CloudMap | Cloud-based Pipeline for Analysis of Mutant Genome Sequences

This user guide serves as a simplified, graphic version of the CloudMap paper for application-oriented end-users. For more details, please see the CloudMap paper. Video versions of these user guides and updates to the pipeline are available at the CloudMap website at: http://usegalaxy.org/cloudmap.

Helpful Galaxy screencasts are available at: http://wiki.g2.bx.psu.edu/Learn/Screencasts

Currently, all of the workflows (with the exception of ***EMS Density Mapping***) should work for any species as long as users provide the appropriate genome reference file (Fasta) where required. Instructions for configuring multi-species support for the ***Hawaiian Variant Mapping with WGS Data*** tool is provided in the ***Analyze Your Own Data Using CloudMap Workflows*** section of this user guide.

CONTENTS:

p2 -- ***Table of Contents***

p3 -- ***Workflow Analysis Flowchart.*** This flowchart shows an overview of the workflows used for data analysis based on the type of starting data. A summary of output files is also provided.

p4 -- ***Hawaiian Variant Mapping with WGS Data and Variant Calling Workflow.*** This workflow is used to analyze the *ot266* Proof of principle in the CloudMap paper. Users may apply this workflow to their own SNP mapped data by substituting the *ot266* dataset with their own dataset. In addition to mapping plots, an annotated list of candidate variants is generated at the end of this workflow.

p19 -- ***Unmapped Mutant Workflow***. This workflow performs the same analysis as the mapping workflows without the mapping-specific tools. An annotated list of candidate variants is generated at the end of this workflow.

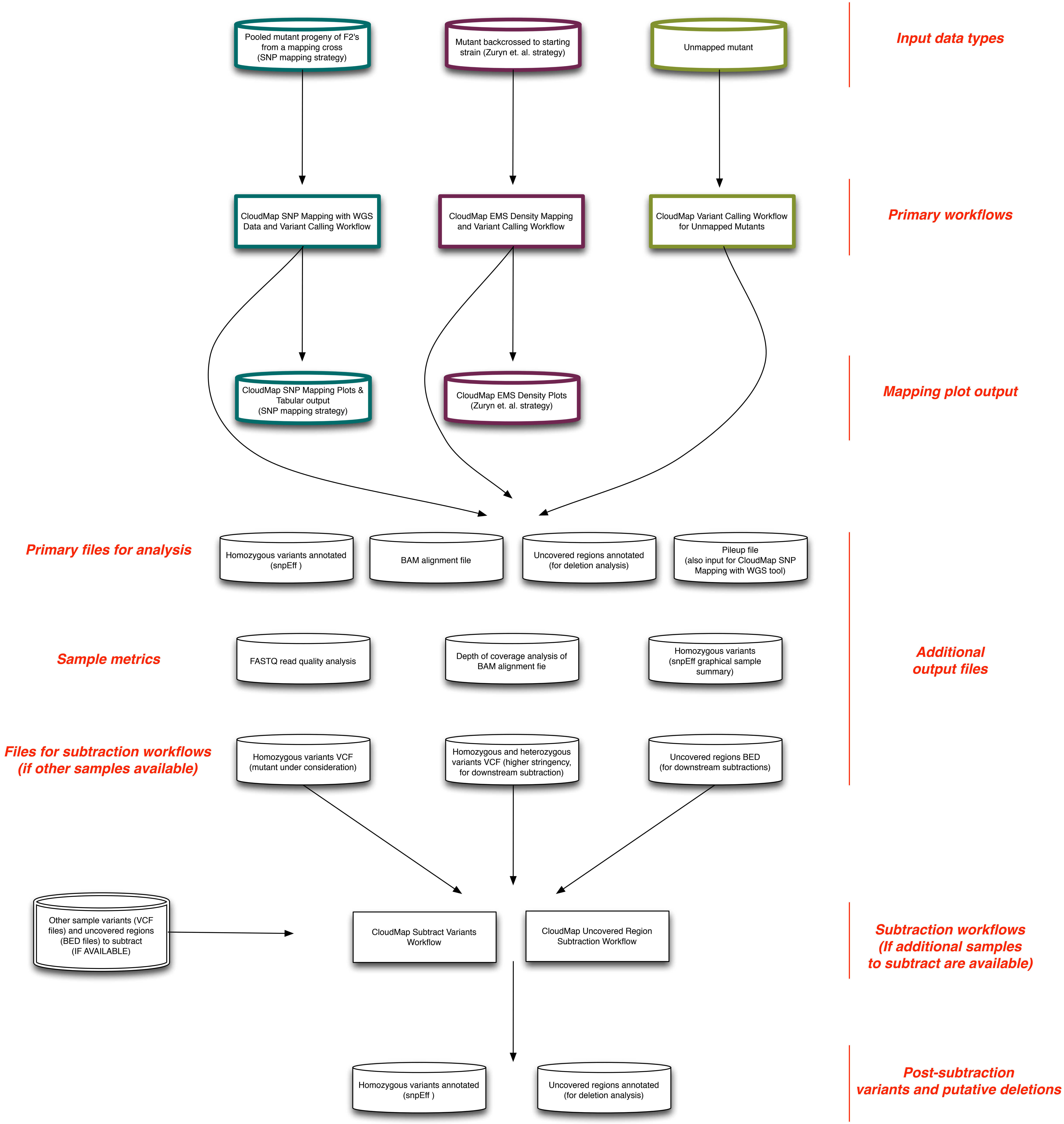
p33 -- ***EMS Density Mapping Workflow.*** This workflow is essentially the same as the ***Unmapped Mutant*** workflow followed by the ***Subtract Variants*** workflow with the addition of an ***EMS density*** plot of the final VCF variants file.

p34 -- ***Subtract Variants Workflow.*** This workflow can be used downstream of primary workflows run on SNP mapped strains, strains backcrossed to their starting strain, or unmapped strains. Here we demonstrate the workflow using the *ot266* example from **Fig. 8** of the CloudMap paper. An annotated list of candidate variants is generated at the end of this workflow.

p50 -- ***Uncovered Region Subtraction Workflow***. This workflow is analogous to the ***Subtract Variants*** workflow except it is performed with uncovered regions. It yields an annotated list of unique uncovered regions in a sample that may be tested for putative deletions with PCR and Sanger sequencing.

p59 -- ***Analyze Your Own Data Using CloudMap Workflows***. This section details how to upload your own datasets, modify CloudMap workflows, and provide support for species other than *C.elegans* or *Arabidopsis*.

p75 -- ***FAQ.*** Frequently Asked Questions


Input data types
Pooled mutant progeny of F2's from a mapping cross (SNP mapping strategy)
Mutant backcrossed to starting strain (Zuryn et. al. strategy)
Unmapped mutant
Primary workflows
CloudMap SNP Mapping with WGS Data and Variant Calling Workflow
CloudMap EMS Density Mapping and Variant Calling Workflow
CloudMap Variant Calling Workflow for Unmapped Mutants
Mapping plot output
CloudMap SNP Mapping Plots & Tabular output (SNP mapping strategy)
CloudMap EMS Density Plots (Zuryn et. al. strategy)
Primary files for analysis
Homozygous variants annotated (snpEff)
BAM alignment file
Uncovered regions annotated (for deletion analysis)
Pileup file (also input for CloudMap SNP Mapping with WGS tool)
Sample metrics
FASTQ read quality analysis
Depth of coverage analysis of BAM alignment fie
Homozygous variants (snpEff graphical sample summary)
Additional output files
Files for subtraction workflows (if other samples available)
Homozygous variants VCF (mutant under consideration)
Homozygous and heterozygous variants VCF (higher stringency, for downstream subtraction)
Uncovered regions BED (for downstream subtractions)
Other sample variants (VCF files) and uncovered regions (BED files) to subtract (IF AVAILABLE)
CloudMap Subtract Variants Workflow
CloudMap Uncovered Region Subtraction Workflow
Subtraction workflows (If additional samples to subtract are available)
Homozygous variants annotated (snpEff)
Uncovered regions annotated (for deletion analysis)
Post-subtraction variants and putative deletions

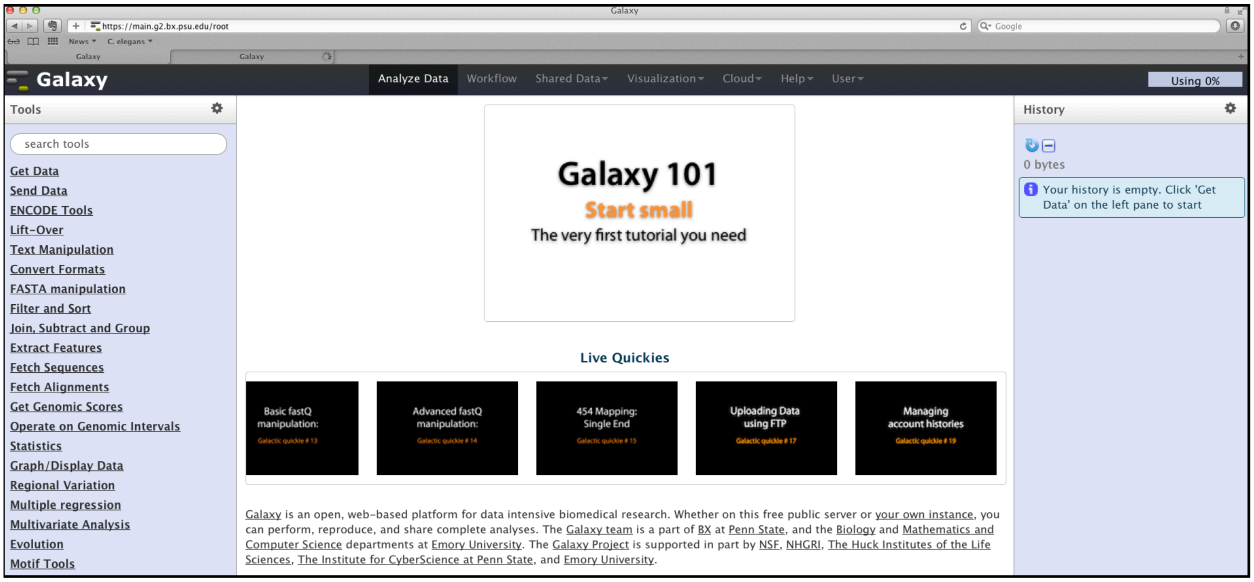
CloudMap Hawaiian Variant Mapping with WGS Data and Variant Calling Workflow (using *ot266* Proof of Principle example from the CloudMap paper). A video version of this user guide is available at: http://usegalaxy.org/cloudmap.

The *ot266* FASTQ file used in this example represents sequencing data from a specific kind of experiment: the *ot266* mutant has been crossed to a mapping strain (CB4856, "Hawaiian") and pooled F2 mutant progeny have been sequenced. This workflow uses single-end FASTQ data but it can be adapted to use paired-end data (see the ***Analyzing Your Own Data*** section of this user guide).

The aim in this user guide is to walk readers through Galaxy-based analysis of the *ot266* mutant using predefined CloudMap workflows which sequentially execute all of the steps required for common mutant analysis functions. This same workflow can be used for analysis of any mutant (from any species) that has been crossed to a mapping strain for which variant information is available.

These workflows provide default function parameters, ensuring that users follow best practices, and allow for automated execution of sequential operations. We provide these workflows as helpful guides, but experienced users may execute functions in any meaningful order they please and may also create and share their own workflows to take advantage of the automation feature. More CloudMap documentation is available at http://usegalaxy.org/cloudmap.

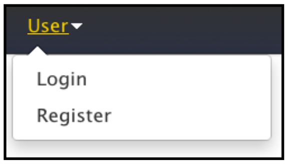
1) Navigate to http://usegalaxy.org (URL will resolve to something like https://main.g2.bx.psu.edu)



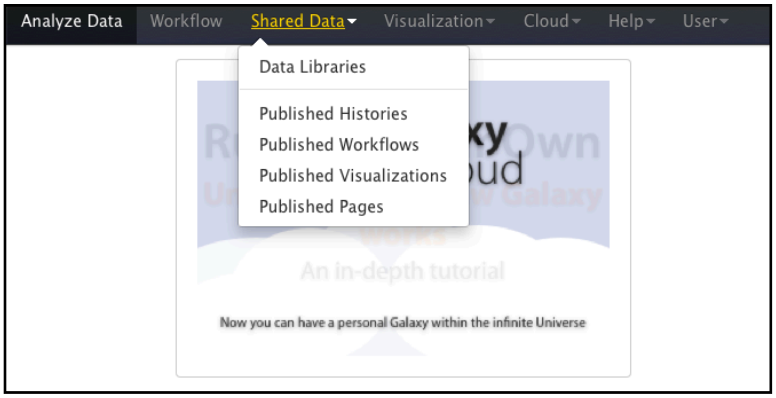
Your history is empty. Click 'Get Data' on the left pane to start

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

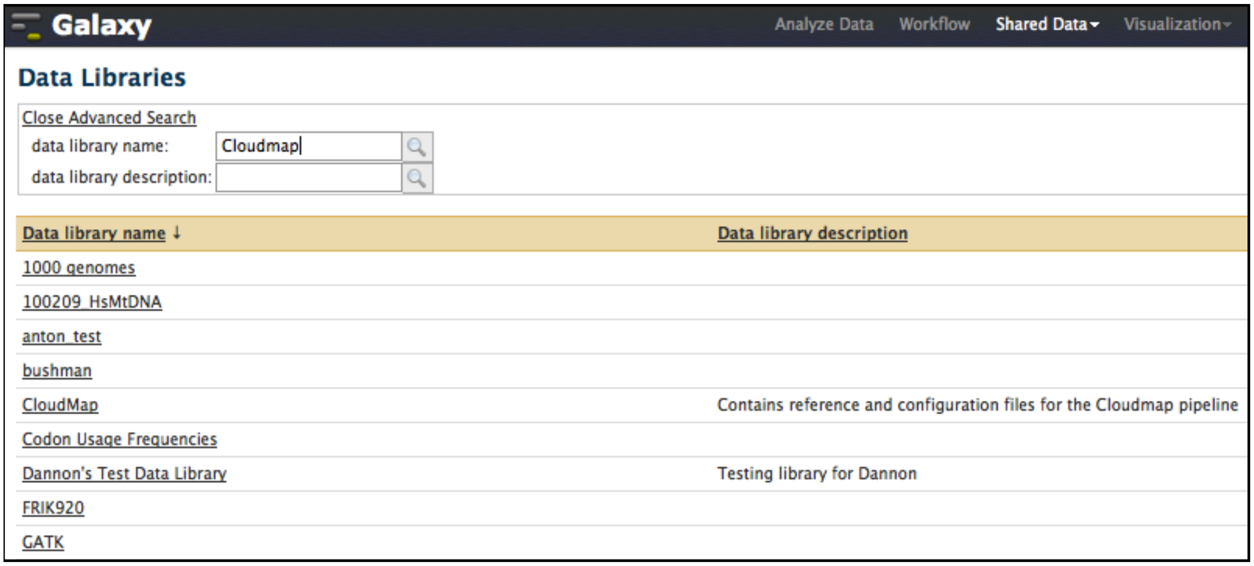
2) Register for an account or login if you already have an account:



3) Once you are logged in using your email address, click on the ***Shared Data*** link at the top of the page:



4) Click on ***Data Libraries*** and search for the CloudMap data library:



Data Libraries

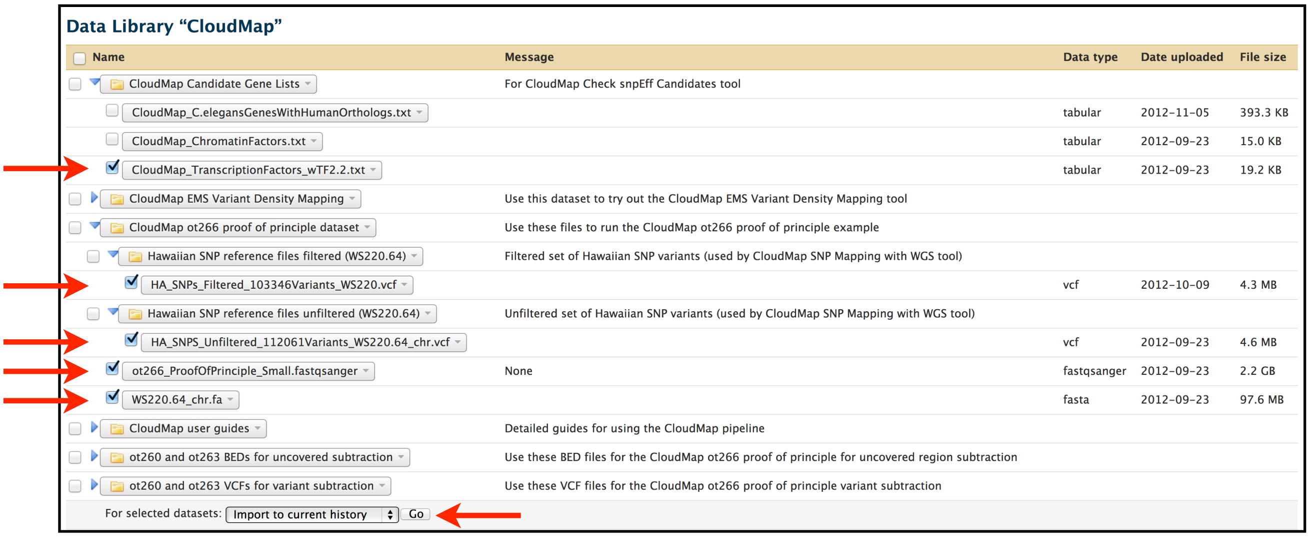
Close Advanced Search

data library name: Cloudmap

data library description:

Data library name ↓	Data library description
1000 genomes	
100209_HsMtDNA	
anton_test	
bushman	
CloudMap	Contains reference and configuration files for the Cloudmap pipeline
Codon Usage Frequencies	
Dannon's Test Data Library	Testing library for Dannon
FRIK920	
GATK	

5) Click on the ***CloudMap*** library and select the 5 data files below for the *ot266* example. Then click “Go” to import these files into your history.



