaxy Analyze Data Workflow

Tools 

search tools 

→ **Get Data**

→ [Upload File](#) from your computer

[UCSC Main](#) table browser

[UCSC Archaea](#) table browser

[EBI SRA](#) ENA SRA

[BioMart](#) Central server

[GrameneMart](#) Central server

[Flymine](#) server

[modENCODE](#) fly server

[modENCODE](#) modMine server

[MouseMine](#) server

[Ratmine](#) server

Galaxy is an open source, we you are new to Galaxy [start here](#)

Now

Load file(s) to Galaxy

The screenshot shows the Galaxy web interface with a modal dialog box titled "Download data directly from web or upload files from your disk". The dialog box contains a large dashed box for file upload with the text "You can Drag & Drop files into this box." Below this, there are two callout boxes: "Choose local file (local copies)" and "Paste/Fetch data (URLs)". The dialog box also features a search bar, a "Type (set all):" dropdown menu set to "Auto-detect", a "Genome (set all):" dropdown menu set to "---- Additional Species ...", and a row of buttons: "Choose local file", "Paste/Fetch data", "Start", "Pause", "Reset", and "Close".

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Tools

search tools

Get Data

Upload File from

UCSC Main table

UCSC Archaea table

EBI SRA ENA SRA

BioMart Central S

GrameneMart Cer

Flymine server

modENCODE fly s

modENCODE mod

MouseMine serve

Ratmine server

YeastMine server

modENCODE wor

WormBase server

ZebrafishMine se

EuPathDB server

GenomeSpace Im

browser

Lift-Over

Text Manipulation

Convert Formats

Download data directly from web or upload files from your disk

You can Drag & Drop files into this box.

Choose local file (local copies)

Paste/Fetch data (URLs)

Type (set all): Auto-detect Q

Genome (set all): ---- Additional Species ...

Choose local file Paste/Fetch data Start Pause Reset Close

Galaxy Project @galaxyproject

GalaxyAdmins Meetup slides & video are now available: Carrie Ganote, @PTI_IU & @guerler, @galaxyproject bit.ly/gxyadm201504 #usegalaxy

44m

Load file(s) to Galaxy

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Using 0%

Tools

search to

Get Data

Upload File

UCSC Ma

UCSC An

EBI SRA

BioMart

Gramene

Flymine

modENC

modENC

MouseM

Ratmine

YeastMir

modENC

WormBas

Zebrafish

EuPathD

Genome browser

Download data directly from web or upload files from your disk

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

Name	Size	Type	Genome	Settings	Status
New File	-	Auto-det...	----- Additional S...		

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

Type (set all): Auto-detect

Genome (set all): ----- Additional Species ...

Choose local file Paste/Fetch data Start Pause Reset Close

Execute

Load file(s) to Galaxy

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Using 0 bytes

Tools

search tools

Get Data

Upload File from

UCSC Main table

UCSC Archaea tab

EBI SRA ENA SRA

BioMart Central s

GrameneMart Cer

Flymine server

modENCODE fly s

modENCODE mod

MouseMine serve

Ratmine server

YeastMine server

modENCODE wor

WormBase server

ZebrafishMine se

EuPathDB server

GenomeSpace im

browser

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Download data directly from web or upload files from your disk

You can Drag & Drop files into this box.

Name	Size	Type	Genome	Settings	Status
New File	0.2 KB	Auto-det...	----- Additional S...		100%

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

http://cbwXX.dyndns.info/module2/NA12878_CBW_chr1_R1.fastq.gz

http://cbwXX.dyndns.info/module2/NA12878_CBW_chr1_R2.fastq.gz

http://cbwXX.dyndns.info/module7/hq19_chr1.fa

Filename is the important identifier for subsequent slides / work

Type (set all): Auto-detect

Genome (set all): ----- Additional Species ...

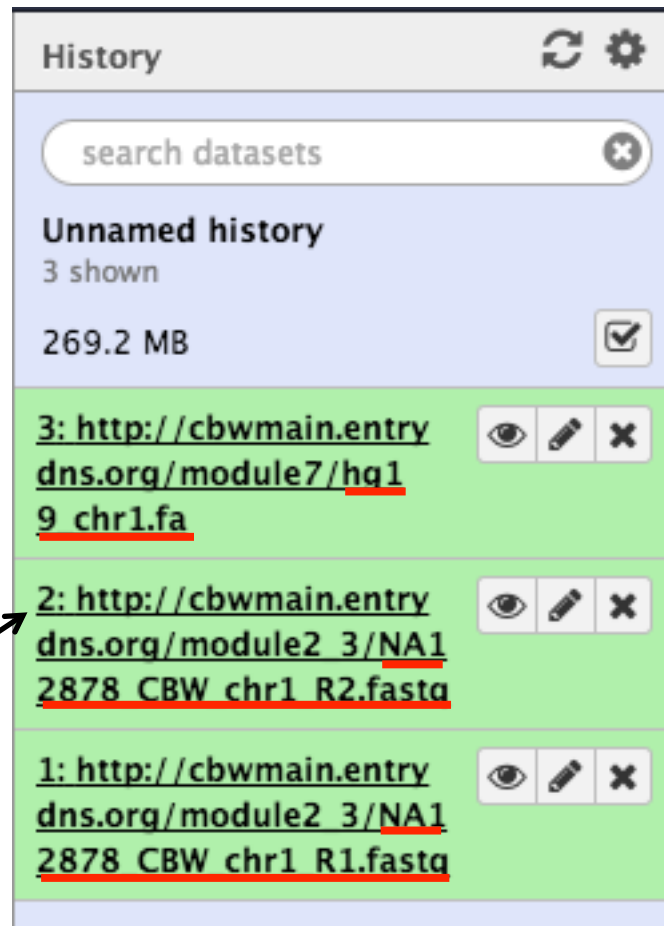
Choose local file Paste/Fetch data Start Pause Reset Close

Galaxy Project @galaxyproject 45m

GalaxyAdmins Meetup slides & video are now available: Carrie Ganote, @PTI_IU & @guerler, @galaxyproject bit.ly/gxyadm201504 #usegalaxy

Expand

Input data loaded (green)



Poke the eye

Edit attribute

Delete

Numbers and paths
may vary with usage

Poke the eye

Analyze Data

Workflow

Shared Data ▾

Visualization ▾

Help ▾

User ▾



This dataset is large and only the first megabyte is shown below.

[Show all](#) | [Save](#)

```
@SN1114:102:D16Y9ACXX:5:1107:11541:88384/1
AGATCCCTAGGACCTGGGGCAGCCCAGTTCGCAGTGGGCATCTGGTCCCTGAGCAACTGCTGGTCAGGTGCTGATGATGTA
+
CCCFHHHHHHHJIIJGIIIIJJJJFHJIIJGHHIIIIJIIJIGGIIJJIGGG)=A?EEHH;@B@D;@ACEECCDDCDAL
@SN1114:102:D16Y9ACXX:6:1211:16647:74484/1
TCTTGCACGCTGCTAGATCCCTAGGACCTGGGGCAGCCCAGTTCGCAGTGGGCATCTGGTCCCTGAGCATCTGCTGGTCAGC
+
?@@FFFFFHFFHHIJJFBHIJIIJEGHEGHEHGGGGGGJJGBFDFGGI;FHGIJF>>EEACF@DEDEECCACCCCCACCD<
@SN1114:102:D16Y9ACXX:6:1114:17422:19388/1
CCAGGTACTGTCATAGGGACTGTCCCGCCTCCTTAGCATACCACCTCCCATCCCTTCCACCACACAAGGGACAACCACGTCC
+
B@@FFDEEHDBHHJJJJIEHJJIIJJIIHIIJI>DAGHIGIGCEHJIIJJB>GAEGIJIEGG=@EEHHFB BBBABCA==??<<?
@SN1114:102:D16Y9ACXX:5:1310:5270:46750/1
TAGGCTTTTGCCCAGGTACTGTCATAGGGACTGTCCCGCCTCCTTAGCATACCACCTCCCATCCCTTCCACCACACAAGGGF
+
@@@=DDEFHHAFDGHIFGGGIHIGIJJGGGIHGIJJGIIJIGJIIHGIG@BCFFHEHIJIIIIJHFAE93?B(;=?AB=25;
@SN1114:102:D16Y9ACXX:4:2208:19114:5975/1
CTAGGCTTTTGCCCAGGTACTGTCATAGGGACTGTCCCGCCTCCTTAGCATACCACCTCCCATCCCTTCCACCACACAAGGG
+
```