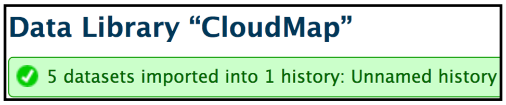
Data Library “CloudMap”

□ Name	Message	Data type	Date uploaded	File size
□ ▼ CloudMap Candidate Gene Lists	For CloudMap Check snpEff Candidates tool			
□ CloudMap_C.elegansGenesWithHumanOrthologs.txt		tabular	2012-11-05	393.3 KB
□ CloudMap_ChromatinFactors.txt		tabular	2012-09-23	15.0 KB
☑ CloudMap_TranscriptionFactors_wTF2.2.txt		tabular	2012-09-23	19.2 KB
□ ▶ CloudMap EMS Variant Density Mapping	Use this dataset to try out the CloudMap EMS Variant Density Mapping tool			
□ ▼ CloudMap ot266 proof of principle dataset	Use these files to run the CloudMap ot266 proof of principle example			
□ ▼ Hawaiian SNP reference files filtered (WS220.64)	Filtered set of Hawaiian SNP variants (used by CloudMap SNP Mapping with WGS tool)			
☑ HA_SNPs_Filtered_103346Variants_WS220.vcf		vcf	2012-10-09	4.3 MB
□ ▼ Hawaiian SNP reference files unfiltered (WS220.64)	Unfiltered set of Hawaiian SNP variants (used by CloudMap SNP Mapping with WGS tool)			
☑ HA_SNPS_Unfiltered_112061Variants_WS220.64_chr.vcf		vcf	2012-09-23	4.6 MB
☑ ot266_ProofOfPrinciple_Small.fastqsanger	None	fastqsanger	2012-09-23	2.2 GB
☑ WS220.64_chr.fa		fasta	2012-09-23	97.6 MB
□ ▶ CloudMap user guides	Detailed guides for using the CloudMap pipeline			
□ ▶ ot260 and ot263 BEDs for uncovered subtraction	Use these BED files for the CloudMap ot266 proof of principle for uncovered region subtraction			
□ ▶ ot260 and ot263 VCFs for variant subtraction	Use these VCF files for the CloudMap ot266 proof of principle variant subtraction			

For selected datasets: Import to current history Go

The filtered “HA_SNPS” file is used to generate SNP mapping plots (details in **Table S1** of the CloudMap paper). The unfiltered “HA_SNPs” VCF is used for variant subtraction as shown in **Fig.8**. of the CloudMap paper.

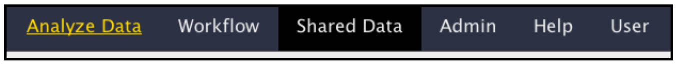
6) You will receive confirmation that the files have been imported into your history:



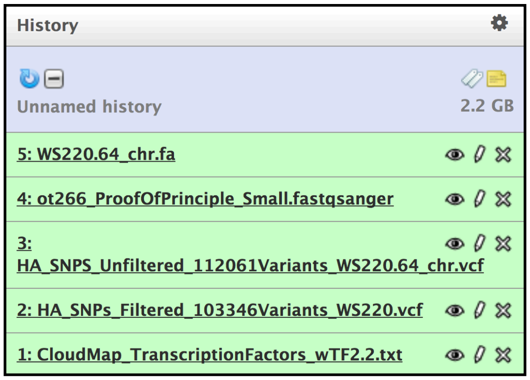
Data Library “CloudMap”

✓ 5 datasets imported into 1 history: Unnamed history

7) Click ***Analyze Data*** on the menu bar to navigate to your history:



8) You will now see that the data files have been added to an unnamed history:



History

Unnamed history 2.2 GB

5: WS220.64_chr.fa

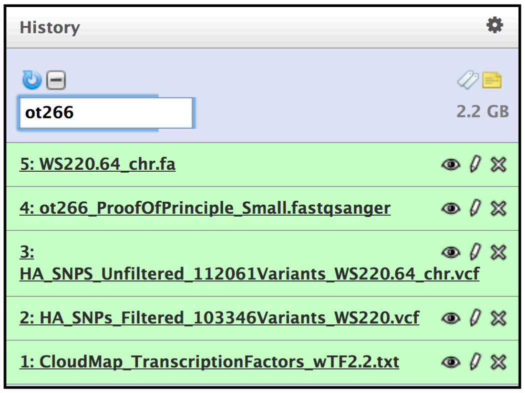
4: ot266_ProofOfPrinciple_Small.fastqsanger

3: HA_SNPS_Unfiltered_112061Variants_WS220.64_chr.vcf

2: HA_SNPs_Filtered_103346Variants_WS220.vcf

1: CloudMap_TranscriptionFactors_wTF2.2.txt

9) Name your history ***ot266*** after the sample that we will be analyzing:



History

ot266 2.2 GB

5: WS220.64_chr.fa

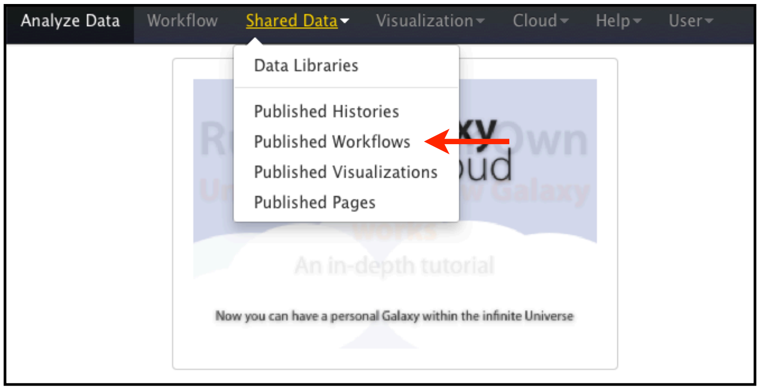
4: ot266_ProofOfPrinciple_Small.fastqsanger

3: HA_SNPS_Unfiltered_112061Variants_WS220.64_chr.vcf

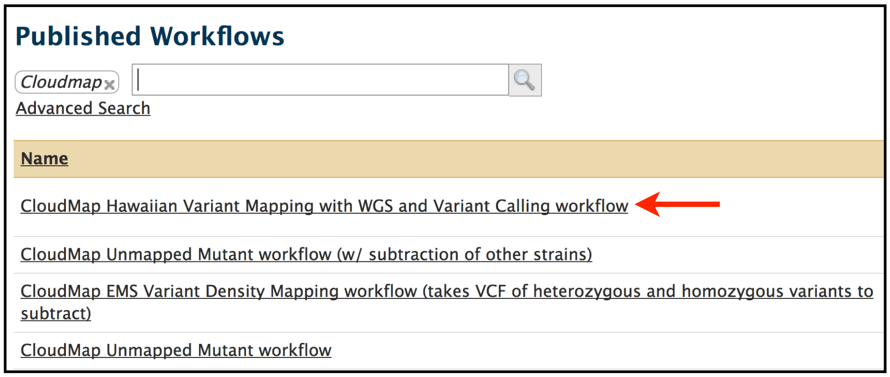
2: HA_SNPs_Filtered_103346Variants_WS220.vcf

1: CloudMap_TranscriptionFactors_wTF2.2.txt

10) Again click on the ***Shared Data*** link at the top of the page and select ***Published Workflows*** :



11) Use the search term “CloudMap” to view the automated workflows. Select the ***CloudMap Hawaiian Variant Mapping with WGS Data and Variant Calling workflow***.



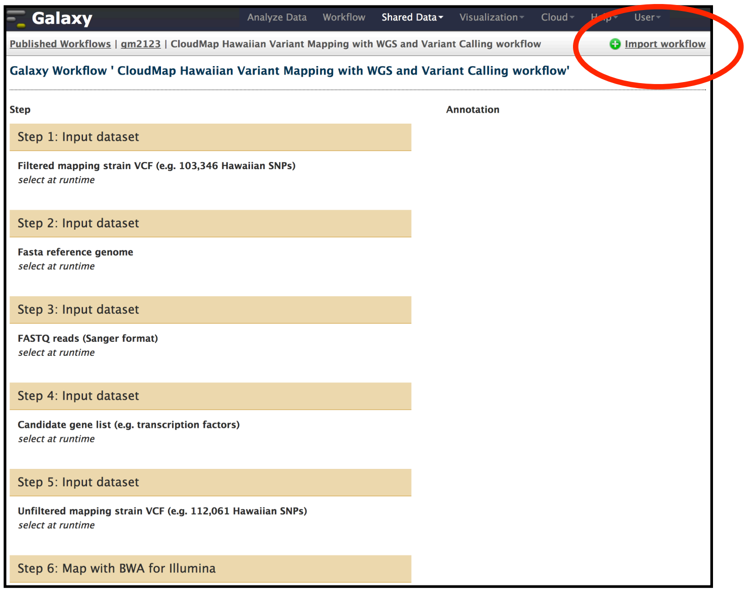
Published Workflows

Cloudmap ×

Advanced Search

Name
CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow
CloudMap Unmapped Mutant workflow (w/ subtraction of other strains)
CloudMap EMS Variant Density Mapping workflow (takes VCF of heterozygous and homozygous variants to subtract)
CloudMap Unmapped Mutant workflow

12) You will now have the option to ***Import workflow***

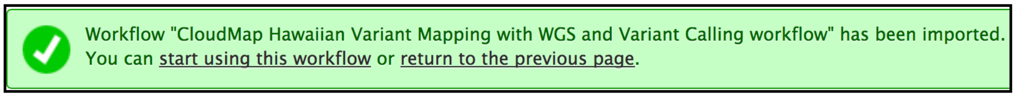


Published Workflows | gm2123 | CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow

Galaxy Workflow ' CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow'

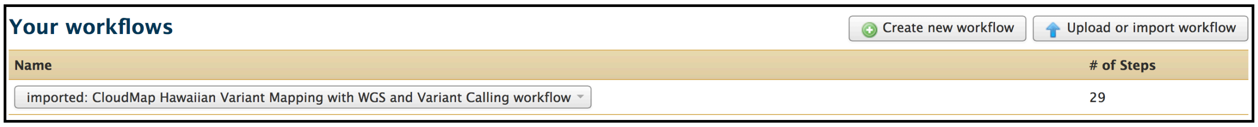
Step	Annotation
Step 1: Input dataset	
Filtered mapping strain VCF (e.g. 103,346 Hawaiian SNPs) *select at runtime*	
Step 2: Input dataset	
Fasta reference genome *select at runtime*	
Step 3: Input dataset	
FASTQ reads (Sanger format) *select at runtime*	
Step 4: Input dataset	
Candidate gene list (e.g. transcription factors) *select at runtime*	
Step 5: Input dataset	
Unfiltered mapping strain VCF (e.g. 112,061 Hawaiian SNPs) *select at runtime*	
Step 6: Map with BWA for Illumina	

13) You will see the message below. Click ***Start using this workflow***.



Workflow "CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow" has been imported.
You can start using this workflow or return to the previous page.

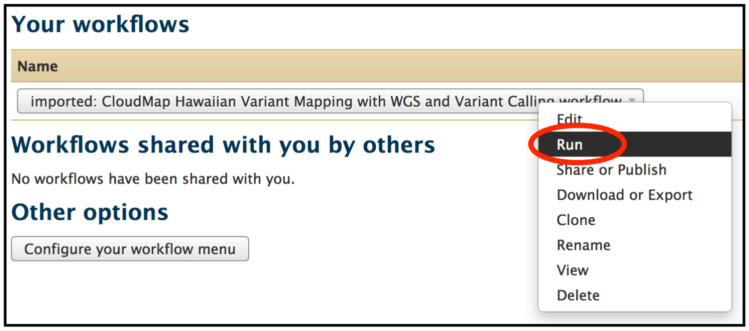
14) You will see that the workflow has been imported. From now on, you can easily access this workflow under the ***Workflow*** tab.



Your workflows

Name	# of Steps
imported: CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow	29

15) Click on the workflow and select ***Run***:



Your workflows

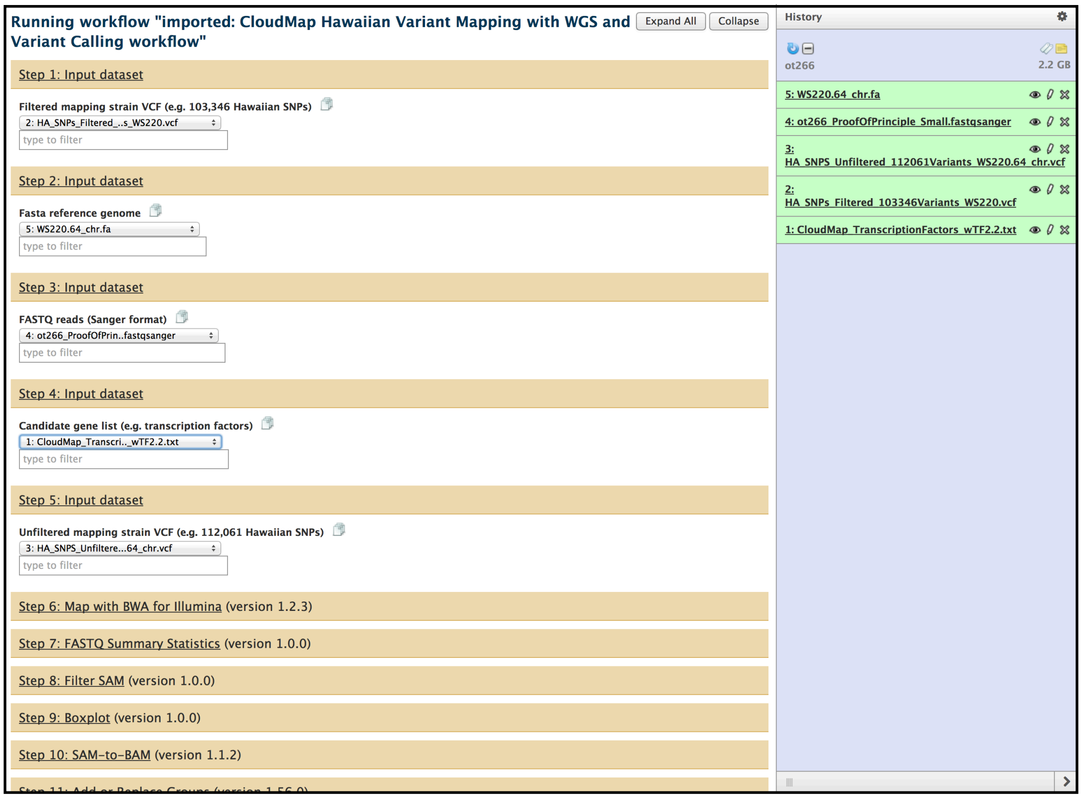
Name
imported: CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow

Workflows shared with you by others

No workflows have been shared with you.

Other options

16) You will see all the steps in the workflow prior to running it. Make sure that each of the input fields corresponds to the appropriate file in your history.