Edit attributes

Analyze Data Workflow Shared Data Visualization Cloud Help User

Attributes Convert Format Datatype Permissions

Edit Attributes

Name:

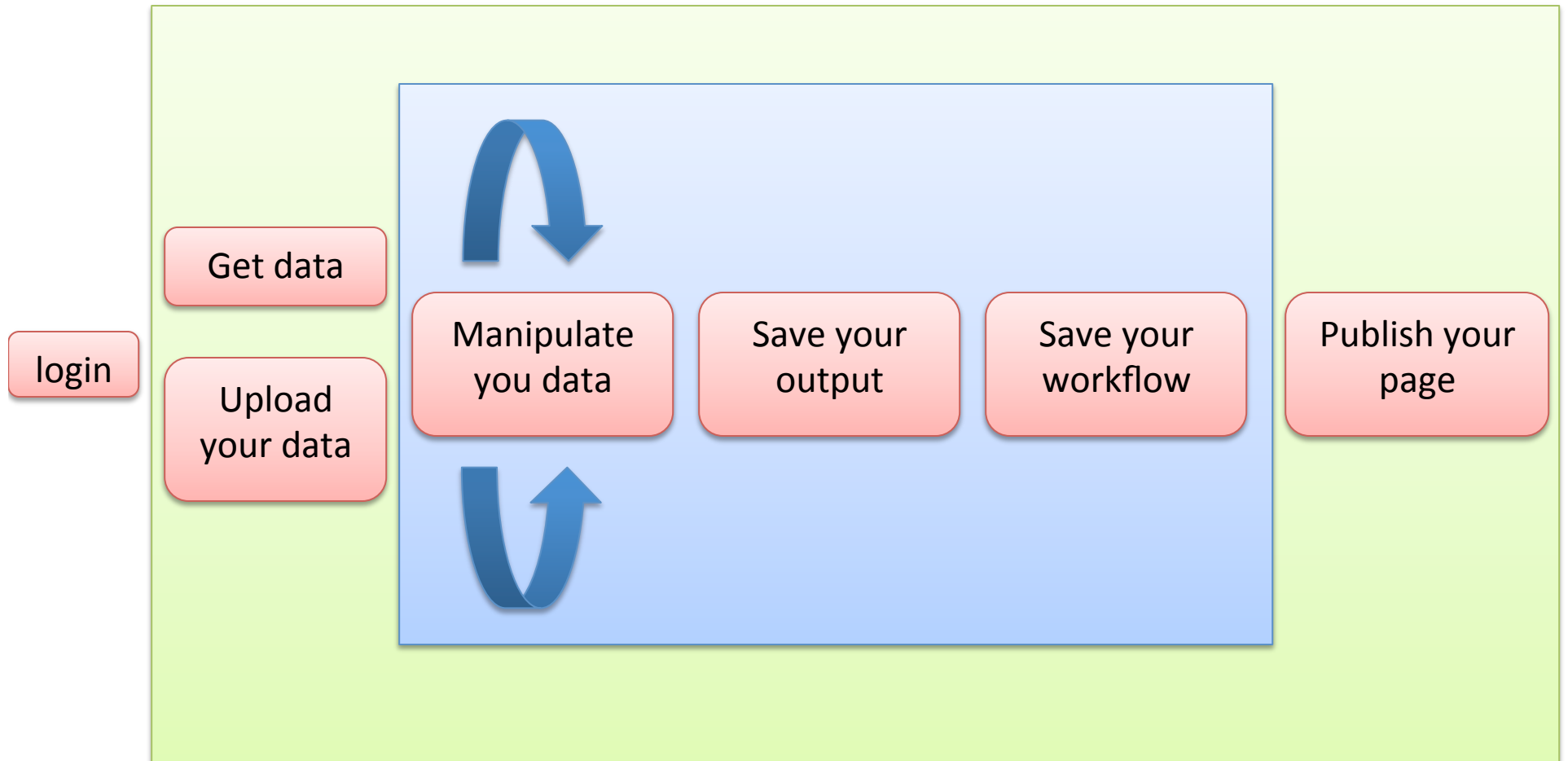
Info:

Database/Build:

This will inspect the dataset and attempt to correct the above column values if they are not accurate.

You can give a shorter meaningful name to your files

General workflow for Galaxy



Use FASTQ Groomer on the 2 fastq files

→ NGS: QC and manipulation


ILLUMINA FASTQ

→ FASTQ Groomer convert between various FASTQ quality formats




FASTQ splitter on joined paired end reads

FASTQ joiner on paired end reads

FASTQ Summary Statistics by column

 FASTQ Groomer convert between various FASTQ quality formats (Galaxy Tool Version 1.0.4) Options

File to groom

   5: http://cbwmain.entrydns.org/module2_3/NA12878_CBW_chr1_R2.fastq


Input FASTQ quality scores type

Sanger & Illumina 1.8+


Advanced Options

Hide Advanced Options

✓ Execute

 **Galaxy**

Analyze DataWorkflowShared Data ▾VisualizationCloud ▾Help ▾User ▾

Tools

[FASTQ Groomer](#) convert between various FASTQ quality formats

[Manipulate FASTQ](#) reads on various attributes

[FASTQ Masker](#) by quality score

[FASTQ joiner](#) on paired end reads

[FASTQ splitter](#) on joined paired end reads

[FASTQ Summary Statistics](#) by column

[FASTQ to FASTA](#) converter


[FASTQ to Tabular](#) converter

[FASTQ Trimmer](#) by column

[FASTQ Quality Trimmer](#) by sliding window



[Tabular to FASTQ](#) converter


[Convert SOLiD output](#) to fastq

 1 job has been successfully added to the queue – resulting in the following datasets:


4: FASTQ Groomer on data 1




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  




search datasets 

Unnamed history
3 shown

269.2 MB 

3: http://cbwmain.entry.dns.org/module7/hq19_chr1.fa   

2: [http://cbwmain.entry.dns.org/module2_3/NA12878 CBW chr1 R2.fastq](http://cbwmain.entry.dns.org/module2_3/NA12878_CBW_chr1_R2.fastq)   

1: [http://cbwmain.entry.dns.org/module2_3/NA12878 CBW chr1 R1.fastq](http://cbwmain.entry.dns.org/module2_3/NA12878_CBW_chr1_R1.fastq)   



History  

search datasets 

Unnamed history
5 shown

296.0 MB 

5: FASTQ Groomer on data 2   

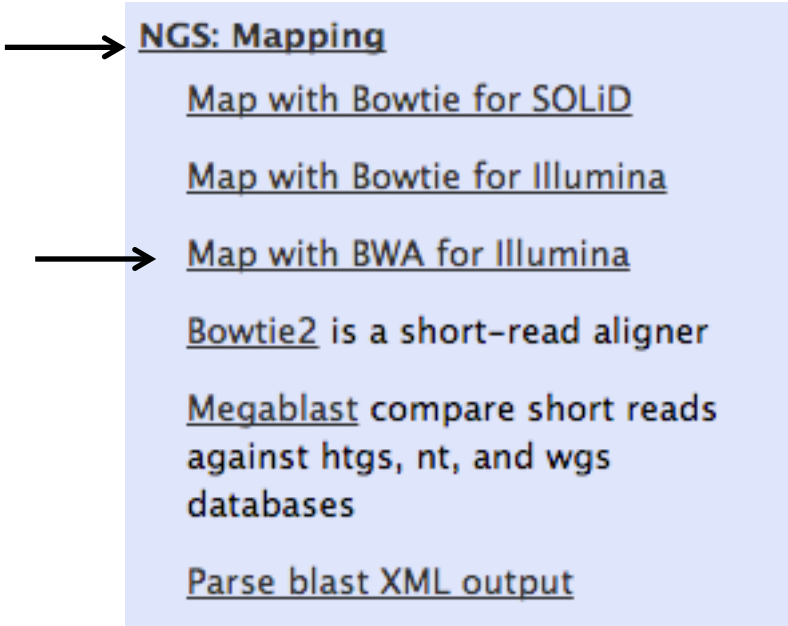
4: FASTQ Groomer on data 1   

3: http://cbwmain.entry.dns.org/module7/hq19_chr1.fa   

2: [http://cbwmain.entry.dns.org/module2_3/NA12878 CBW chr1 R2.fastq](http://cbwmain.entry.dns.org/module2_3/NA12878_CBW_chr1_R2.fastq)   

1: [http://cbwmain.entry.dns.org/module2_3/NA12878 CBW chr1 R1.fastq](http://cbwmain.entry.dns.org/module2_3/NA12878_CBW_chr1_R1.fastq)   

Alignment with BWA



NGS: Mapping

Map with Bowtie for SOLiD

Map with Bowtie for Illumina

Map with BWA for Illumina

Bowtie2 is a short-read aligner

Megablast compare short reads
against htgs, nt, and wgs
databases

Parse blast XML output

Alignment with BWA

Map with BWA for Illumina (Galaxy Tool Version 1.2.3) Options

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

Use one from the history

Select a reference from history

3: http://cbwmain.entrydns.org/module7/hg19_chr1.fa

Is this library mate-paired?

Paired-end

Forward FASTQ file

4: FASTQ Groomer on data 1

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

Reverse FASTQ file

5: FASTQ Groomer on data 2

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

Specify the read group for this file? (samse/sampe -r)

Yes

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

NA12878

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

Library name (LB)

LIB1

Required if RG specified

Platform/technology used to produce the reads (PL)

ILLUMINA

Required if RG specified. Valid values : CAPILLARY, LS454, ILLUMINA, SOLID, HELICOS, IONTORRENT and PACBIO

Sample (SM)

NA12878



Required if RG specified. Use pool name where a pool is being sequenced





1 job has been successfully added to the queue – resulting in the following datasets:





6: Map with BWA for Illumina on data 5, data 4, and data 3: mapped reads




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  

search datasets 



Unnamed history
6 shown
296.0 MB 


 6: Map with BWA for I   
llumina on data 5, data
4, and data 3: mapped reads


5: FASTQ Groomer on da   
ta 2




4: FASTQ Groomer on da   
ta 1




This step will take some time to run....




History  

search datasets 

Unnamed history
6 shown
336.1 MB 

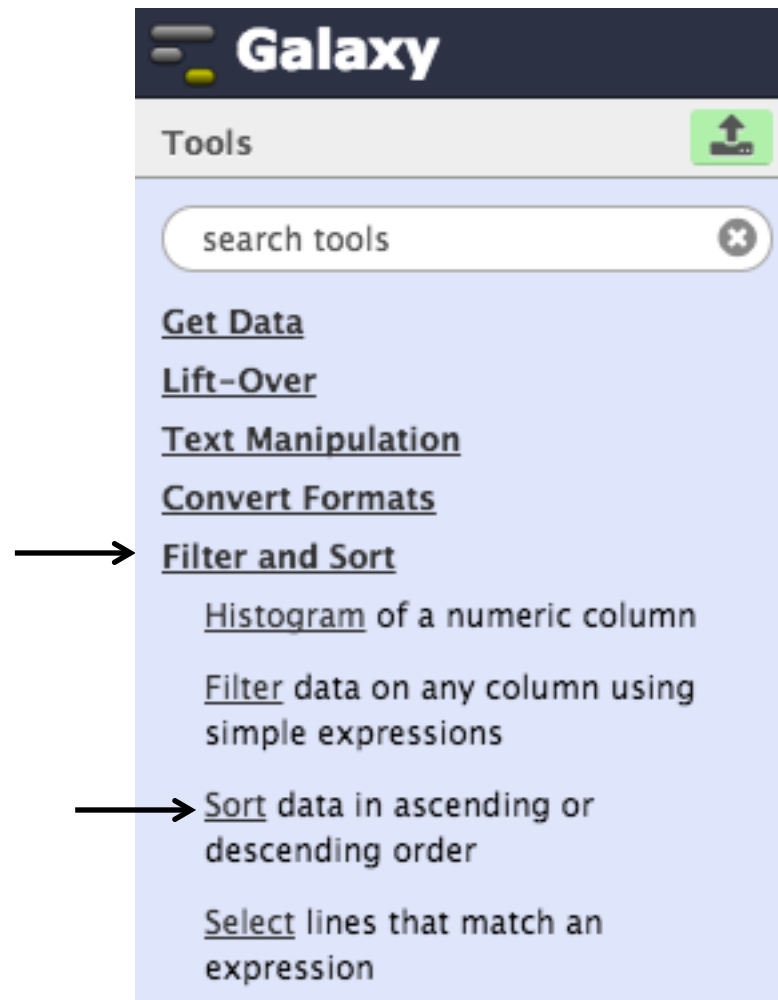
6: Map with BWA for Illu   
mina on data 5, data 4,
and data 3: mapped reads

5: FASTQ Groomer on da   
ta 2

4: FASTQ Groomer on da   
ta 1




QNAME	FLAG	RNAME	POS	MAPQ	CIGAR
@SQ SN:chr1 LN:249250621					
@RG ID:NA12878 LB:LIB1 PL:ILLUMINA SM:NA12878					
@PG ID:bwa PN:bwa VN:0.5.9-r16					
SN1114:102:D16Y9ACXX:5:1107:11541:88384	99	chr1	17704814	60	101M
SN1114:102:D16Y9ACXX:5:1107:11541:88384	147	chr1	17704937	60	101M
SN1114:102:D16Y9ACXX:6:1211:16647:74484	99	chr1	17704800	60	101M
SN1114:102:D16Y9ACXX:6:1211:16647:74484	147	chr1	17704963	60	101M
SN1114:102:D16Y9ACXX:6:1114:17422:19388	83	chr1	17704968	60	101M
SN1114:102:D16Y9ACXX:6:1114:17422:19388	163	chr1	17704782	60	101M
SN1114:102:D16Y9ACXX:5:1310:5270:46750	83	chr1	17704979	60	101M
SN1114:102:D16Y9ACXX:5:1310:5270:46750	163	chr1	17704867	60	101M

Sort the SAM file



🔧 Sort data in ascending or descending order (Galaxy Tool Version 1.0.3) ▼ Options

Sort Dataset

→    6: Map with BWA for Illumina on data 5, data 4, and data 3: mapped reads ▼

on column

→ Column: 3 Column 3: chromosome ▼

with flavor

→ Alphabetical sort ▼

everything in

→ Ascending order ▼




Column selection

→ + Insert Column selection Add new column selection

✓ Execute

🔧 Sort data in ascending or descending order (Galaxy Tool Version 1.0.3) ▼ Options

Sort Dataset

→    6: Map with BWA for Illumina on data 5, data 4, and data 3: mapped reads ▼

on column

→ Column: 3 ▼


with flavor

→ Alphabetical sort ▼

everything in

→ Ascending order ▼

Column selection

1: Column selection 

→ **on column**

Column: 4 ▼

with flavor


→ Numerical sort ▼

everything in



→ Ascending order ▼

Column 4: position

+ Insert Column selection

 **Sort data in ascending or descending order (Galaxy Tool Version 1.0.3)** ▼ Options

Sort Dataset

   6: Map with BWA for Illumina on data 5, data 4, and data 3: mapped reads ▼

on column

Column: 3 ▼


with flavor

Alphabetical sort ▼

everything in

Ascending order ▼

Column selection

1: Column selection 

on column


Column: 4 ▼


with flavor

Numerical sort ▼

everything in

Ascending order ▼

 Insert Column selection



 **Execute**





1 job has been successfully added to the queue – resulting in the following datasets:




7: Sort on data 6




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  

search datasets 

Unnamed history
7 shown
376.3 MB 

7: Sort on data 6   

6: Map with BWA for Illumina on data 5, data 4, and data 3: mapped reads   

Remove duplicated reads



NGS: SAM Tools

BAM to fastq Convert BAM file to fastq

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

Remove duplicated reads

Filter SAM (version 1.0.0)

Select dataset to filter:

7: Sort on data 6

Flags

Flag 1

Type:

The read is a PCR or optical duplicate

Set the states for this flag:

☒ No

☐ Yes

Remove Flag 1

Add new Flag

Execute

Add header to sorted and filtered SAM file

Filter and Sort

Filter data on any column using simple expressions




Sort data in ascending or descending order

Select lines that match an expression

- 1) Select the header lines
- 2) Concatenate the header lines and the sorted/filtered SAM file

🔑 Select lines that match an expression (Galaxy Tool Version 1.0.1) Options

Select lines from

→    6: Map with BWA for Illumina on data 5, data 4, and data 3: mapped reads

that

→ Matching

the pattern

→ ^@

here you can enter text or regular expression (for syntax check lower part of this frame)

→



Add header to sorted and filtered SAM file

→ Text Manipulation


- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Cut columns from a table
- Merge Columns together

🔧 Concatenate datasets tail-to-head (Galaxy Tool Version 1.0.0) ▼ Options



Concatenate Dataset

→   9: Select on data 6 ▼


Dataset



1: Dataset 


Select


→   8: Filter SAM on data 7 ▼




+ Insert Dataset




→ 




History  




search datasets 




Unnamed history
10 shown
416.4 MB 

10: Concatenate datasets on data 9 and data 8   

9: Select on data 6   

8: Filter SAM on data 7   

7: Sort on data 6   

6: Map with BWA for Illumina on data 5, data 4, and data 3: mapped reads   



Input for GATK
RealignerTargetCreator

Create an interval

→ **Text Manipulation**

- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Cut columns from a table
- Merge Columns together
- Convert delimiters to TAB
- Create single interval as a new dataset
- Change Case of selected columns

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

Chromosome
chr1

Start position
17704860

End position
18004860

Name
Interval_chr1




Strand
plus
If your interval is strandless set strand to plus




→ **✓ Execute**

GATK RealignerTargetCreator


→ NGS: GATK Tools (beta)

- Validate Variants
- Select Variants from VCF files
- Variant Recalibrator
- Variant Filtration on VCF files
- Eval Variants
- Combine Variants
- Apply Variant Recalibration
- Variant Annotator
- Unified Genotyper SNP and indel caller
- Realigner Target Creator for use in local realignment

11: Create single interval   

10: Concatenate datasets on data 9 and data 8   




GATK RealignerTargetCreator

 Realigner Target Creator for use in local realignment (Galaxy Tool Version 0.0.4) Options

Choose the source for the reference list




History

BAM file

   10: Concatenate datasets on data 9 and data 8


-I,--input_file <input_file>

Using reference file

   3: http://cbwmain.entrydns.org/module7/hg19_chr1.fa

-R,--reference_sequence <reference_sequence>

Binding for reference-ordered data

 Insert Binding for reference-ordered data

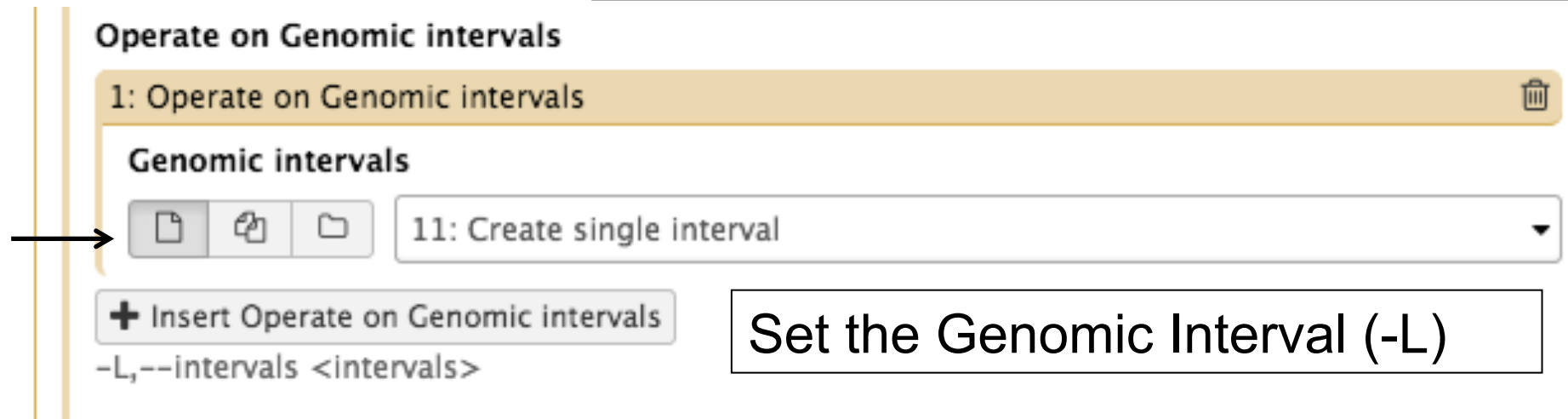
-known,--known <known>

Basic or Advanced GATK options

Advanced

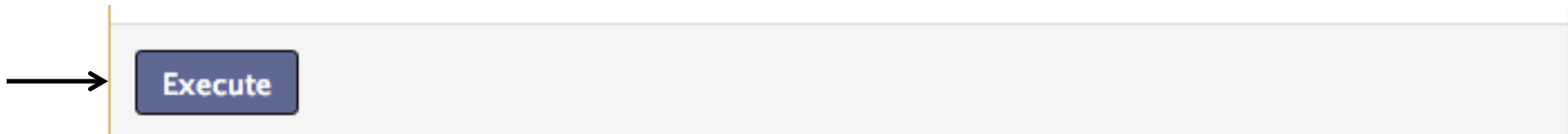
GATK RealignerTargetCreator

Add new Operate on Genomic intervals



The screenshot shows the 'Operate on Genomic intervals' section of the GATK RealignerTargetCreator interface. It features a list of operations, with '1: Operate on Genomic intervals' selected. Below this, there are three icons (document, copy, folder) and a dropdown menu showing '11: Create single interval'. An arrow points to the first icon. Below the icons is a button labeled '+ Insert Operate on Genomic intervals' and a text input field containing '-L,--intervals <intervals>'. A text box to the right of the button says 'Set the Genomic Interval (-L)'.

Keep default values for other parameters



The screenshot shows the bottom of the GATK RealignerTargetCreator interface, featuring a large 'Execute' button. An arrow points to the button.

GATK RealignerTargetCreator



1 job has been successfully added to the queue – resulting in the following datasets:

12: Realigner Target Creator on data 3, data 10, and data 11 (GATK intervals)

13: Realigner Target Creator on data 3, data 10, and data 11 (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.

GATK IndelRealigner

→ NGS: GATK Tools (beta)

Validate Variants

Select Variants from VCF files

Variant Recalibrator

Variant Filtration on VCF files

Eval Variants

Combine Variants

Apply Variant Recalibration




Variant Annotator




Unified Genotyper SNP and indel caller

Realigner Target Creator for use in local realignment

Print Reads from BAM files

→ Indel Realigner – perform local realignment

13: Realigner Target   
Creator on data 3, data 10, and data 11 (log)

12: Realigner Target   
Creator on data 3, data 10, and data 11 (GATK intervals)

Indel Realigner – perform local realignment (Galaxy Tool Version 0.0.6) Options

Choose the source for the reference list

History

BAM file

10: Concatenate datasets on data 9 and data 8

-l,--input_file <input_file>

Using reference file

3: http://cbwmain.entrydns.org/module7/hg19_chr1.fa

-R,--reference_sequence <reference_sequence>

Restrict realignment to provided intervals

12: Realigner Target Creator on data 3, data 10, and data 11 (GATK inter...