Running workflow "imported: CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow"

Step 1: Input dataset

Filtered mapping strain VCF (e.g. 103,346 Hawaiian SNPs)
2: HA_SNPs_Filtered_..s_WS220.vcf
type to filter

Step 2: Input dataset

Fasta reference genome
5: WS220.64_chr.fa
type to filter

Step 3: Input dataset

FASTQ reads (Sanger format)
4: ot266_ProofOfPrin..fastqsanger
type to filter

Step 4: Input dataset

Candidate gene list (e.g. transcription factors)
1: CloudMap_Transcri..._wTF2.2.txt
type to filter

Step 5: Input dataset

Unfiltered mapping strain VCF (e.g. 112,061 Hawaiian SNPs)
3: HA_SNPS_Unfiltere...64_chr.vcf
type to filter

Step 6: Map with BWA for Illumina (version 1.2.3)

Step 7: FASTQ Summary Statistics (version 1.0.0)

Step 8: Filter SAM (version 1.0.0)

Step 9: Boxplot (version 1.0.0)

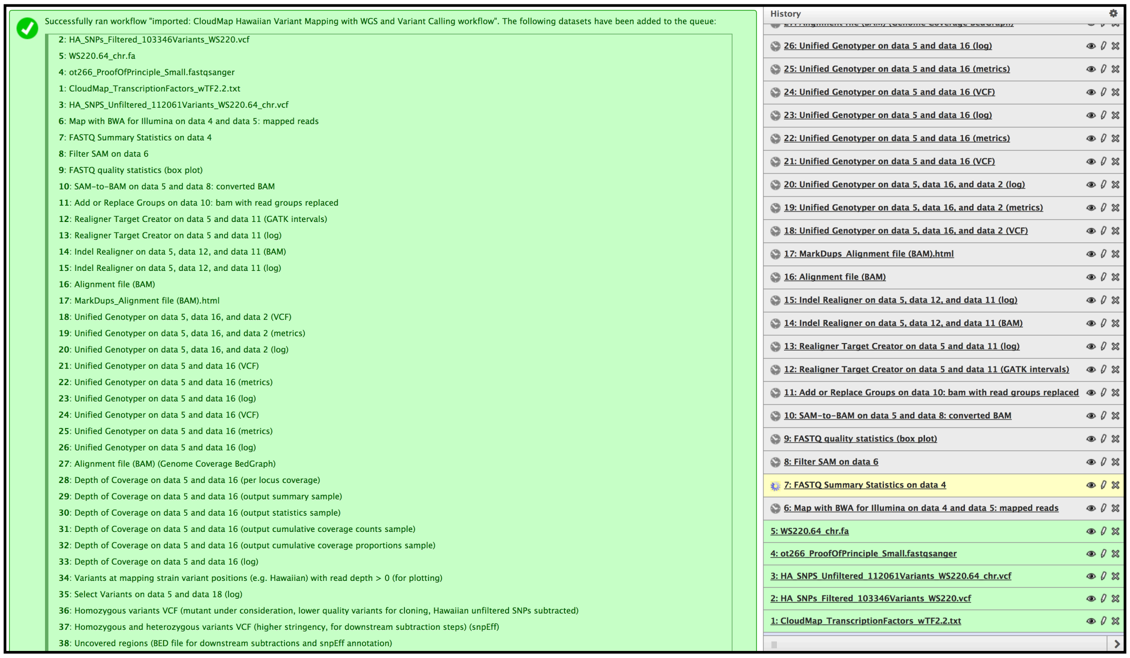
Step 10: SAM-to-BAM (version 1.1.2)

Step 11: Add or Replace Groups (version 1.56.0)

History
ot266 2.2 GB

5: WS220.64_chr.fa
4: ot266_ProofOfPrinciple_Small.fastqsanger
3: HA_SNPS_Unfiltered_112061Variants_WS220.64_chr.vcf
2: HA_SNPs_Filtered_103346Variants_WS220.vcf
1: CloudMap_TranscriptionFactors_wTF2.2.txt

17) All of the automated functions have the appropriate default parameters configured, although experienced users may want to modify these prior to running (see the ***Analyzing Your Own Data Using CloudMap Workflows*** section of this user guide). Once you are ready to run the workflow, press ***Run Workflow*** at the bottom of the page and the workflow will start (this step takes a minute or two to begin, be patient and don't hit the ***Run Workflow*** button repeatedly). You will receive an email when the workflow is completed:



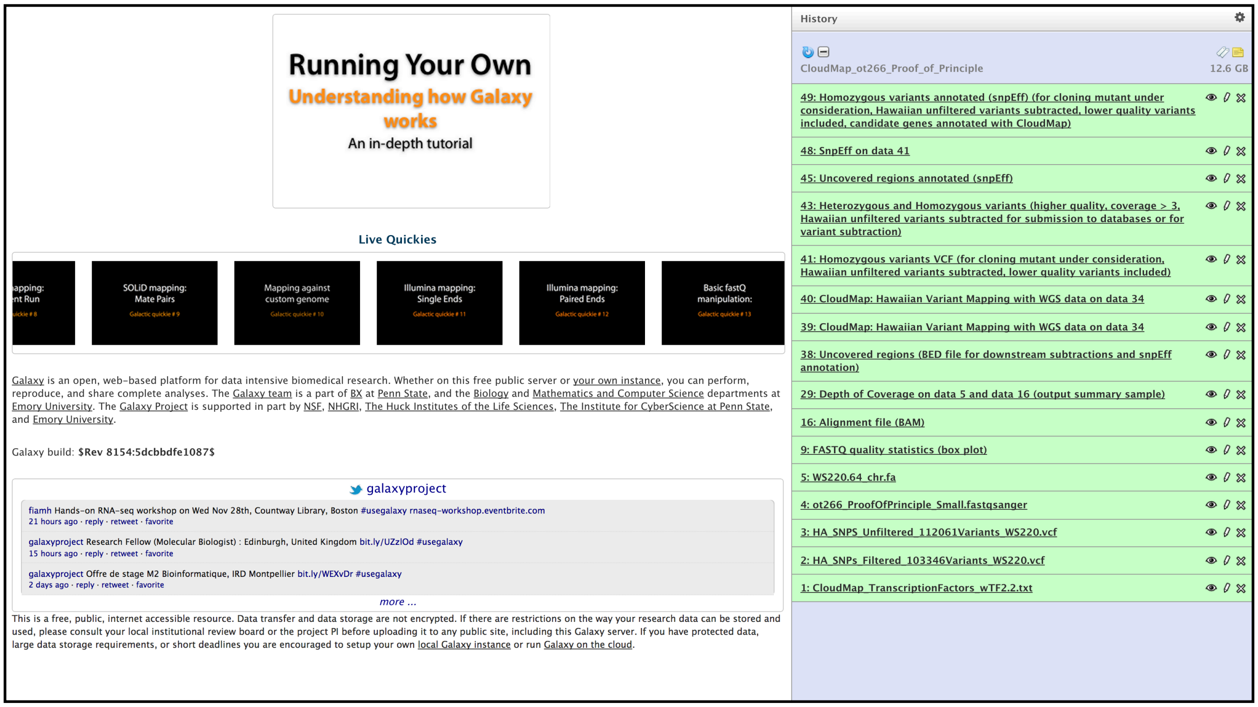
Successfully ran workflow "imported: CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow". The following datasets have been added to the queue:

```
2: HA_SNPs_Filtered_103346Variants_WS220.vcf
5: WS220.64_chr.fa
4: ot266_ProofOfPrinciple_Small.fastqsanger
1: CloudMap_TranscriptionFactors_wTF2.2.txt
3: HA_SNPS_Unfiltered_112061Variants_WS220.64_chr.vcf
6: Map with BWA for Illumina on data 4 and data 5: mapped reads
7: FASTQ Summary Statistics on data 4
8: Filter SAM on data 6
9: FASTQ quality statistics (box plot)
10: SAM-to-BAM on data 5 and data 8: converted BAM
11: Add or Replace Groups on data 10: bam with read groups replaced
12: Realigner Target Creator on data 5 and data 11 (GATK intervals)
13: Realigner Target Creator on data 5 and data 11 (log)
14: Indel Realigner on data 5, data 12, and data 11 (BAM)
15: Indel Realigner on data 5, data 12, and data 11 (log)
16: Alignment file (BAM)
17: MarkDups_Alignment file (BAM).html
18: Unified Genotyper on data 5, data 16, and data 2 (VCF)
19: Unified Genotyper on data 5, data 16, and data 2 (metrics)
20: Unified Genotyper on data 5, data 16, and data 2 (log)
21: Unified Genotyper on data 5 and data 16 (VCF)
22: Unified Genotyper on data 5 and data 16 (metrics)
23: Unified Genotyper on data 5 and data 16 (log)
24: Unified Genotyper on data 5 and data 16 (VCF)
25: Unified Genotyper on data 5 and data 16 (metrics)
26: Unified Genotyper on data 5 and data 16 (log)
27: Alignment file (BAM) (Genome Coverage BedGraph)
28: Depth of Coverage on data 5 and data 16 (per locus coverage)
29: Depth of Coverage on data 5 and data 16 (output summary sample)
30: Depth of Coverage on data 5 and data 16 (output statistics sample)
31: Depth of Coverage on data 5 and data 16 (output cumulative coverage counts sample)
32: Depth of Coverage on data 5 and data 16 (output cumulative coverage proportions sample)
33: Depth of Coverage on data 5 and data 16 (log)
34: Variants at mapping strain variant positions (e.g. Hawaiian) with read depth > 0 (for plotting)
35: Select Variants on data 5 and data 18 (log)
36: Homozygous variants VCF (mutant under consideration, lower quality variants for cloning, Hawaiian unfiltered SNPs subtracted)
37: Homozygous and heterozygous variants VCF (higher stringency, for downstream subtraction steps) (snpEff)
38: Uncovered regions (BED file for downstream subtractions and snpEff annotation)
```

History

26: Unified Genotyper on data 5 and data 16 (log)
25: Unified Genotyper on data 5 and data 16 (metrics)
24: Unified Genotyper on data 5 and data 16 (VCF)
23: Unified Genotyper on data 5 and data 16 (log)
22: Unified Genotyper on data 5 and data 16 (metrics)
21: Unified Genotyper on data 5 and data 16 (VCF)
20: Unified Genotyper on data 5, data 16, and data 2 (log)
19: Unified Genotyper on data 5, data 16, and data 2 (metrics)
18: Unified Genotyper on data 5, data 16, and data 2 (VCF)
17: MarkDups_Alignment file (BAM).html
16: Alignment file (BAM)
15: Indel Realigner on data 5, data 12, and data 11 (log)
14: Indel Realigner on data 5, data 12, and data 11 (BAM)
13: Realigner Target Creator on data 5 and data 11 (log)
12: Realigner Target Creator on data 5 and data 11 (GATK intervals)
11: Add or Replace Groups on data 10: bam with read groups replaced
10: SAM-to-BAM on data 5 and data 8: converted BAM
9: FASTQ quality statistics (box plot)
8: Filter SAM on data 6
7: FASTQ Summary Statistics on data 4
6: Map with BWA for Illumina on data 4 and data 5: mapped reads
5: WS220.64_chr.fa
4: ot266_ProofOfPrinciple_Small.fastqsanger
3: HA_SNPS_Unfiltered_112061Variants_WS220.64_chr.vcf
2: HA_SNPs_Filtered_103346Variants_WS220.vcf
1: CloudMap_TranscriptionFactors_wTF2.2.txt

18) Once the workflow has finished running, you can view the resulting output:



Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Galaxy build: **$Rev 8154:5dcbbdfe1087$**

galaxyproject

fiamh Hands-on RNA-seq workshop on Wed Nov 28th, Countway Library, Boston #usegalaxy rnaseq-workshop.eventbrite.com
21 hours ago · reply · retweet · favorite

galaxyproject Research Fellow (Molecular Biologist) : Edinburgh, United Kingdom bit.ly/UZzlOd #usegalaxy
15 hours ago · reply · retweet · favorite

galaxyproject Offre de stage M2 Bioinformatique, IRD Montpellier bit.ly/WEXvDr #usegalaxy
2 days ago · reply · retweet · favorite

more ...

This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own local Galaxy instance or run Galaxy on the cloud.

CloudMap_ot266_Proof_of_Principle 12.6 GB

49: Homozygous variants annotated (snpEff) (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included, candidate genes annotated with CloudMap)

48: SnpEff on data 41

45: Uncovered regions annotated (snpEff)

43: Heterozygous and Homozygous variants (higher quality, coverage > 3, Hawaiian unfiltered variants subtracted for submission to databases or for variant subtraction)

41: Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)

40: CloudMap: Hawaiian Variant Mapping with WGS data on data 34

39: CloudMap: Hawaiian Variant Mapping with WGS data on data 34

38: Uncovered regions (BED file for downstream subtractions and snpEff annotation)

29: Depth of Coverage on data 5 and data 16 (output summary sample)

16: Alignment file (BAM)

9: FASTQ quality statistics (box plot)

5: WS220.64_chr.fa

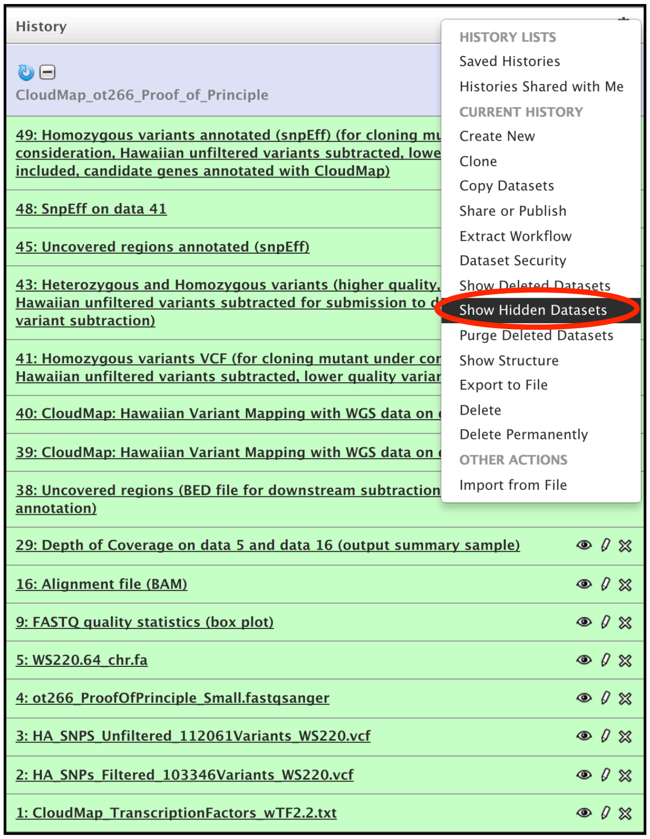
4: ot266_ProofOfPrinciple_Small.fastqsanger

3: HA_SNPS_Unfiltered_112061Variants_WS220.vcf

2: HA_SNPs_Filtered_103346Variants_WS220.vcf

1: CloudMap_TranscriptionFactors_wTF2.2.txt

19) You will notice that while over 40 output files were generated during the course of the workflow (output files are sequentially numbered), only some output files remain visible while others are hidden. The visible files are most important for analysis of the mutant under consideration or downstream analysis. In order to view hidden files, click ***Show Hidden Datasets*** in the History menu:



CloudMap_ot266_Proof_of_Principle

49: Homozygous variants annotated (snpEff) (for cloning mu consideration, Hawaiian unfiltered variants subtracted, lowe included, candidate genes annotated with CloudMap)

48: SnpEff on data 41

45: Uncovered regions annotated (snpEff)

43: Heterozygous and Homozygous variants (higher quality, Hawaiian unfiltered variants subtracted for submission to d variant subtraction)

41: Homozygous variants VCF (for cloning mutant under cor Hawaiian unfiltered variants subtracted, lower quality variar

40: CloudMap: Hawaiian Variant Mapping with WGS data on

39: CloudMap: Hawaiian Variant Mapping with WGS data on

38: Uncovered regions (BED file for downstream subtraction annotation)

29: Depth of Coverage on data 5 and data 16 (output summary sample)

16: Alignment file (BAM)

9: FASTQ quality statistics (box plot)

5: WS220.64_chr.fa

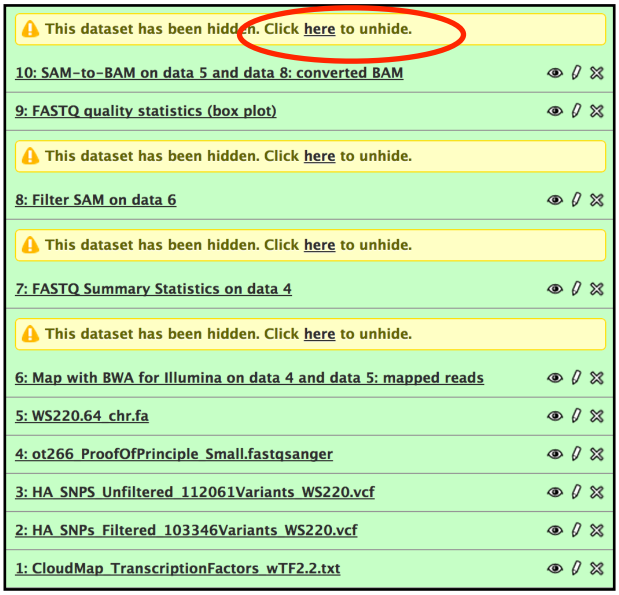
4: ot266_ProofOfPrinciple_Small.fastqsanger

3: HA_SNPS_Unfiltered_112061Variants_WS220.vcf

2: HA_SNPs_Filtered_103346Variants_WS220.vcf

1: CloudMap_TranscriptionFactors_wTF2.2.txt

20) You may unhide any files that are hidden:



This dataset has been hidden. Click here to unhide.

10: SAM-to-BAM on data 5 and data 8: converted BAM

9: FASTQ quality statistics (box plot)

This dataset has been hidden. Click here to unhide.

8: Filter SAM on data 6

This dataset has been hidden. Click here to unhide.

7: FASTQ Summary Statistics on data 4

This dataset has been hidden. Click here to unhide.