-targetIntervals,--targetIntervals <targetIntervals>

Binding for reference-ordered data

+ Insert Binding for reference-ordered data

-known,--knownAlleles <knownAlleles>

LOD threshold above which the realigner will proceed to realign

5

-LOD,--LODThresholdForCleaning <LODThresholdForCleaning>

Use only known indels provided as RODs

Yes No

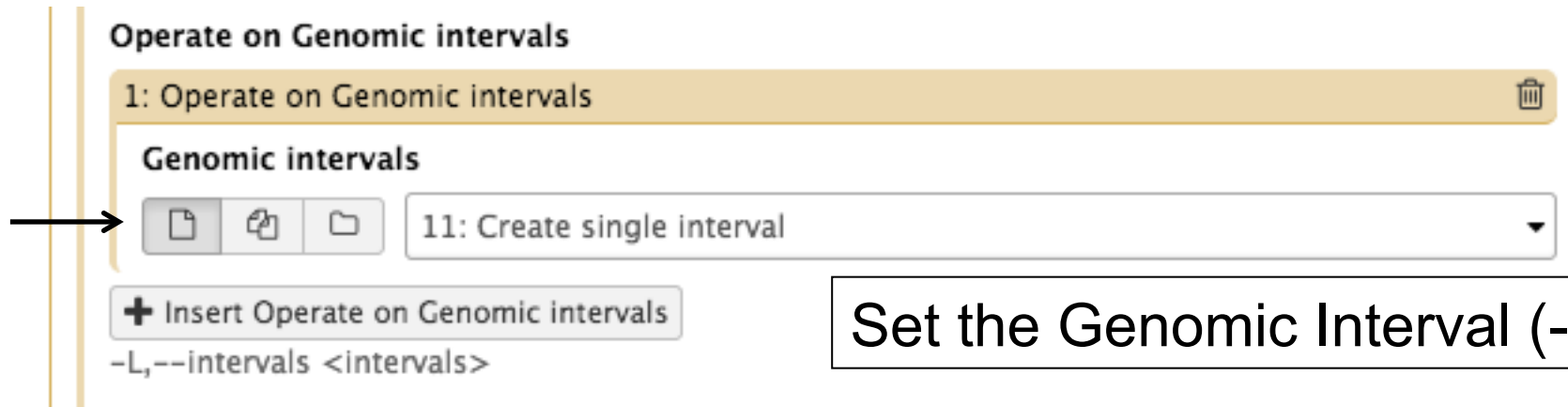
-knownsOnly

Basic or Advanced GATK options

Advanced

GATK IndelRealigner

Add new Operate on Genomic intervals



Operate on Genomic intervals

1: Operate on Genomic intervals

Genomic intervals

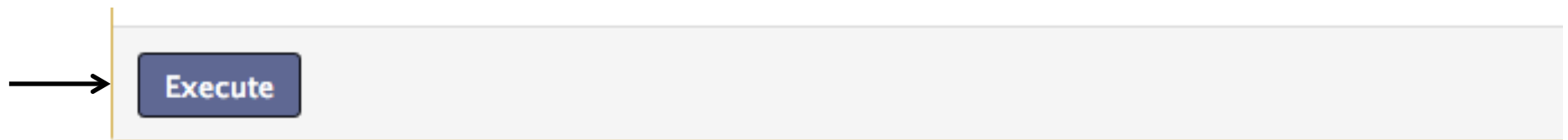
11: Create single interval

+ Insert Operate on Genomic intervals

-L,--intervals <intervals>

Set the Genomic Interval (-L)

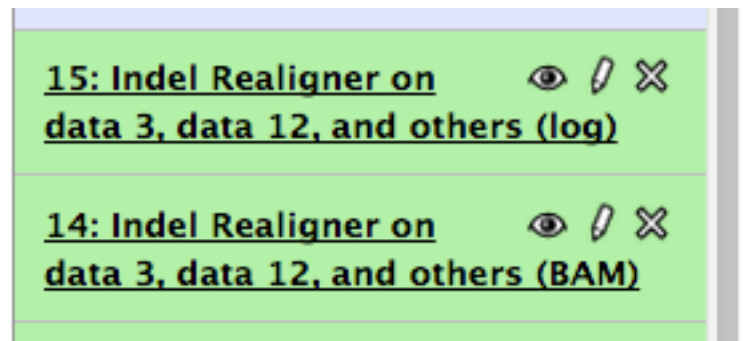
Keep default values for other parameter



Execute

GATK UnifiedGenotype

- NGS: GATK Tools (beta)
 - Validate Variants
 - Select Variants from VCF files
 - Variant Recalibrator
 - Variant Filtration on VCF files
 - Eval Variants
 - Combine Variants
 - Apply Variant Recalibration
 - Variant Annotator
 - Unified Genotyper SNP and indel caller



Choose the source for the reference list

History

BAM file

1: BAM file

BAM file

14: Indel Realigner on data 3, data 12, and others (BAM)

+ Insert BAM file

-I,--input_file <input_file>

Using reference file

3: http://cbwmain.entrydns.org/module7/hg19_chr1.fa

-R,--reference_sequence <reference_sequence>

Binding for reference-ordered data

+ Insert Binding for reference-ordered data

-D,--dbsnp <dbsnp>

Genotype likelihoods calculation model to employ

BOTH

-glm,--genotype_likelihoods_model <genotype_likelihoods_model>

The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be called

30

-stand_call_conf,--standard_min_confidence_threshold_for_calling <standard_min_confidence_threshold_for_calling>

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)

10

-stand_emit_conf,--standard_min_confidence_threshold_for_emitting <standard_min_confidence_threshold_for_emitting>

Basic or Advanced GATK options:

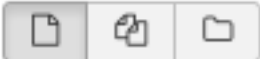
Advanced

Add new Operate on Genomic intervals

Operate on Genomic intervals

1: Operate on Genomic intervals

Genomic intervals





11: Create single interval


+ Insert Operate on Genomic intervals


-L,--intervals <intervals>




Set the Genomic Interval (-L)




Execute




History  

search datasets 

Unnamed history
18 shown, 1 [hidden](#)
477.5 MB 

18: Unified Genotyper on data 3, data 14, and data 11 (log)   

17: Unified Genotyper on data 3, data 14, and data 11 (metrics)   

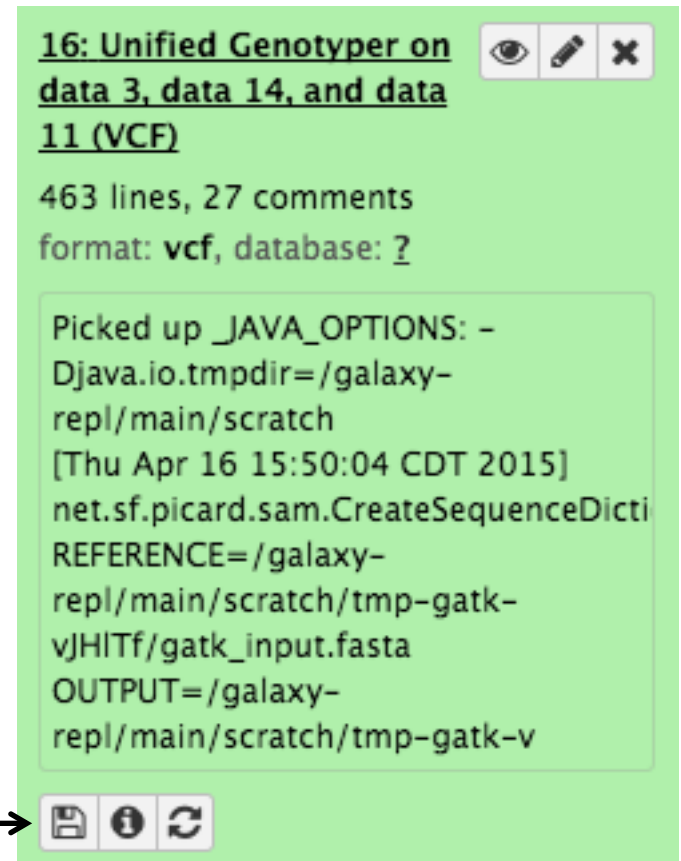
16: Unified Genotyper on data 3, data 14, and data 11 (VCF)   

← vcf file, will require further filtering and annotations (and can be downloaded)

GATK VariantFiltration & VariantAnnotator

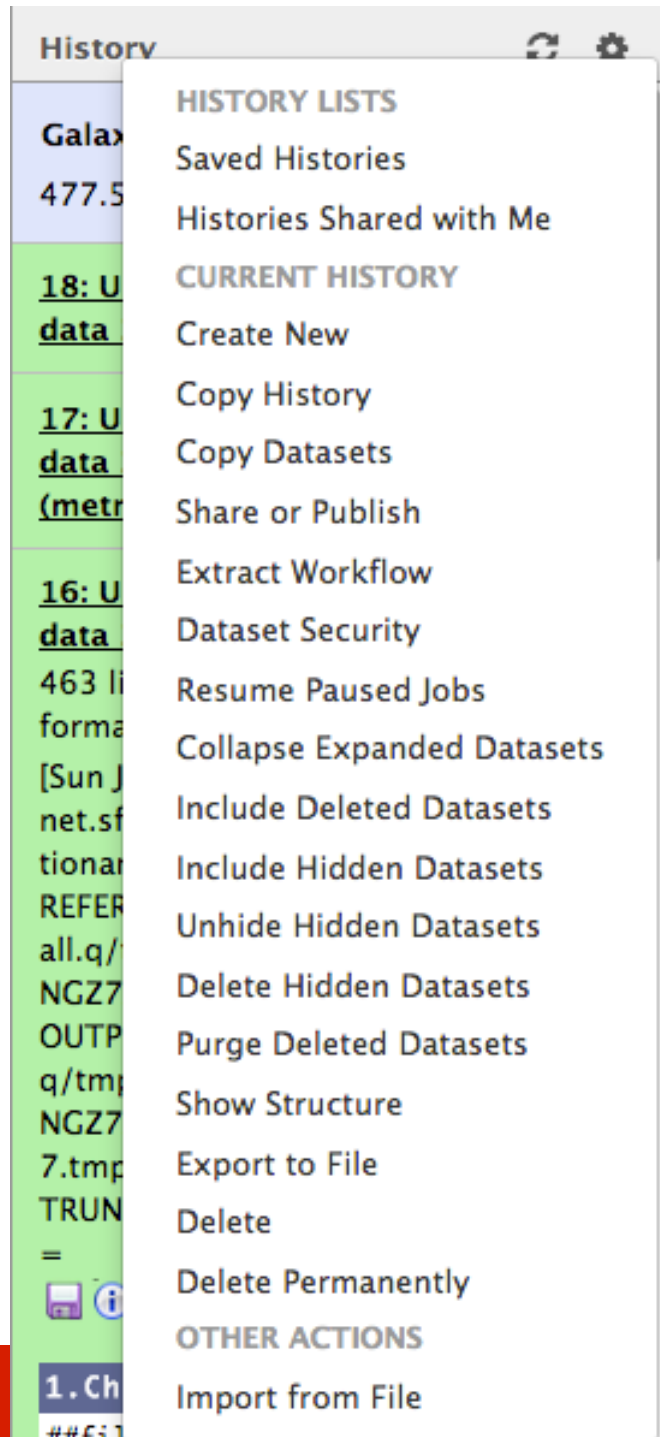
- Tools not available on the Galaxy Cloud instance (for the moment)
- Present on usegalaxy.com under NGS: GATK Tools (beta)

Save your vcf file on your computer



The screenshot shows a Galaxy tool interface for 'Unified Genotyper on data 3, data 14, and data 11 (VCF)'. The interface has a light green background. At the top, the tool name is displayed in a dark font, followed by '463 lines, 27 comments' and 'format: vcf, database: ?'. Below this, a text box contains the following text: 'Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/galaxy-repl/main/scratch [Thu Apr 16 15:50:04 CDT 2015] net.sf.picard.sam.CreateSequenceDicti REFERENCE=/galaxy-repl/main/scratch/tmp-gatk-vJHITf/gatk_input.fasta OUTPUT=/galaxy-repl/main/scratch/tmp-gatk-v'. At the bottom of the interface, there are three icons: a floppy disk (save), an information icon (i), and a refresh icon (circular arrow).

```
16: Unified Genotyper on  
data 3, data 14, and data  
11 (VCF)  
463 lines, 27 comments  
format: vcf, database: ?  
  
Picked up _JAVA_OPTIONS: -  
Djava.io.tmpdir=/galaxy-  
repl/main/scratch  
[Thu Apr 16 15:50:04 CDT 2015]  
net.sf.picard.sam.CreateSequenceDicti  
REFERENCE=/galaxy-  
repl/main/scratch/tmp-gatk-  
vJHITf/gatk_input.fasta  
OUTPUT=/galaxy-  
repl/main/scratch/tmp-gatk-v
```



- Share history with neighbor
- Extract workflow

sharing


Saved Histories

search history names and tags

[Advanced Search](#)

<input type="checkbox"/>	<u>Name</u>	<u>Datasets</u>	<u>Tags</u>	<u>Sharing</u>	<u>Size on Disk</u>	<u>Created</u>	<u>Last Updated</u> ↑	<u>Status</u>
<input type="checkbox"/>	Galaxy_lab	18	0 Tags		477.5 MB	~ 12 hours ago	~ 10 hours ago	current history
<input type="checkbox"/>	t		0 Tags		1.0 GB	2 days ago	1 day ago	
	For	name	Delete	Delete Permanently	Undelete	Histories more than a time period specified by the Galaxy administrator(s) may be permanently deleted.		