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

5: WS220.64_chr.fa

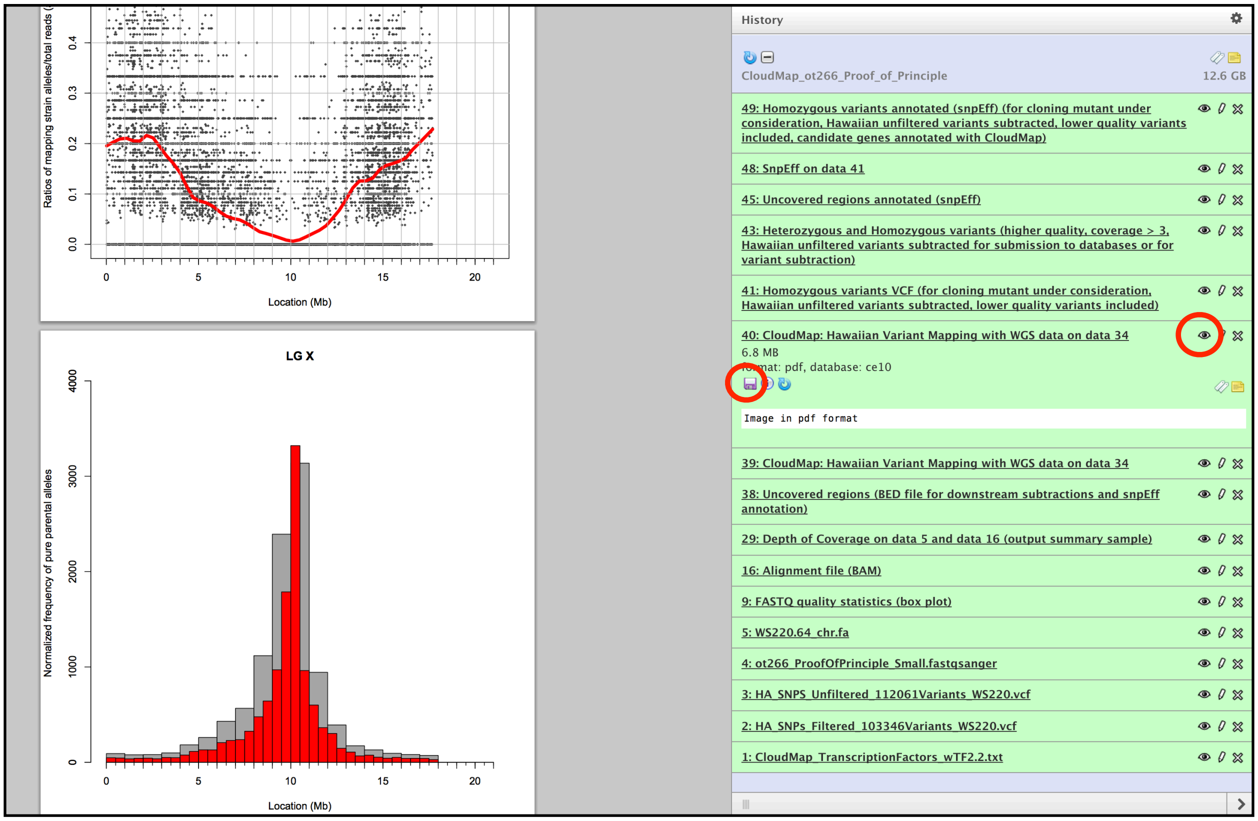
4: ot266_ProofOfPrinciple_Small.fastqsanger

3: HA_SNPS_Unfiltered_112061Variants_WS220.vcf

2: HA_SNPs_Filtered_103346Variants_WS220.vcf

1: CloudMap_TranscriptionFactors_wTF2.2.txt

21) Click on a file to view more information on that file or to download the file:



Ratios of mapping strain alleles/total reads (
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
Location (Mb)
LG X
Normalized frequency of pure parental alleles
0
1000
2000
3000
4000
0
5
10
15
20
Location (Mb)
History

CloudMap_ot266_Proof_of_Principle 12.6 GB

49: Homozygous variants annotated (snpEff) (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included, candidate genes annotated with CloudMap)

48: SnpEff on data 41

45: Uncovered regions annotated (snpEff)

43: Heterozygous and Homozygous variants (higher quality, coverage > 3, Hawaiian unfiltered variants subtracted for submission to databases or for variant subtraction)

41: Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)

40: CloudMap: Hawaiian Variant Mapping with WGS data on data 34
6.8 MB
format: pdf, database: ce10

```
Image in pdf format
```

39: CloudMap: Hawaiian Variant Mapping with WGS data on data 34

38: Uncovered regions (BED file for downstream subtractions and snpEff annotation)

29: Depth of Coverage on data 5 and data 16 (output summary sample)

16: Alignment file (BAM)

9: FASTQ quality statistics (box plot)

5: WS220.64_chr.fa

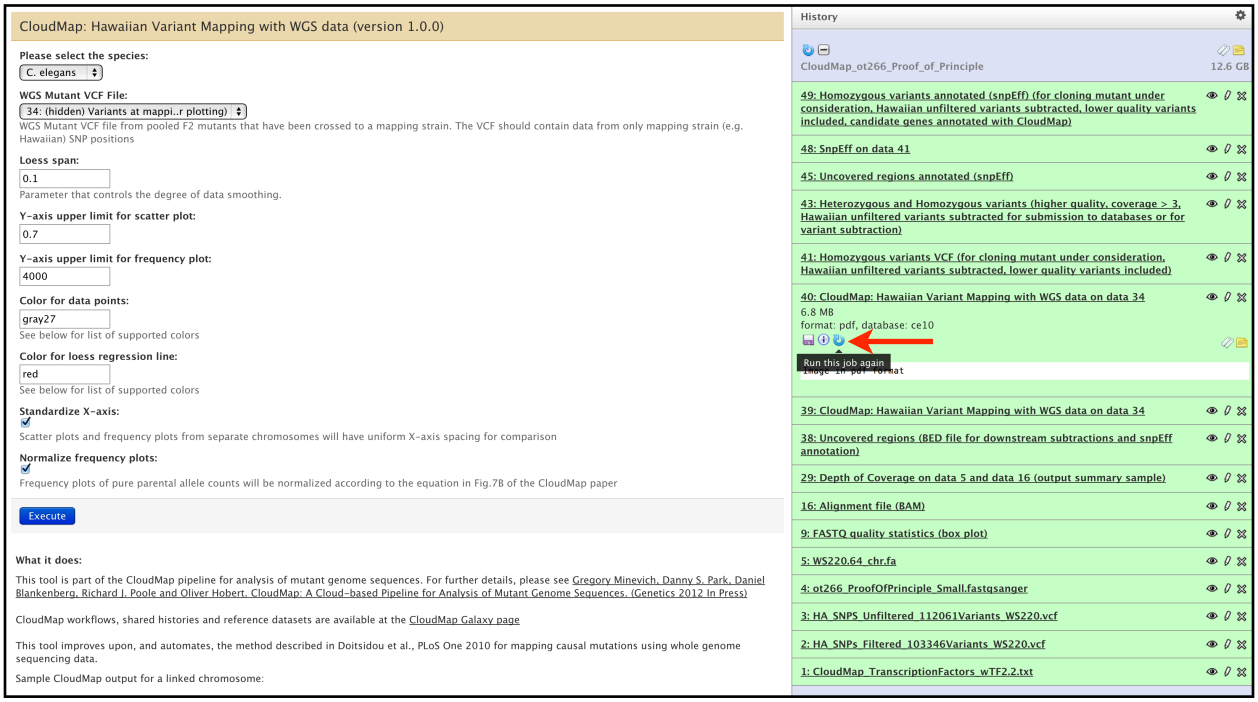
4: ot266_ProofOfPrinciple_Small.fastqsanger

3: HA_SNPS_Unfiltered_112061Variants_WS220.vcf

2: HA_SNPs_Filtered_103346Variants_WS220.vcf

1: CloudMap_TranscriptionFactors_wTF2.2.txt

If you want to rerun a tool with different parameters, click the ***run this job again*** arrow. To rerun a tool on a hidden dataset, make sure to unhide the hidden dataset first. If a tool fails (it will turn red) for no apparent reason when it has previously worked successfully, try running it again before submitting a bug report to Galaxy.



CloudMap: Hawaiian Variant Mapping with WGS data (version 1.0.0)

Please select the species:
C. elegans

WGS Mutant VCF File:
34: (hidden) Variants at mappi..r plotting)
WGS Mutant VCF file from pooled F2 mutants that have been crossed to a mapping strain. The VCF should contain data from only mapping strain (e.g. Hawaiian) SNP positions

Loess span:
0.1
Parameter that controls the degree of data smoothing.

Y-axis upper limit for scatter plot:
0.7

Y-axis upper limit for frequency plot:
4000

Color for data points:
gray27
See below for list of supported colors

Color for loess regression line:
red
See below for list of supported colors

Standardize X-axis:
☑
Scatter plots and frequency plots from separate chromosomes will have uniform X-axis spacing for comparison

Normalize frequency plots:
☑
Frequency plots of pure parental allele counts will be normalized according to the equation in Fig.7B of the CloudMap paper

Execute

What it does:

This tool is part of the CloudMap pipeline for analysis of mutant genome sequences. For further details, please see Gregory Minevich, Danny S. Park, Daniel Blankenberg, Richard J. Poole and Oliver Hobert. CloudMap: A Cloud-based Pipeline for Analysis of Mutant Genome Sequences. (Genetics 2012 In Press)

CloudMap workflows, shared histories and reference datasets are available at the CloudMap Galaxy page

This tool improves upon, and automates, the method described in Doitsidou et al., PLoS One 2010 for mapping causal mutations using whole genome sequencing data.

Sample CloudMap output for a linked chromosome:

History

CloudMap_ot266_Proof_of_Principle 12.6 GB

49: Homozygous variants annotated (snpEff) (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included, candidate genes annotated with CloudMap)

48: SnpEff on data 41

45: Uncovered regions annotated (snpEff)

43: Heterozygous and Homozygous variants (higher quality, coverage > 3, Hawaiian unfiltered variants subtracted for submission to databases or for variant subtraction)

41: Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)

40: CloudMap: Hawaiian Variant Mapping with WGS data on data 34
6.8 MB
format: pdf, database: ce10
Run this job again
Image in pdf format

39: CloudMap: Hawaiian Variant Mapping with WGS data on data 34

38: Uncovered regions (BED file for downstream subtractions and snpEff annotation)

29: Depth of Coverage on data 5 and data 16 (output summary sample)

16: Alignment file (BAM)

9: FASTQ quality statistics (box plot)

5: WS220.64_chr.fa

4: ot266_ProofOfPrinciple_Small.fastqsanger

3: HA_SNPS_Unfiltered_112061Variants_WS220.vcf

2: HA_SNPs_Filtered_103346Variants_WS220.vcf

1: CloudMap_TranscriptionFactors_wTF2.2.txt

22) Several ***sample metric*** files are created as part of the workflow (more details on following pages):

1. A ***FASTQ quality statistics*** file summarizes the quality of all reads before they are aligned to the reference genome (*Galaxy's FASTQ manipulation tools*).

2. A ***Depth of Coverage*** file gives a summary of overall read depth in the BAM alignment file (*GATK*).

3. A ***graphical summary of all the variants*** in the sample (*snpEff)*. This file must be downloaded to be viewed properly. It will not appear correctly if viewed within Galaxy using the "peek" (eye) icon. (For more information on file format, see: http://snpeff.sourceforge.net/)

23) A ***primary set of files for analysis*** are created as part of the workflow:

1. A CloudMap-generated ***Hawaiian Variant Mapping plot*** that narrows down the region of genome containing the causal variant(s) and a ***tabular file containing the data used to make the plots***.

2. An ***annotated set of homozygous variants*** in the entire sample (*snpEff*) including annotation of candidate genes with CloudMap. (For more information on file format, see: http://snpeff.sourceforge.net/)

3. A ***BAM alignment file*** that can be viewed in your choice of alignment viewers (*SAMtools*). (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat.html)

4. A list of ***annotated uncovered regions*** (BED file) that may be putative deletions (*BEDtools* & *snpEff)*. (For more information on file format, see: http://snpeff.sourceforge.net/)

24) Additional files that can be used for ***downstream subtraction workflows*** are generated (for more details see the ***Subtract Variants*** and ***Uncovered Region Subtraction*** workflows):

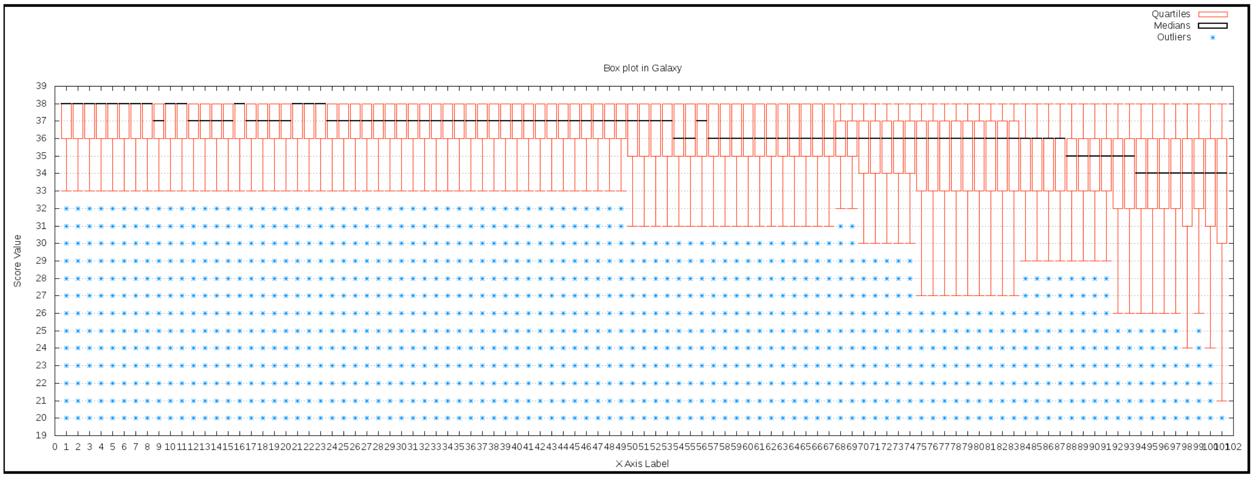
1. A ***set of homozygous variants*** (VCF file) in the entire sample that can be further filtered by subtracting variants present in other samples using the ***CloudMap Subtract Variants*** workflow (*GATK*). This VCF file is used as input into snpEff to generate the ***annotated list of homozygous variants*** mentioned in the section above. It has Hawaiian unfiltered variants subtracted and includes variants that pass a low quality filtering threshold. This file should be downloaded to be easily viewed in its entirety. The first several lines in any VCF file are header lines starting with "#" so users who wish to filter or sort these files in Excel are advised to remove the header lines. (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat.html)

2. A ***set of homozygous and heterozygous variants*** (VCF file) in the entire sample (run at higher quality stringency) that can be used as a set of variants to subtract from other samples (GATK). It has Hawaiian unfiltered variants subtracted and includes variants that pass a higher quality filtering threshold (read mapping quality ≥ 30 and coverage ≥ 3). In an effort to subtract as many variants as possible, users may subtract not only homozygous variants from other strains, but also heterozygous variants. Such a strategy assumes that phenotype-inducing homozygous mutant variants in the strain under analysis are unlikely to be heterozygous in strains that will be used for subtraction. It is especially important to apply this strategy when subtracting variant lists generated using the *Hawaiian Variant Mapping with WGS Data* approach (see section "**CloudMap *Hawaiian Variant Mapping with WGS Data*** tool"), since background variants will be present in a heterozygous state in these pooled samples as a consequence of the mapping cross. (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat.html)

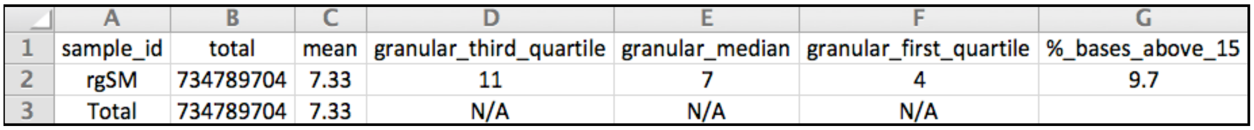
3. A set of ***uncovered regions*** (BED file) used to generate the annotated uncovered regions mentioned in the section above. This list of uncovered regions can be used in two ways. It can be further filtered by subtracting uncovered regions present in other samples using the ***CloudMap Uncovered Region Subtraction*** workflow to find uncovered regions unique to the sample under analysis. The resultant file can then be annotated using snpEff. Alternatively, these uncovered regions can be used to subtract from the set of uncovered regions in other samples (using *BEDtools*). (for more details see the ***Subtract Variants*** and ***Uncovered Region Subtraction*** workflows) (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat.html)

Examples of ***sample metric*** files (mentioned in section 22 above):

22.1) ***FASTQ quality statistics*** file (*Galaxy's FASTQ manipulation tools*)