Work on the workflow

 Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Using 1.5 GB

Your workflows

Create new workflow Upload or import workflow

Name	# of Steps
CBW_Galaxy_job_workflow	15
Module2_3_H	0

Workflow

No workflows have been shared with you by others

Other options

Configure your workflow

Edit

Run

Share or Publish

Download or Export

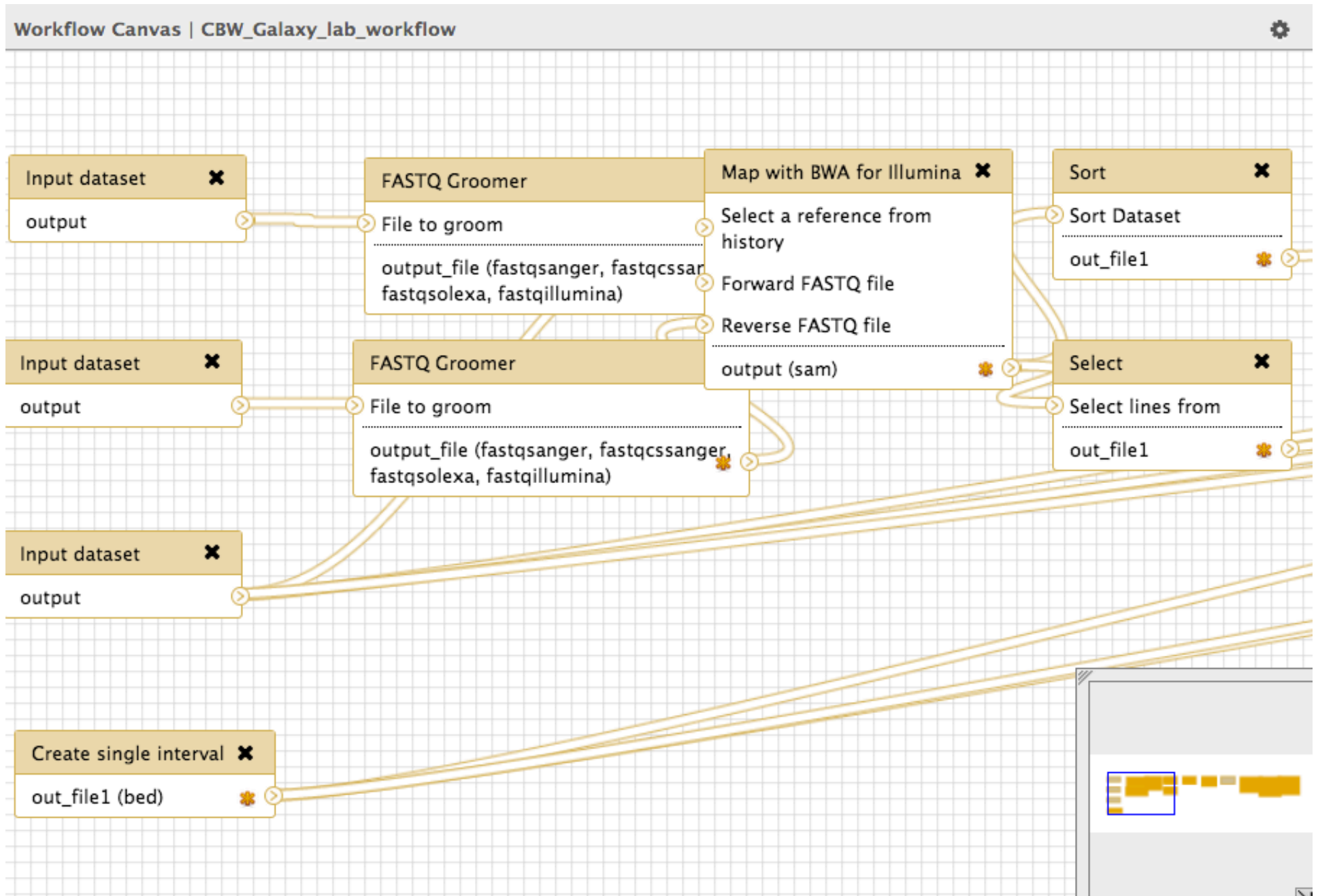
Copy

Rename





View

Delete













- Edit workflow



You can Drag & Drop files into this box.

Name	Size	Type	Genome	Settings	Status
 New File	59 b	Auto-det... ▼	 ----- Additional S... ▼		100% 
You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.					
<div>http://cbwXX.dyndns.info/module2/<u>dbSNP_135_chr1.vcf.gz</u></div>					

Input files

<u>19: http://cbwmain.entrydns.org/module2_3/db_SNP_135_chr1.vcf</u>	  
<u>18: Unified Genotyper on data 3, data 14, and data 11 (log)</u>	  
<u>17: Unified Genotyper on data 3, data 14, and data 11 (metrics)</u>	  
<u>16: Unified Genotyper on data 3, data 14, and data 11 (VCF)</u>	  
463 lines, 27 comments format: vcf, database: ?	

Use GATK Tools

→	<u>NGS: GATK Tools (beta)</u>
	<u>Validate Variants</u>
	<u>Select Variants</u> from VCF files
	<u>Variant Recalibrator</u>
→	<u>Variant Filtration</u> on VCF files
	<u>Eval Variants</u>
	<u>Combine Variants</u>
	<u>Apply Variant Recalibration</u>
	<u>Variant Annotator</u>
	<u>Unified Genotyper</u> SNP and indel caller

NOTE: The best practice when using GATK is to use the VariantRecalibrator. In our data set, we had too few variants to accurately use the variant recalibrator and therefore we used the VariantFiltration tool instead.

GATK VariantFiltration

Variant Filtration on VCF files (Galaxy Tool Version 0.0.5) Options

Choose the source for the reference list

History

Variant file to annotate

16: Unified Genotyper on data 3, data 14, and data 11 (VCF)

-V,--variant <variant>

Using reference file

3: http://cbwmain.entrydns.org/module7/hg19_chr1.fa

-R,--reference_sequence <reference_sequence>

Variant Filters

+ Insert Variant Filters

The variant filters we want to set here are:

```
java -Xmx2g -jar /usr/local/GATK/GenomeAnalysisTK.jar -T VariantFiltration -R hg19.fa --variant  
NA12878.vcf -o NA12878_filter.vcf --filterExpression "QD < 2.0" --filterExpression "FS > 200.0" --  
filterExpression "MQ < 40.0" --filterName QDFilter --filterName FSFilter --filterName MQFilter
```

GATK VariantFiltration

The screenshot shows the 'Variant Filters' section of the GATK VariantFiltration tool. It contains a list of filters, currently with one filter named 'QDFilter'. The filter's expression is 'QD < 2.0'. Below the expression, there is a note: 'JEXL formatted expressions (-filter, --filterExpression <filterExpression>)'. The filter name is 'QDFilter', with a note: '-filterName, --filterName <filterName>'. There are two radio buttons for 'Use filter at the individual sample level', with 'Yes' selected. Below this, there is a note: 'Use -G_filter, --genotypeFilterExpression <genotypeFilterExpression> and -G_filterName, --genotypeFilterName <genotypeFilterName> for filter type'. At the bottom, there is a button labeled '+ Insert Variant Filters'.

Variant Filters

1: Variant Filters

Filter expression

QD < 2.0

JEXL formatted expressions (-filter, --filterExpression <filterExpression>)

Filter name

QDFilter

-filterName, --filterName <filterName>

Use filter at the individual sample level

Yes No

Use -G_filter, --genotypeFilterExpression <genotypeFilterExpression> and -G_filterName, --genotypeFilterName <genotypeFilterName> for filter type

+ Insert Variant Filters

The variant filters we want to set here are:

```
java -Xmx2g -jar /usr/local/GATK/GenomeAnalysisTK.jar -T VariantFiltration -R hg19.fa --variant NA12878.vcf -o NA12878_filter.vcf --filterExpression "QD < 2.0" --filterExpression "FS > 200.0" --filterExpression "MQ < 40.0" --filterName QDFilter --filterName FSFilter --filterName MQFilter
```

GATK VariantFiltration

The screenshot shows the '2: Variant Filters' panel in the GATK interface. It contains three main sections: 'Filter expression', 'Filter name', and 'Use filter at the individual sample level'. The 'Filter expression' field contains 'FS > 200.0'. The 'Filter name' field contains 'FSFilter'. The 'Use filter at the individual sample level' section has 'Yes' and 'No' buttons, with 'Yes' selected. Below these fields is a button labeled '+ Insert Variant Filters'. Arrows point to each of these elements: a black arrow to the filter expression, a black arrow to the filter name, and a red arrow to the 'Insert Variant Filters' button.

2: Variant Filters

Filter expression

FS > 200.0

JEXL formatted expressions (-filter,--filterExpression <filterExpression>)

Filter name

FSFilter

-filterName,--filterName <filterName>

Use filter at the individual sample level

Yes No

Use -G_filter,--genotypeFilterExpression <genotypeFilterExpression> and -G_filterName,--genotypeFilterName <genotypeFilterName> for filter type

+ Insert Variant Filters

The variant filters we want to set here are:

```
java -Xmx2g -jar /usr/local/GATK/GenomeAnalysisTK.jar -T VariantFiltration -R hg19.fa --variant NA12878.vcf -o NA12878_filter.vcf --filterExpression "QD < 2.0" --filterExpression "FS > 200.0" --filterExpression "MQ < 40.0" --filterName QDFilter --filterName FSFilter --filterName MQFilter
```

GATK VariantFiltration

The screenshot shows the '3: Variant Filters' panel in the GATK interface. It contains three main sections: 'Filter expression', 'Filter name', and 'Use filter at the individual sample level'. The 'Filter expression' field contains 'MQ < 40.0'. The 'Filter name' field contains 'MQFilter'. The 'Use filter at the individual sample level' section has 'Yes' and 'No' buttons, with 'Yes' selected. Below these fields is a text box with instructions: 'Use -G_filter,--genotypeFilterExpression <genotypeFilterExpression> and -G_filterName,--genotypeFilterName <genotypeFilterName> for filter type'. At the bottom of the panel is an 'Execute' button. Arrows point to the 'Filter expression' field, the 'Filter name' field, and the 'Execute' button.