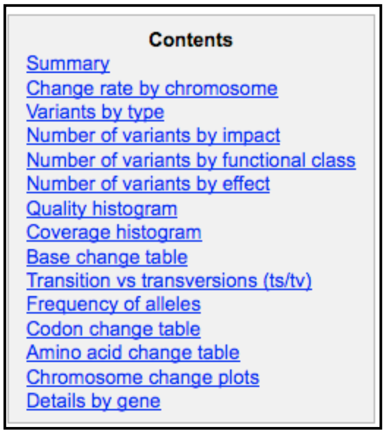


22.2) ***Depth of Coverage*** file (*GATK)*



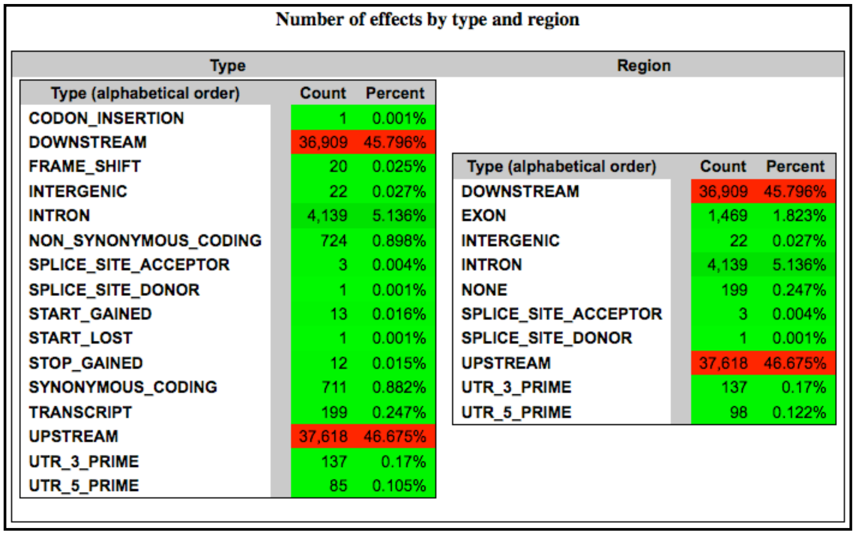
	A	B	C	D	E	F	G
1	sample_id	total	mean	granular_third_quartile	granular_median	granular_first_quartile	%_bases_above_15
2	rgSM	734789704	7.33	11	7	4	9.7
3	Total	734789704	7.33	N/A	N/A	N/A	

22.3) ***Graphical summary of all the variants*** in the sample (html file from *snpEff)*. Note: this file is very comprehensive and only excerpts of it are shown here:



Contents

Summary
Change rate by chromosome
Variants by type
Number of variants by impact
Number of variants by functional class
Number of variants by effect
Quality histogram
Coverage histogram
Base change table
Transition vs transversions (ts/tv)
Frequency of alleles
Codon change table
Amino acid change table
Chromosome change plots
Details by gene



Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
CODON_INSERTION	1	0.001%
DOWNSTREAM	36,909	45.796%
FRAME_SHIFT	20	0.025%
INTERGENIC	22	0.027%
INTRON	4,139	5.136%
NON_SYNONYMOUS_CODING	724	0.898%
SPLICE_SITE_ACCEPTOR	3	0.004%
SPLICE_SITE_DONOR	1	0.001%
START_GAINED	13	0.016%
START_LOST	1	0.001%
STOP_GAINED	12	0.015%
SYNONYMOUS_CODING	711	0.882%
TRANSCRIPT	199	0.247%
UPSTREAM	37,618	46.675%
UTR_3_PRIME	137	0.17%
UTR_5_PRIME	85	0.105%

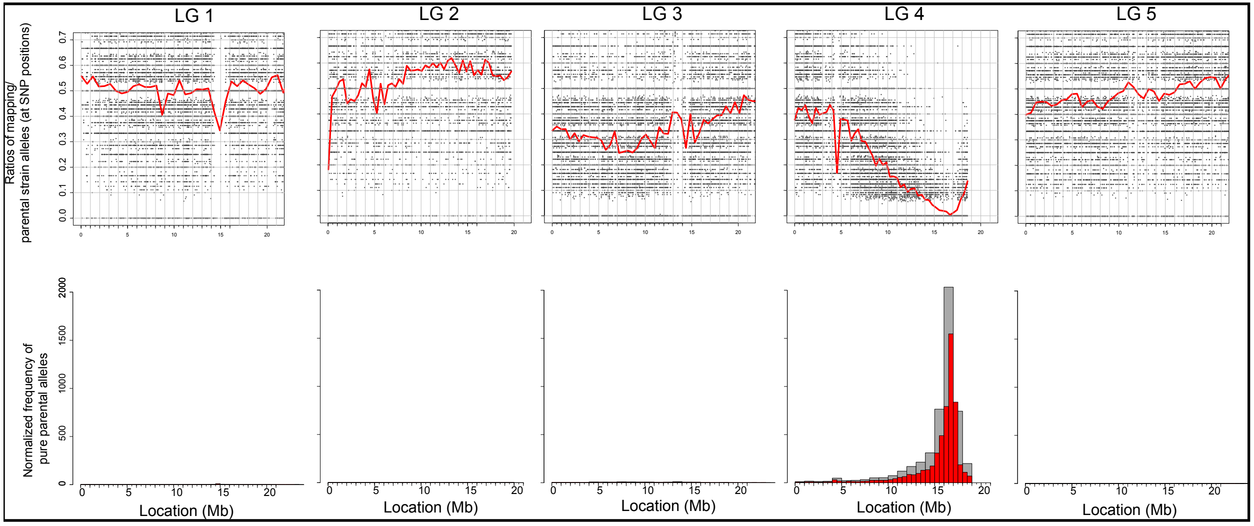
Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	36,909	45.796%
EXON	1,469	1.823%
INTERGENIC	22	0.027%
INTRON	4,139	5.136%
NONE	199	0.247%
SPLICE_SITE_ACCEPTOR	3	0.004%
SPLICE_SITE_DONOR	1	0.001%
UPSTREAM	37,618	46.675%
UTR_3_PRIME	137	0.17%
UTR_5_PRIME	98	0.122%

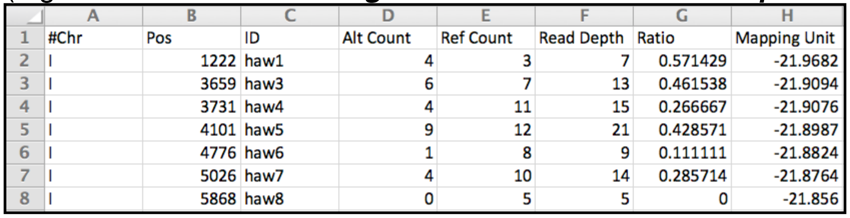
Examples of ***primary set of files for analysis*** (mentioned in step 23 above):

23.1) ***Hawaiian Variant Mapping plot*** and ***tabular file containing the data used to make the plots*** (*CloudMap*)

(e.g. ***Hawaiian Variant Mapping plot:*** Fig.10 *Arabidopsis*)



(e.g. ***Tabular file containing the data used to make the plots:*** *C. elegans*)



	A	B	C	D	E	F	G	H
1	#Chr	Pos	ID	Alt Count	Ref Count	Read Depth	Ratio	Mapping Unit
2	I	1222	haw1	4	3	7	0.571429	-21.9682
3	I	3659	haw3	6	7	13	0.461538	-21.9094
4	I	3731	haw4	4	11	15	0.266667	-21.9076
5	I	4101	haw5	9	12	21	0.428571	-21.8987
6	I	4776	haw6	1	8	9	0.111111	-21.8824
7	I	5026	haw7	4	10	14	0.285714	-21.8764
8	I	5868	haw8	0	5	5	0	-21.856

23.2) ***Annotated set of homozygous variants*** (Fig.4) (*snpEff*)

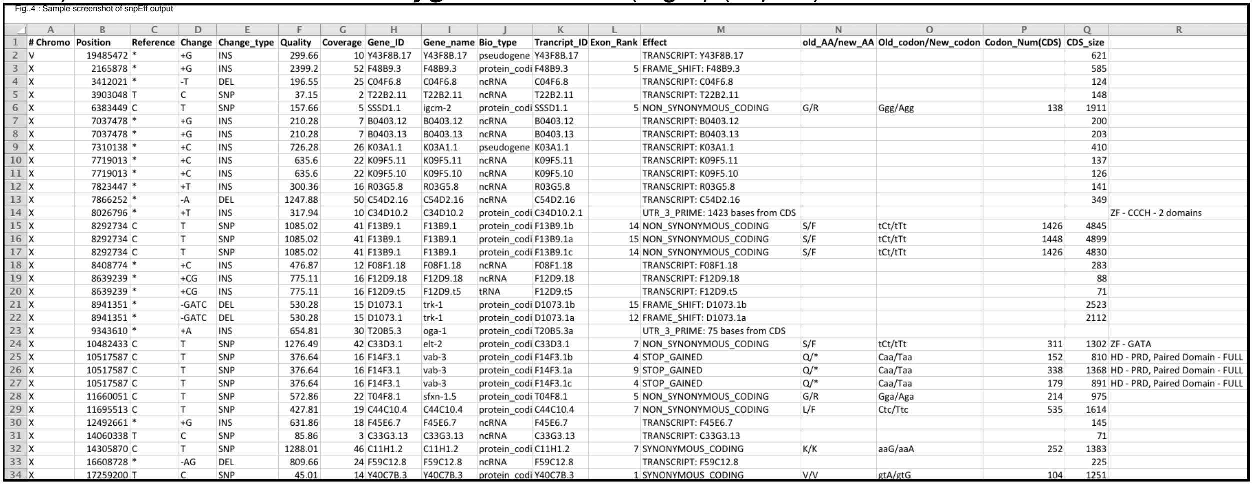
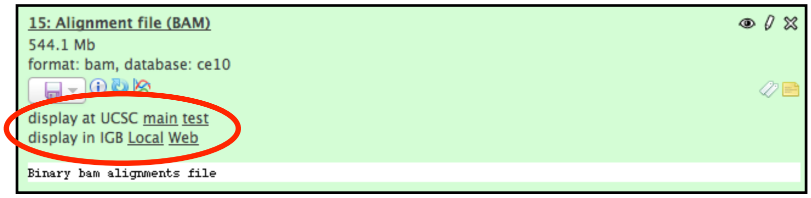


Fig..4 : Sample screenshot of snpEff output

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	# Chromo	Position	Reference	Change	Change_type	Quality	Coverage	Gene_ID	Gene_name	Bio_type	Trancript_ID	Exon_Rank	Effect	old_AA/new_AA	Old_codon/New_codon	Codon_Num(CDS)	CDS_size	
2	V	19485472	*	+G	INS	299.66	10	Y43F8B.17	Y43F8B.17	pseudogene	Y43F8B.17		TRANSCRIPT: Y43F8B.17				621	
3	X	2165878	*	+G	INS	2399.2	52	F48B9.3	F48B9.3	protein_codi	F48B9.3	5	FRAME_SHIFT: F48B9.3				585	
4	X	3412021	*	-T	DEL	196.55	25	C04F6.8	C04F6.8	ncRNA	C04F6.8		TRANSCRIPT: C04F6.8				124	
5	X	3903048	T	C	SNP	37.15	2	T22B2.11	T22B2.11	ncRNA	T22B2.11		TRANSCRIPT: T22B2.11				148	
6	X	6383449	C	T	SNP	157.66	5	SSSD1.1	igcm-2	protein_codi	SSSD1.1	5	NON_SYNONYMOUS_CODING	G/R	Ggg/Agg	138	1911	
7	X	7037478	*	+G	INS	210.28	7	B0403.12	B0403.12	ncRNA	B0403.12		TRANSCRIPT: B0403.12				200	
8	X	7037478	*	+G	INS	210.28	7	B0403.13	B0403.13	ncRNA	B0403.13		TRANSCRIPT: B0403.13				203	
9	X	7310138	*	+C	INS	726.28	26	K03A1.1	K03A1.1	pseudogene	K03A1.1		TRANSCRIPT: K03A1.1				410	
10	X	7719013	*	+C	INS	635.6	22	K09F5.11	K09F5.11	ncRNA	K09F5.11		TRANSCRIPT: K09F5.11				137	
11	X	7719013	*	+C	INS	635.6	22	K09F5.10	K09F5.10	ncRNA	K09F5.10		TRANSCRIPT: K09F5.10				126	
12	X	7823447	*	+T	INS	300.36	16	R03G5.8	R03G5.8	ncRNA	R03G5.8		TRANSCRIPT: R03G5.8				141	
13	X	7866252	*	-A	DEL	1247.88	50	C54D2.16	C54D2.16	ncRNA	C54D2.16		TRANSCRIPT: C54D2.16				349	
14	X	8026796	*	+T	INS	317.94	10	C34D10.2	C34D10.2	protein_codi	C34D10.2.1		UTR_3_PRIME: 1423 bases from CDS					ZF - CCCH - 2 domains
15	X	8292734	C	T	SNP	1085.02	41	F13B9.1	F13B9.1	protein_codi	F13B9.1b	14	NON_SYNONYMOUS_CODING	S/F	tCt/tTt	1426	4845	
16	X	8292734	C	T	SNP	1085.02	41	F13B9.1	F13B9.1	protein_codi	F13B9.1a	15	NON_SYNONYMOUS_CODING	S/F	tCt/tTt	1448	4899	
17	X	8292734	C	T	SNP	1085.02	41	F13B9.1	F13B9.1	protein_codi	F13B9.1c	14	NON_SYNONYMOUS_CODING	S/F	tCt/tTt	1426	4830	
18	X	8408774	*	+C	INS	476.87	12	F08F1.18	F08F1.18	ncRNA	F08F1.18		TRANSCRIPT: F08F1.18				283	
19	X	8639239	*	+CG	INS	775.11	16	F12D9.18	F12D9.18	ncRNA	F12D9.18		TRANSCRIPT: F12D9.18				88	
20	X	8639239	*	+CG	INS	775.11	16	F12D9.t5	F12D9.t5	tRNA	F12D9.t5		TRANSCRIPT: F12D9.t5				71	
21	X	8941351	*	-GATC	DEL	530.28	15	D1073.1	trk-1	protein_codi	D1073.1b	15	FRAME_SHIFT: D1073.1b				2523	
22	X	8941351	*	-GATC	DEL	530.28	15	D1073.1	trk-1	protein_codi	D1073.1a	12	FRAME_SHIFT: D1073.1a				2112	
23	X	9343610	*	+A	INS	654.81	30	T20B5.3	oga-1	protein_codi	T20B5.3a		UTR_3_PRIME: 75 bases from CDS					
24	X	10482433	C	T	SNP	1276.49	42	C33D3.1	elt-2	protein_codi	C33D3.1	7	NON_SYNONYMOUS_CODING	S/F	tCt/tTt	311	1302	ZF - GATA
25	X	10517587	C	T	SNP	376.64	16	F14F3.1	vab-3	protein_codi	F14F3.1b	4	STOP_GAINED	Q/*	Caa/Taa	152	810	HD - PRD, Paired Domain - FULL
26	X	10517587	C	T	SNP	376.64	16	F14F3.1	vab-3	protein_codi	F14F3.1a	9	STOP_GAINED	Q/*	Caa/Taa	338	1368	HD - PRD, Paired Domain - FULL
27	X	10517587	C	T	SNP	376.64	16	F14F3.1	vab-3	protein_codi	F14F3.1c	4	STOP_GAINED	Q/*	Caa/Taa	179	891	HD - PRD, Paired Domain - FULL
28	X	11660051	C	T	SNP	572.86	22	T04F8.1	sfxn-1.5	protein_codi	T04F8.1	5	NON_SYNONYMOUS_CODING	G/R	Gga/Aga	214	975	
29	X	11695513	C	T	SNP	427.81	19	C44C10.4	C44C10.4	protein_codi	C44C10.4	7	NON_SYNONYMOUS_CODING	L/F	Ctc/Ttc	535	1614	
30	X	12492661	*	+G	INS	631.86	18	F45E6.7	F45E6.7	ncRNA	F45E6.7		TRANSCRIPT: F45E6.7				145	
31	X	14060338	T	C	SNP	85.86	3	C33G3.13	C33G3.13	ncRNA	C33G3.13		TRANSCRIPT: C33G3.13				71	
32	X	14305870	C	T	SNP	1288.01	46	C11H1.2	C11H1.2	protein_codi	C11H1.2	7	SYNONYMOUS_CODING	K/K	aaG/aaA	252	1383	
33	X	16608728	*	-AG	DEL	809.66	24	F59C12.8	F59C12.8	ncRNA	F59C12.8		TRANSCRIPT: F59C12.8				225	
34	X	17259200	T	C	SNP	45.01	14	Y40C7B.3	Y40C7B.3	protein_codi	Y40C7B.3	1	SYNONYMOUS_CODING	V/V	gtA/gtG	104	1251	

23.3) ***BAM alignment*** file (*SAMtools*) (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat.html)