3: Variant Filters

Filter expression

MQ < 40.0

JEXL formatted expressions (-filter,--filterExpression <filterExpression>)

Filter name

MQFilter

-filterName,--filterName <filterName>

Use filter at the individual sample level

Yes No

Use -G_filter,--genotypeFilterExpression <genotypeFilterExpression> and -G_filterName,--genotypeFilterName <genotypeFilterName> for filter type

Execute

The variant filters we want to set here are:

```
java -Xmx2g -jar /usr/local/GATK/GenomeAnalysisTK.jar -T VariantFiltration -R hg19.fa --variant NA12878.vcf -o NA12878_filter.vcf --filterExpression "QD < 2.0" --filterExpression "FS > 200.0" --filterExpression "MQ < 40.0" --filterName QDFilter --filterName FSFilter --filterName MQFilter
```

21: Variant Filtration on data 3 and data 16 (log)



20: Variant Filtration on data 3 and data 16 (Variant File)



19: http://cbwmain.entrydns.org/module2_3/db_SNP_135_chr1.vcf



18: Unified Genotyper on data 3, data 14, and data 11 (log)



17: Unified Genotyper on data 3, data 14, and data 11 (metrics)



You can look at the output vcf file that contains some filter annotation


snpEff

→ snpEff



SnpSift Filter Filter variants using arbitrary expressions

→ SnpEff Variant effect and annotation

CloudMap: Check snpEff Candidates Marks up a snpEff output file with matches to a gene candidate list.

 **SnpEff Variant effect and annotation (Galaxy Tool Version 1.0)** Options

Sequence changes (SNPs, MNPs, InDels)

  20: Variant Filtration on data 3 and data 16 (Variant File)

Input format

VCF

Output format

Tabular

Genome

Caenorhabditis elegans : WS220.64

Caenorhabditis elegans : WS220.64

Brachypodium_distachyon : Bd21_Bdistachyon_283_assembly_v2.0

Only C.elegans genome ???

SnEff on Galaxy

- Currently cannot use another genome than C.elegans or use your own genome on the main public server usegalaxy.org
- Looked for “Galaxy SnEff genome” found related answers on the “Galaxy Development List Archive” and on “biostar.usegalaxy.org”



Question: Need help with "SnEff" tool

SnpEff on Galaxy

From Jennifer Jackson (Galaxy team) on Feb 18, 2014 [Galaxy Development List Archive]

“Reference Genome in snpEff Tool”

“There are no current plans to include additional genomes to the SnpEff tool on the public Main Galaxy instance at <http://usegalaxy.org>.

The best solution is to either run a local Galaxy (with sufficient resources) or what is probably easier and more practical for many scientific end users, a cloud Galaxy or possibly a Slipstream Appliance. The tool wrapper is in the Tool Shed, so it can be installed and used within your Galaxy, where you can add in any genome that you want that has the appropriate reference data available.

Help to get started is in these links:

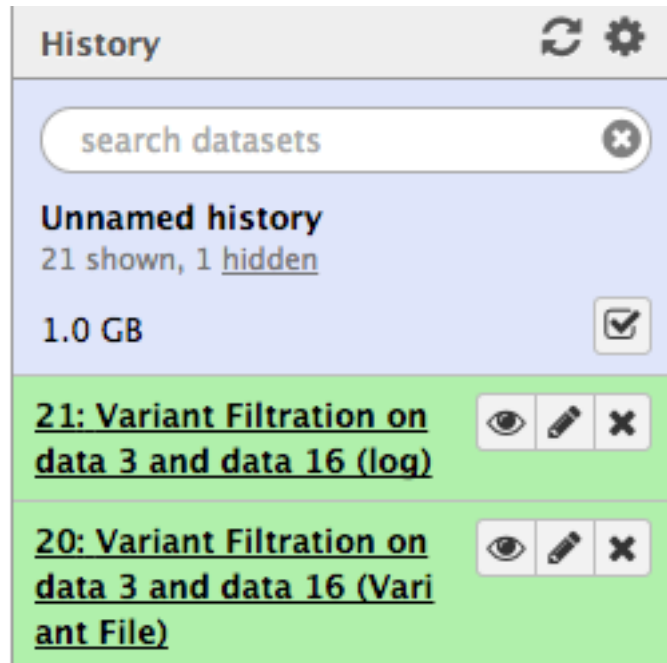
<https://wiki.galaxyproject.org/BigPicture/Choices>

<https://wiki.galaxyproject.org/Tool%20Shed>

Hopefully one of these solutions will work out for both of you!

-> We won't run it for this lab exercise

GATK VariantAnnotation (with dbSNPs)



NGS: GATK Tools (beta)

Table Recalibration on BAM files

Validate Variants

Select Variants from VCF files

Variant Recalibrator

Variant Filtration on VCF files

Eval Variants

Combine Variants

Apply Variant Recalibration



Variant Annotator

Unified Genotyper SNP and indel caller

Realigner Target Creator for use in local realignment

Print Reads from BAM files

Indel Realigner - perform local realignment




GATK VariantAnnotation (with dbSNPs)

Variant Annotator (Galaxy Tool Version 0.0.5) Options

Choose the source for the reference list

History

Variant file to annotate

   20: Variant Filtration on data 3 and data 16 (Variant File)

-V,--variant <variant>




Increase efficiency for small variant files.

☒ Yes ☐ No

--intervals

+ BAM file

Using reference file

   3: http://cbwmain.entrydns.org/module7/hg19_chr1.fa




-R,--reference_sequence <reference_sequence>

GATK VariantAnnotation (with dbSNPs)

Provide a dbSNP reference-ordered data file

→ Set dbSNP ▼
-D,--dbsnp <dbsnp>

ROD file

→    19: http://cbwmain.entrydns.org/module2_3/dbSNP_135_chr1.vcf ▼

→ Execute

History

search datasets

Unnamed history
23 shown, 1 hidden
1.0 GB

23: Variant Annotator on data 3, data 19, and data 20 (log)

22: Variant Annotator on data 3, data 19, and data 20 (Variant File)

21: Variant Filtration on data 3 and data 16 (log)

20: Variant Filtration on data 3 and data 16 (Variant File)

19: http://cbwmain.entr ydns.org/module2_3/db SNP_135_chr1.vcf

Save your workflow
Share your workflow

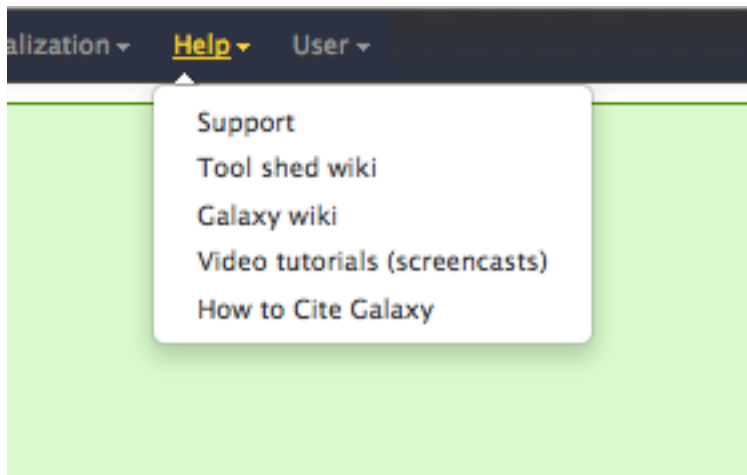
Download your filtered and annotated variant calls as vcf files

Continue using the tools on galaxy...

Continue exploring/analysing/visualizing your data and results..

Remember, lots of tutorials, videos, mailing list, twitter etc ...

- <https://vimeo.com/galaxyproject>



We are on a Coffee Break &
Networking Session