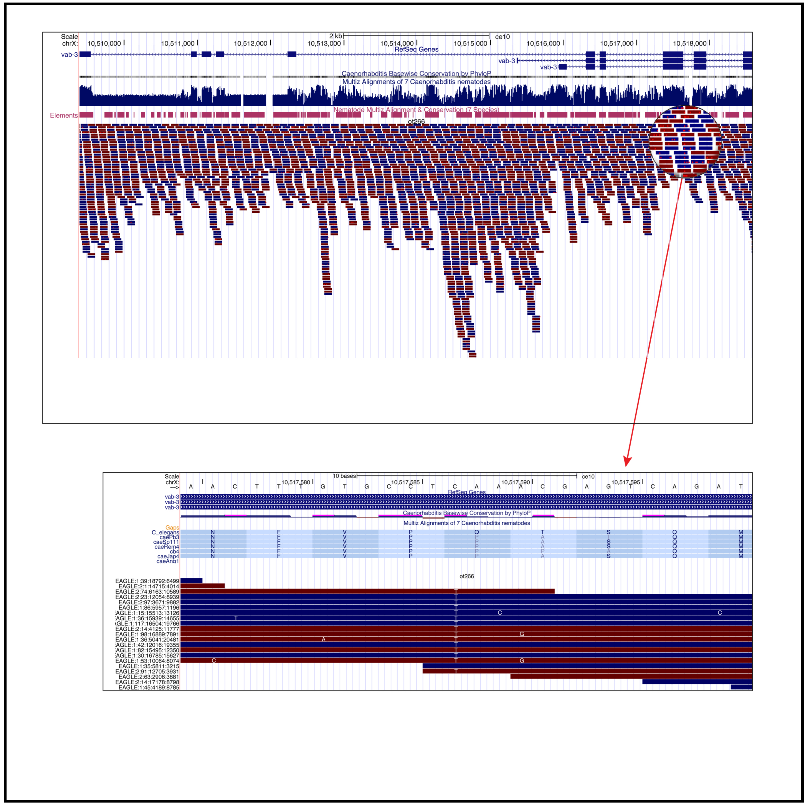
Click on the "***display in"*** link in your history or download the BAM file to view it in your alignment viewer of choice:



15: Alignment file (BAM)
544.1 Mb
format: bam, database: ce10
display at UCSC main test
display in IGB Local Web

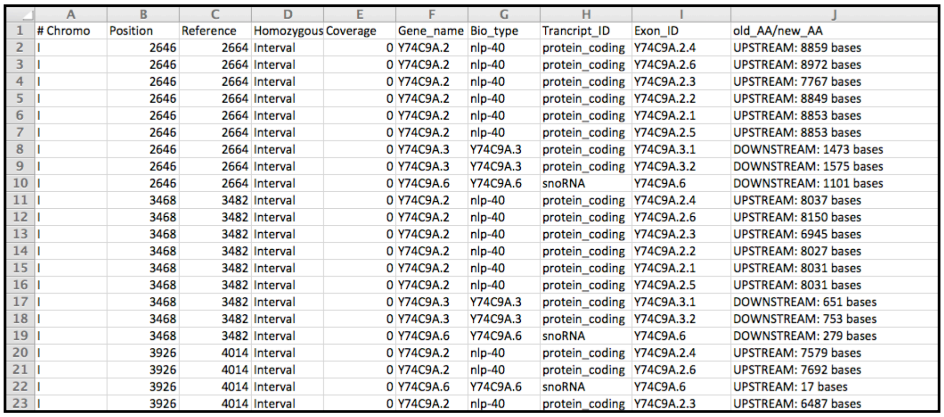
```
Binary bam alignments file
```

(e.g. Fig.9 UCSC Genome Browser)


Scale
chrX:
2 kb
ce10
vab-3
RefSeq Genes
Caenorhabditis Basewise Conservation by PhyloP
Multiz Alignments of 7 Caenorhabditis nematodes
Nematode Multiz Alignment & Conservation (7 Species)
Elements
ot266
10 bases
Gaps

Note: Information displayed in alignment viewers often will not exactly match that in variant files (VCFs) or lists of annotated variants (snpEff). This is because read mapping qualities and base qualities are incorporated into which variants are ultimately called. Most alignment viewers have filter settings that can be used to only display reads with mapping quality scores above a certain value. Applying these filters should result in alignments that more closely approximate variant lists.

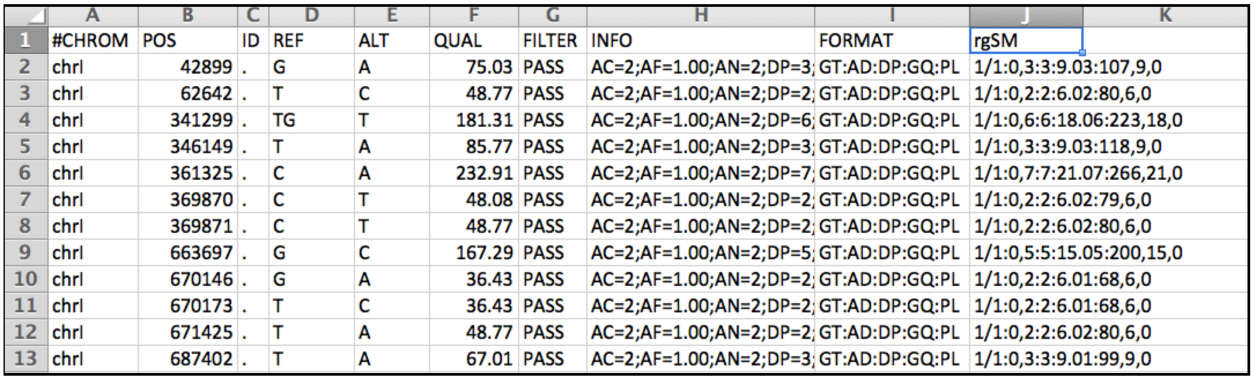
23.4) A list of ***annotated uncovered regions*** (BED file) (*BEDtools* & *snpEff)* (For more information on file format, see: http://snpeff.sourceforge.net/)



	A	B	C	D	E	F	G	H	I	J
1	# Chromo	Position	Reference	Homozygous	Coverage	Gene_name	Bio_type	Trancript_ID	Exon_ID	old_AA/new_AA
2	I	2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.4	UPSTREAM: 8859 bases
3	I	2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.6	UPSTREAM: 8972 bases
4	I	2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.3	UPSTREAM: 7767 bases
5	I	2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.2	UPSTREAM: 8849 bases
6	I	2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.1	UPSTREAM: 8853 bases
7	I	2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.5	UPSTREAM: 8853 bases
8	I	2646	2664	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.1	DOWNSTREAM: 1473 bases
9	I	2646	2664	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.2	DOWNSTREAM: 1575 bases
10	I	2646	2664	Interval	0	Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6	DOWNSTREAM: 1101 bases
11	I	3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.4	UPSTREAM: 8037 bases
12	I	3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.6	UPSTREAM: 8150 bases
13	I	3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.3	UPSTREAM: 6945 bases
14	I	3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.2	UPSTREAM: 8027 bases
15	I	3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.1	UPSTREAM: 8031 bases
16	I	3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.5	UPSTREAM: 8031 bases
17	I	3468	3482	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.1	DOWNSTREAM: 651 bases
18	I	3468	3482	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.2	DOWNSTREAM: 753 bases
19	I	3468	3482	Interval	0	Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6	DOWNSTREAM: 279 bases
20	I	3926	4014	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.4	UPSTREAM: 7579 bases
21	I	3926	4014	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.6	UPSTREAM: 7692 bases
22	I	3926	4014	Interval	0	Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6	UPSTREAM: 17 bases
23	I	3926	4014	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.3	UPSTREAM: 6487 bases

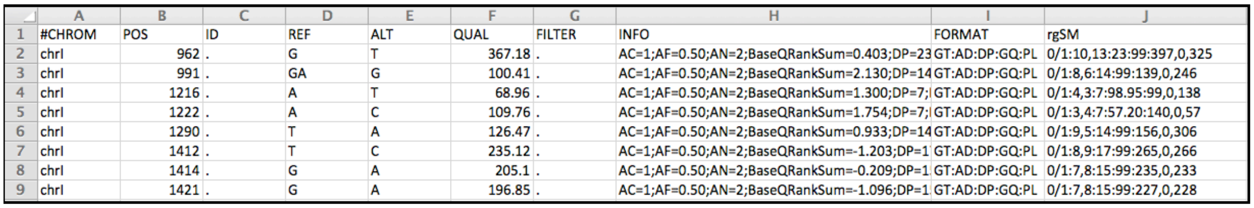
Additional files that can be used for ***downstream subtraction workflows*** (mentioned in step 24 above):

24.1) ***Set of homozygous variants*** (VCF file generated by *GATK*). Header lines starting with "#" have been removed in Excel. (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat)



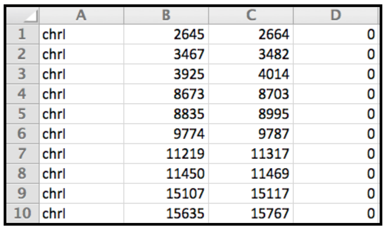
	A	B	C	D	E	F	G	H	I	J	K
1	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	rgSM	
2	chrI	42899	.	G	A	75.03	PASS	AC=2;AF=1.00;AN=2;DP=3;	GT:AD:DP:GQ:PL	1/1:0,3:3:9.03:107,9,0	
3	chrI	62642	.	T	C	48.77	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:80,6,0	
4	chrI	341299	.	TG	T	181.31	PASS	AC=2;AF=1.00;AN=2;DP=6;	GT:AD:DP:GQ:PL	1/1:0,6:6:18.06:223,18,0	
5	chrI	346149	.	T	A	85.77	PASS	AC=2;AF=1.00;AN=2;DP=3;	GT:AD:DP:GQ:PL	1/1:0,3:3:9.03:118,9,0	
6	chrI	361325	.	C	A	232.91	PASS	AC=2;AF=1.00;AN=2;DP=7;	GT:AD:DP:GQ:PL	1/1:0,7:7:21.07:266,21,0	
7	chrI	369870	.	C	T	48.08	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:79,6,0	
8	chrI	369871	.	C	T	48.77	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:80,6,0	
9	chrI	663697	.	G	C	167.29	PASS	AC=2;AF=1.00;AN=2;DP=5;	GT:AD:DP:GQ:PL	1/1:0,5:5:15.05:200,15,0	
10	chrI	670146	.	G	A	36.43	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.01:68,6,0	
11	chrI	670173	.	T	C	36.43	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.01:68,6,0	
12	chrI	671425	.	T	A	48.77	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:80,6,0	
13	chrI	687402	.	T	A	67.01	PASS	AC=2;AF=1.00;AN=2;DP=3;	GT:AD:DP:GQ:PL	1/1:0,3:3:9.01:99,9,0	

24.2) ***Set of homozygous and heterozygous variants*** (VCF file generated by *GATK*). Header lines starting with "#" have been removed in Excel. (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat)



	A	B	C	D	E	F	G	H	I	J
1	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	rgSM
2	chrI	962	.	G	T	367.18	.	AC=1;AF=0.50;AN=2;BaseQRankSum=0.403;DP=23	GT:AD:DP:GQ:PL	0/1:10,13:23:99:397,0,325
3	chrI	991	.	GA	G	100.41	.	AC=1;AF=0.50;AN=2;BaseQRankSum=2.130;DP=14	GT:AD:DP:GQ:PL	0/1:8,6:14:99:139,0,246
4	chrI	1216	.	A	T	68.96	.	AC=1;AF=0.50;AN=2;BaseQRankSum=1.300;DP=7;	GT:AD:DP:GQ:PL	0/1:4,3:7:98.95:99,0,138
5	chrI	1222	.	A	C	109.76	.	AC=1;AF=0.50;AN=2;BaseQRankSum=1.754;DP=7;	GT:AD:DP:GQ:PL	0/1:3,4:7:57.20:140,0,57
6	chrI	1290	.	T	A	126.47	.	AC=1;AF=0.50;AN=2;BaseQRankSum=0.933;DP=14	GT:AD:DP:GQ:PL	0/1:9,5:14:99:156,0,306
7	chrI	1412	.	T	C	235.12	.	AC=1;AF=0.50;AN=2;BaseQRankSum=-1.203;DP=1	GT:AD:DP:GQ:PL	0/1:8,9:17:99:265,0,266
8	chrI	1414	.	G	A	205.1	.	AC=1;AF=0.50;AN=2;BaseQRankSum=-0.209;DP=1	GT:AD:DP:GQ:PL	0/1:7,8:15:99:235,0,233
9	chrI	1421	.	G	A	196.85	.	AC=1;AF=0.50;AN=2;BaseQRankSum=-1.096;DP=1	GT:AD:DP:GQ:PL	0/1:7,8:15:99:227,0,228

24.3) ***Set of uncovered regions (***BED file***)*** (*BEDtools*). (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat)



	A	B	C	D
1	chrI	2645	2664	0
2	chrI	3467	3482	0
3	chrI	3925	4014	0
4	chrI	8673	8703	0
5	chrI	8835	8995	0
6	chrI	9774	9787	0
7	chrI	11219	11317	0
8	chrI	11450	11469	0
9	chrI	15107	15117	0
10	chrI	15635	15767	0

Note: We strongly suggest that users employ the ***Subtract Variants*** and ***Uncovered Region Subtraction*** workflows if additional strains are available for this purpose. The general concept is shown in **Fig.5** of the CloudMap paper.

CloudMap | Cloud-based Pipeline for Analysis of Mutant Genome Sequences

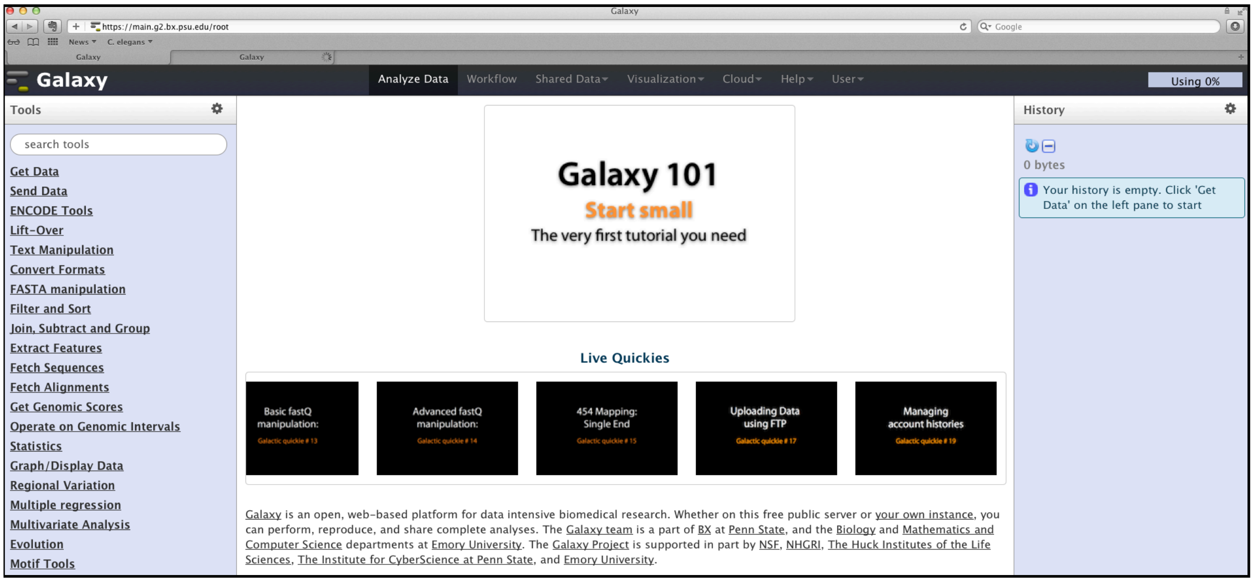
CloudMap UnMapped Mutant Workflow

This workflow performs the same analysis as the ***Hawaiian Variant Mapping with WGS data and Variant Calling workflow*** without the mapping-specific tools and input reference files. The workflow should be used for data generated from a single mutant, not from pooled mutants resulting from a cross to a mapping strain. This workflow uses single-end FASTQ data but it can be adapted to use paired-end data (see the ***Analyzing Your Own Data*** section of this user guide). A video version of this user guide is available at: http://usegalaxy.org/cloudmap.

These workflows provide default function parameters, ensuring that users follow best practices, and allow for automated execution of sequential operations. We provide these workflows as helpful guides, but experienced users may execute functions in any meaningful order they please and may also create and share their own workflows to take advantage of the automation feature. More CloudMap documentation is available at http://usegalaxy.org/cloudmap.

The *ot266* FASTQ file used in this example represents Hawaiian variant mapped data but for the purposes of this user guide, we perform an unmapped analysis. Users wishing to run their own unmapped data should also view the ***Analyzing Your Own Data*** section of this user guide before proceeding.

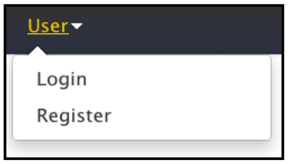
1) Navigate to http://usegalaxy.org (URL will resolve to something like https://main.g2.bx.psu.edu)



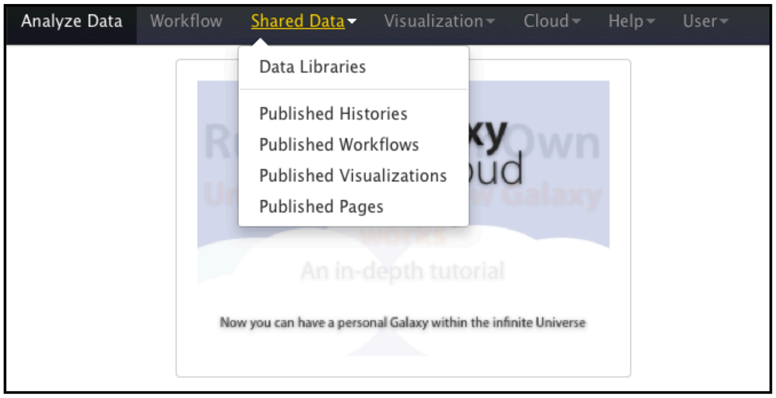
Your history is empty. Click 'Get Data' on the left pane to start

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

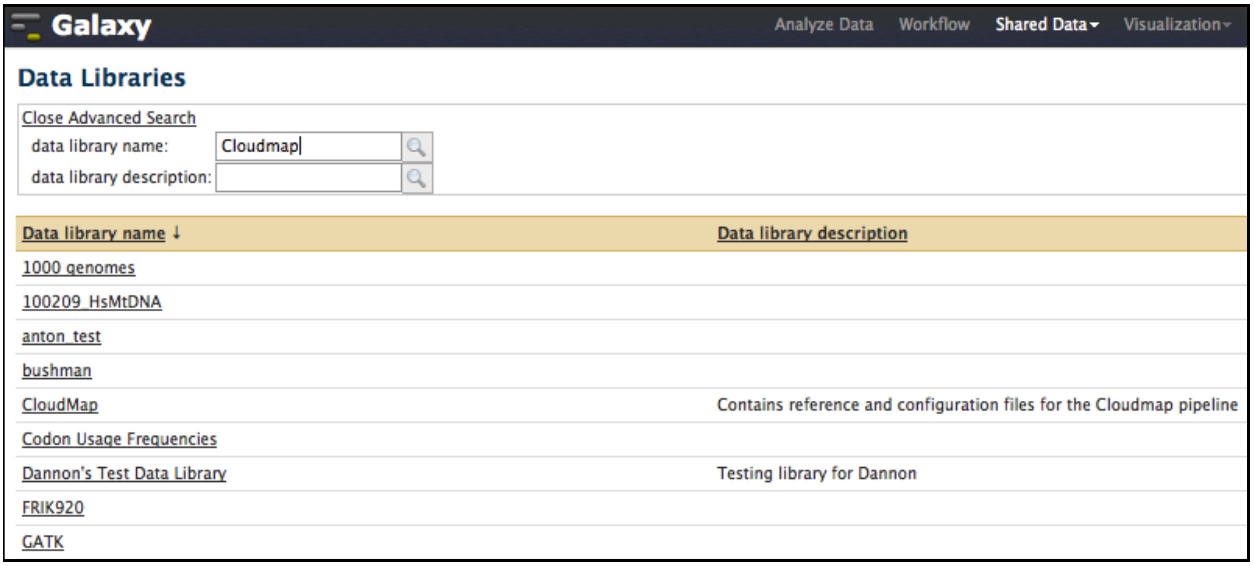
2) Register for an account or login if you already have an account:



3) Once you are logged in using your email address, click on the ***Shared Data*** link at the top of the page:



4) Click on ***Data Libraries*** and search for the CloudMap data library:



Data Libraries

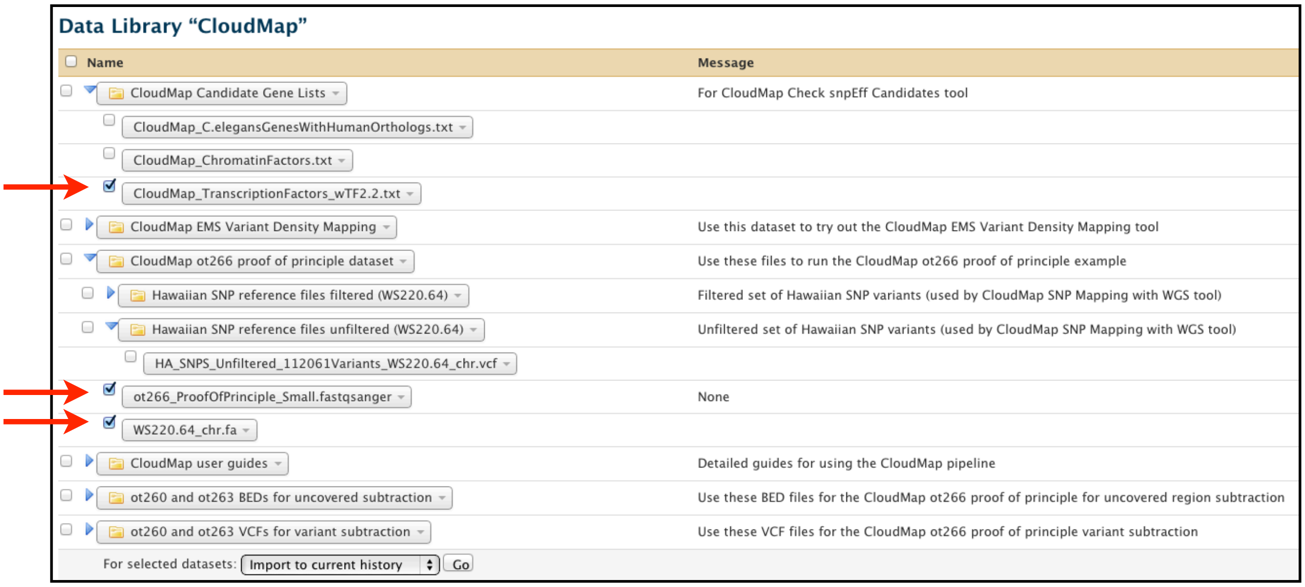
Close Advanced Search

data library name: Cloudmap

data library description:

Data library name ↓	Data library description
1000 genomes	
100209_HsMtDNA	
anton_test	
bushman	
CloudMap	Contains reference and configuration files for the Cloudmap pipeline
Codon Usage Frequencies	
Dannon's Test Data Library	Testing library for Dannon
FRIK920	
GATK	

5) Click on the ***CloudMap*** library and select the 5 data files below for the *ot266* example. Then click “Go” to import these files into your history.

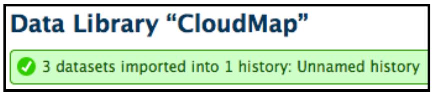


Data Library “CloudMap”

☐ Name	Message
☐ CloudMap Candidate Gene Lists	For CloudMap Check snpEff Candidates tool
☐ CloudMap_C.elegansGenesWithHumanOrthologs.txt	
☐ CloudMap_ChromatinFactors.txt	
☑ CloudMap_TranscriptionFactors_wTF2.2.txt	
☐ CloudMap EMS Variant Density Mapping	Use this dataset to try out the CloudMap EMS Variant Density Mapping tool
☐ CloudMap ot266 proof of principle dataset	Use these files to run the CloudMap ot266 proof of principle example
☐ Hawaiian SNP reference files filtered (WS220.64)	Filtered set of Hawaiian SNP variants (used by CloudMap SNP Mapping with WGS tool)
☐ Hawaiian SNP reference files unfiltered (WS220.64)	Unfiltered set of Hawaiian SNP variants (used by CloudMap SNP Mapping with WGS tool)
☐ HA_SNPS_Unfiltered_112061Variants_WS220.64_chr.vcf	
☑ ot266_ProofOfPrinciple_Small.fastqsanger	None
☑ WS220.64_chr.fa	
☐ CloudMap user guides	Detailed guides for using the CloudMap pipeline
☐ ot260 and ot263 BEDs for uncovered subtraction	Use these BED files for the CloudMap ot266 proof of principle for uncovered region subtraction
☐ ot260 and ot263 VCFs for variant subtraction	Use these VCF files for the CloudMap ot266 proof of principle variant subtraction

For selected datasets: Import to current history Go

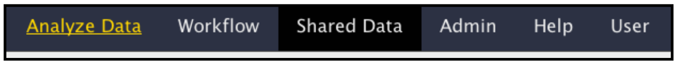
6) You will receive confirmation that the files have been imported into your history:



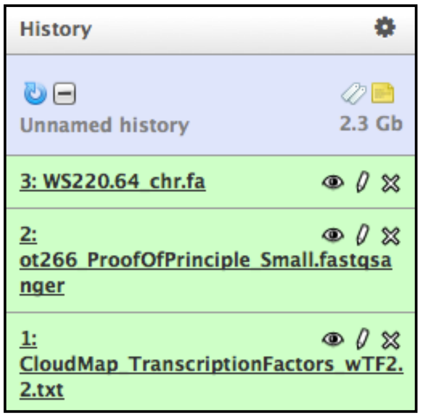
Data Library “CloudMap”

✓ 3 datasets imported into 1 history: Unnamed history

7) Click ***Analyze Data*** on the menu bar to navigate to your history:



8) You will now see that the data files have been added to an unnamed history:



History

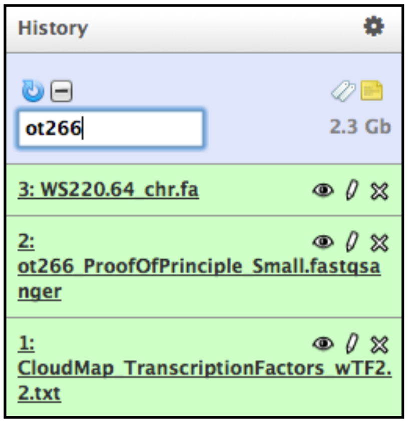
Unnamed history 2.3 Gb

3: WS220.64_chr.fa

2: ot266_ProofOfPrinciple_Small.fastqsanger

1: CloudMap_TranscriptionFactors_wTF2.2.txt

9) Name your history ***ot266*** after the sample that we will be analyzing:



History

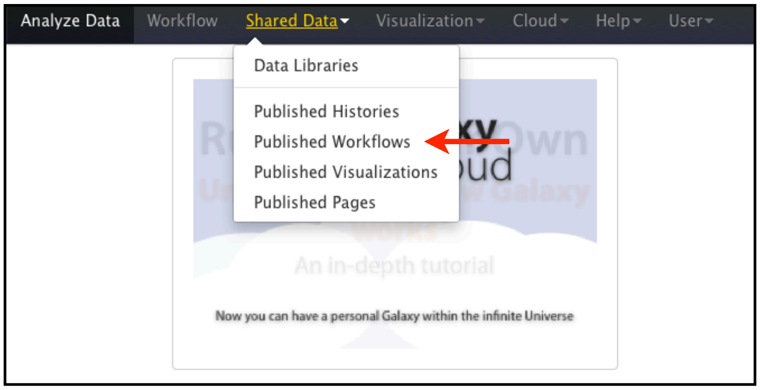
ot266 2.3 Gb

3: WS220.64_chr.fa

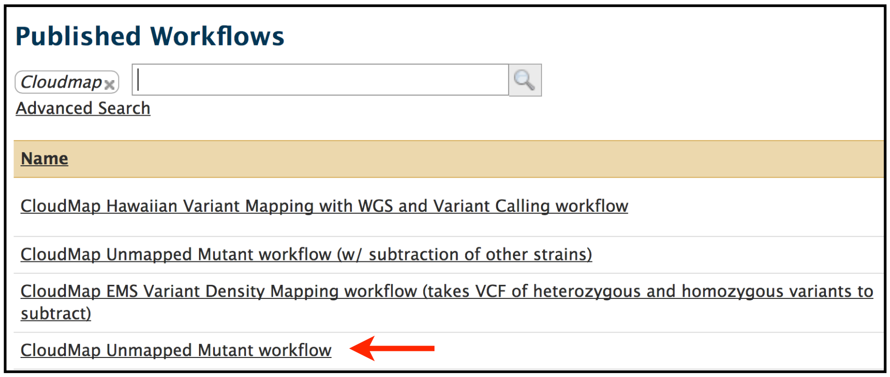
2: ot266_ProofOfPrinciple_Small.fastqsanger

1: CloudMap_TranscriptionFactors_wTF2.2.txt

10) Again click on the ***Shared Data*** link at the top of the page and select ***Published Workflows*** :



11) Use the search term "CloudMap" to view the automated workflows. Select the ***CloudMap Unmapped Mutant workflow.***



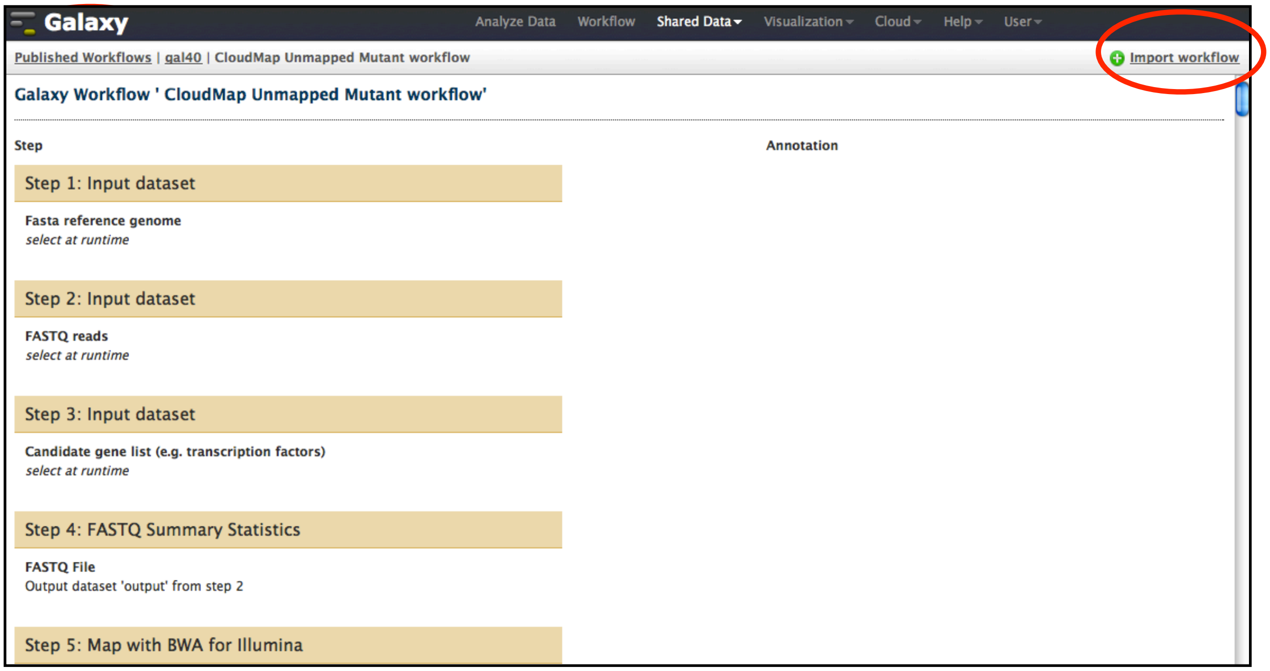
Published Workflows

Cloudmap

Advanced Search

Name
CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow
CloudMap Unmapped Mutant workflow (w/ subtraction of other strains)
CloudMap EMS Variant Density Mapping workflow (takes VCF of heterozygous and homozygous variants to subtract)
CloudMap Unmapped Mutant workflow

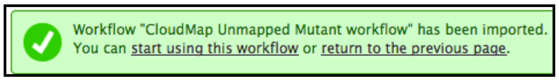
12) You will now have the option to ***Import workflow***



Galaxy Workflow ' CloudMap Unmapped Mutant workflow'

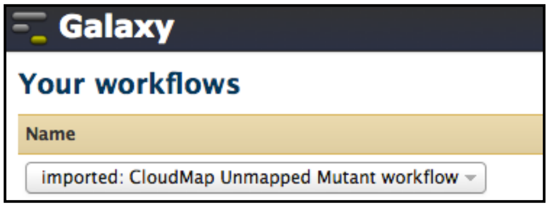
Step	Annotation
Step 1: Input dataset	
Fasta reference genome *select at runtime*	
Step 2: Input dataset	
FASTQ reads *select at runtime*	
Step 3: Input dataset	
Candidate gene list (e.g. transcription factors) *select at runtime*	
Step 4: FASTQ Summary Statistics	
FASTQ File Output dataset 'output' from step 2	
Step 5: Map with BWA for Illumina	

13) You will see the message below. Click ***Start using this workflow***.



Workflow "CloudMap Unmapped Mutant workflow" has been imported.
You can start using this workflow or return to the previous page.

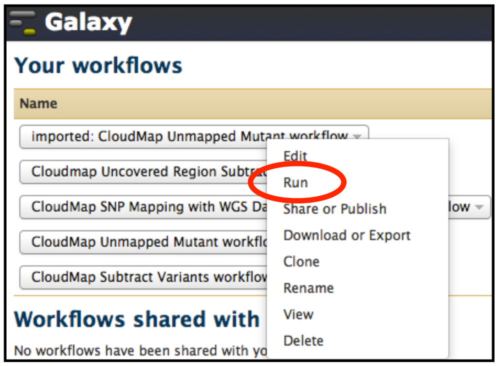
14) You will see that the workflow has been imported. From now on, you can easily access this workflow under the ***Workflow*** tab.



Your workflows

Name
imported: CloudMap Unmapped Mutant workflow

15) Click on the workflow and select ***Run***:



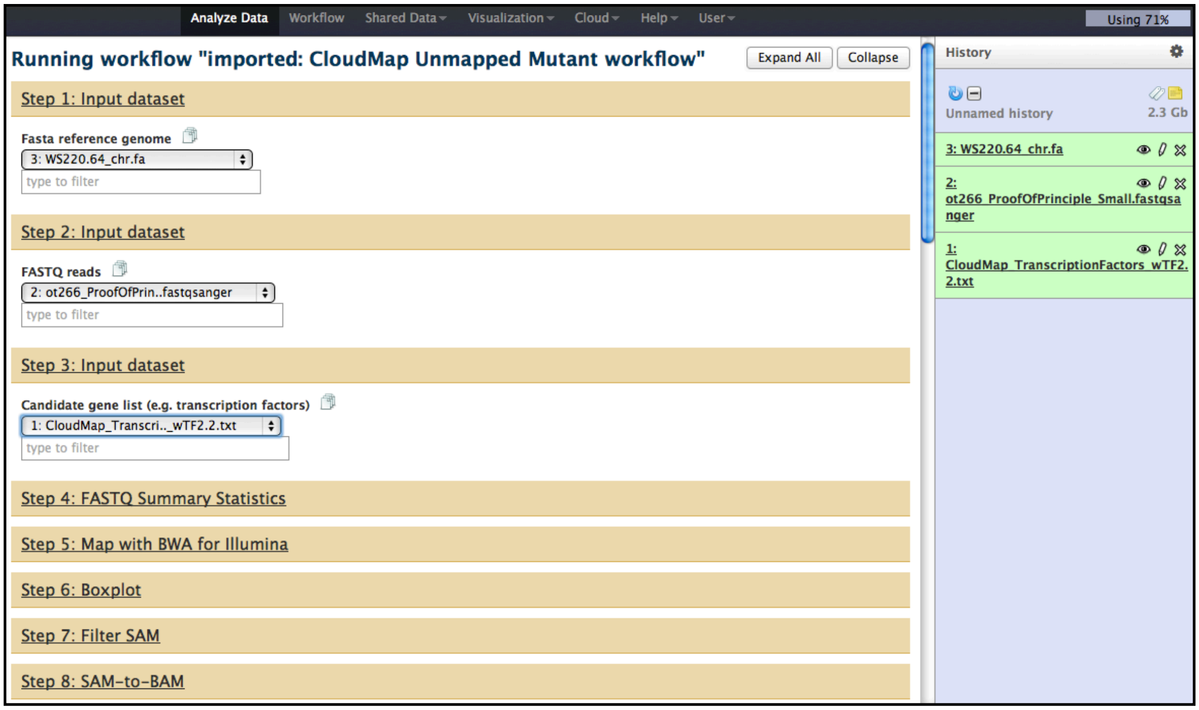
Your workflows

Name
imported: CloudMap Unmapped Mutant workflow
Cloudmap Uncovered Region Subt
CloudMap SNP Mapping with WGS Da ... low
CloudMap Unmapped Mutant workflo
CloudMap Subtract Variants workflov

Workflows shared with

No workflows have been shared with yo

16) You will see all the steps in the workflow prior to running it. Make sure that each of the input fields corresponds to the appropriate file in your history.



Running workflow "imported: CloudMap Unmapped Mutant workflow"

Step 1: Input dataset

Fasta reference genome: 3: WS220.64_chr.fa

type to filter

Step 2: Input dataset

FASTQ reads: 2: ot266_ProofOfPrin..fastqsanger

type to filter

Step 3: Input dataset

Candidate gene list (e.g. transcription factors): 1: CloudMap_Transcri..._wTF2.2.txt

type to filter

Step 4: FASTQ Summary Statistics

Step 5: Map with BWA for Illumina

Step 6: Boxplot

Step 7: Filter SAM

Step 8: SAM-to-BAM

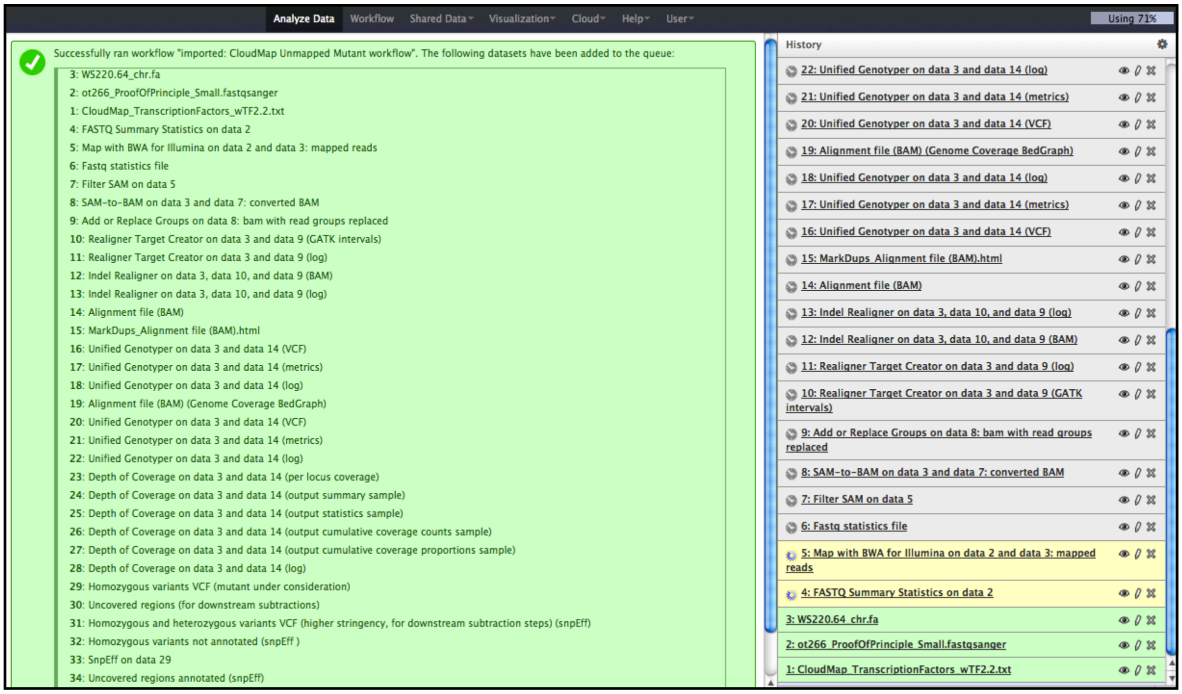
History:

3: WS220.64_chr.fa

2: ot266_ProofOfPrinciple_Small.fastqsanger

1: CloudMap_TranscriptionFactors_wTF2.2.txt

17) All of the automated functions have the appropriate default parameters configured, although experienced users may want to modify these prior to running (see the ***Analyzing Your Own Data Using CloudMap Workflows*** section of this user guide). Once you are ready to run the workflow, press ***Run Workflow*** at the bottom of the page and the workflow will start (this step takes a minute or two to begin, be patient and don't hit the ***Run Workflow*** button repeatedly). You will receive an email when the workflow is completed:



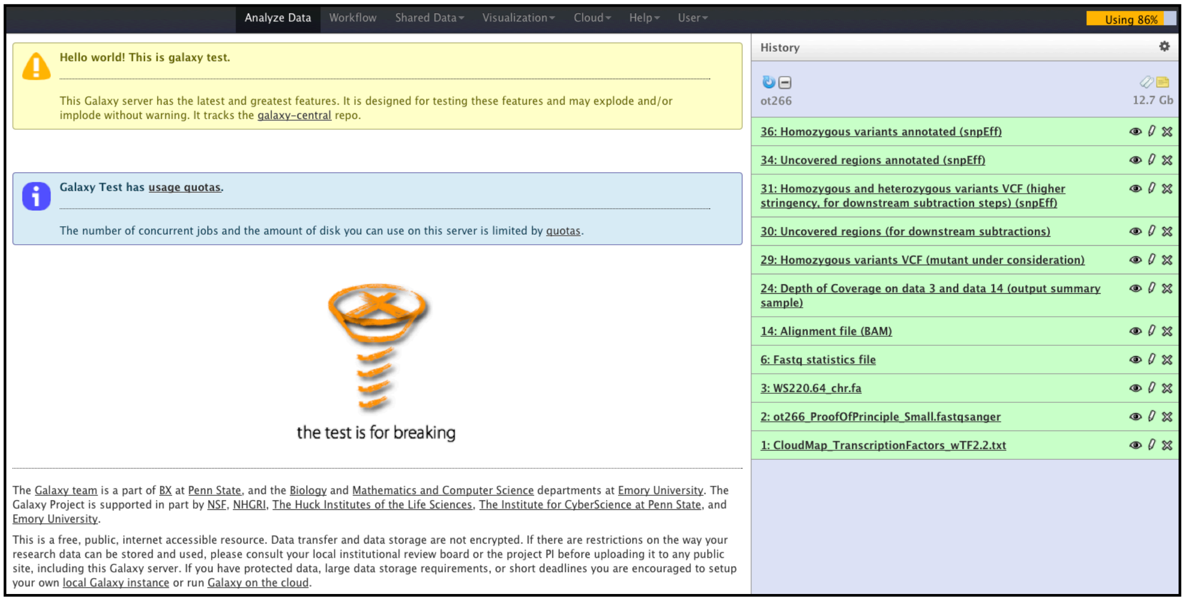
Successfully ran workflow "imported: CloudMap Unmapped Mutant workflow". The following datasets have been added to the queue:

3: WS220.64_chr.fa
2: ot266_ProofOfPrinciple_Small.fastqsanger
1: CloudMap_TranscriptionFactors_wTF2.2.txt
4: FASTQ Summary Statistics on data 2
5: Map with BWA for Illumina on data 2 and data 3: mapped reads
6: Fastq statistics file
7: Filter SAM on data 5
8: SAM-to-BAM on data 3 and data 7: converted BAM
9: Add or Replace Groups on data 8: bam with read groups replaced
10: Realigner Target Creator on data 3 and data 9 (GATK intervals)
11: Realigner Target Creator on data 3 and data 9 (log)
12: Indel Realigner on data 3, data 10, and data 9 (BAM)
13: Indel Realigner on data 3, data 10, and data 9 (log)
14: Alignment file (BAM)
15: MarkDups_Alignment file (BAM).html
16: Unified Genotyper on data 3 and data 14 (VCF)
17: Unified Genotyper on data 3 and data 14 (metrics)
18: Unified Genotyper on data 3 and data 14 (log)
19: Alignment file (BAM) (Genome Coverage BedGraph)
20: Unified Genotyper on data 3 and data 14 (VCF)
21: Unified Genotyper on data 3 and data 14 (metrics)
22: Unified Genotyper on data 3 and data 14 (log)
23: Depth of Coverage on data 3 and data 14 (per locus coverage)
24: Depth of Coverage on data 3 and data 14 (output summary sample)
25: Depth of Coverage on data 3 and data 14 (output statistics sample)
26: Depth of Coverage on data 3 and data 14 (output cumulative coverage counts sample)
27: Depth of Coverage on data 3 and data 14 (output cumulative coverage proportions sample)
28: Depth of Coverage on data 3 and data 14 (log)
29: Homozygous variants VCF (mutant under consideration)
30: Uncovered regions (for downstream subtractions)
31: Homozygous and heterozygous variants VCF (higher stringency, for downstream subtraction steps) (snpEff)
32: Homozygous variants not annotated (snpEff)
33: SnpEff on data 29
34: Uncovered regions annotated (snpEff)

History:

22: Unified Genotyper on data 3 and data 14 (log)
21: Unified Genotyper on data 3 and data 14 (metrics)
20: Unified Genotyper on data 3 and data 14 (VCF)
19: Alignment file (BAM) (Genome Coverage BedGraph)
18: Unified Genotyper on data 3 and data 14 (log)
17: Unified Genotyper on data 3 and data 14 (metrics)
16: Unified Genotyper on data 3 and data 14 (VCF)
15: MarkDups_Alignment file (BAM).html
14: Alignment file (BAM)
13: Indel Realigner on data 3, data 10, and data 9 (log)
12: Indel Realigner on data 3, data 10, and data 9 (BAM)
11: Realigner Target Creator on data 3 and data 9 (log)
10: Realigner Target Creator on data 3 and data 9 (GATK intervals)
9: Add or Replace Groups on data 8: bam with read groups replaced
8: SAM-to-BAM on data 3 and data 7: converted BAM
7: Filter SAM on data 5
6: Fastq statistics file
5: Map with BWA for Illumina on data 2 and data 3: mapped reads
4: FASTQ Summary Statistics on data 2
3: WS220.64_chr.fa
2: ot266_ProofOfPrinciple_Small.fastqsanger
1: CloudMap_TranscriptionFactors_wTF2.2.txt

18) Once the workflow has finished running, you can view the resulting output:



Hello world! This is galaxy test.

This Galaxy server has the latest and greatest features. It is designed for testing these features and may explode and/or implode without warning. It tracks the galaxy-central repo.

Galaxy Test has usage quotas.

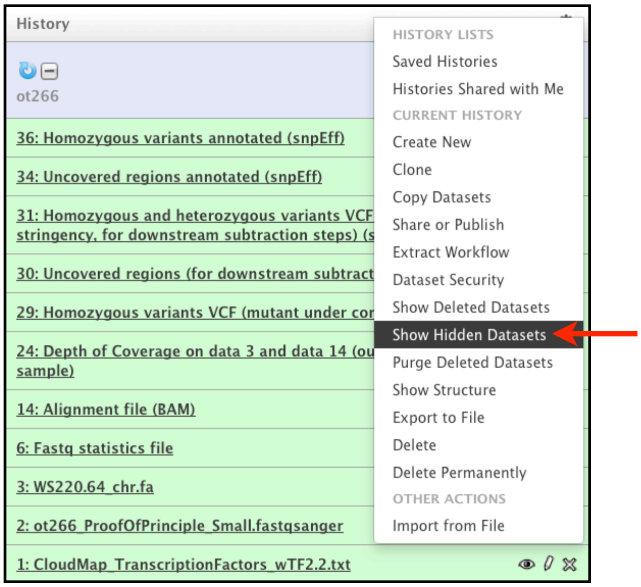
The number of concurrent jobs and the amount of disk you can use on this server is limited by quotas.

The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

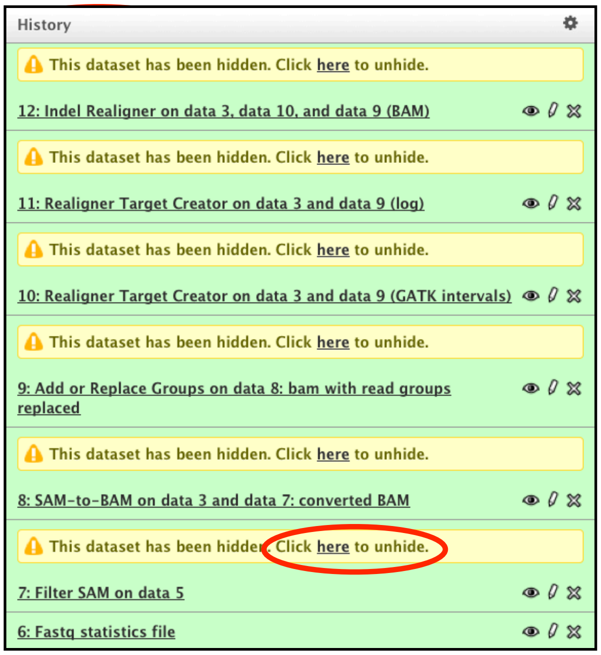
This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own local Galaxy instance or run Galaxy on the cloud.

36: Homozygous variants annotated (snpEff)
34: Uncovered regions annotated (snpEff)
31: Homozygous and heterozygous variants VCF (higher stringency, for downstream subtraction steps) (snpEff)
30: Uncovered regions (for downstream subtractions)
29: Homozygous variants VCF (mutant under consideration)
24: Depth of Coverage on data 3 and data 14 (output summary sample)
14: Alignment file (BAM)
6: Fastq statistics file
3: WS220.64_chr.fa
2: ot266_ProofOfPrinciple_Small.fastqsanger
1: CloudMap_TranscriptionFactors_wTF2.2.txt

19) You will notice that while over 30 output files were generated during the course of the workflow (output files are sequentially numbered), only some output files remain visible while others are hidden. The visible files are most important for analysis of the mutant under consideration or downstream analysis. In order to view hidden files, click ***Show Hidden Datasets*** in the History menu:


History
ot266
36: Homozygous variants annotated (snpEff)
34: Uncovered regions annotated (snpEff)
31: Homozygous and heterozygous variants VCF
stringency, for downstream subtraction steps) (s
30: Uncovered regions (for downstream subtract
29: Homozygous variants VCF (mutant under cor
24: Depth of Coverage on data 3 and data 14 (ou
sample)
14: Alignment file (BAM)
6: Fastq statistics file
3: WS220.64_chr.fa
2: ot266_ProofOfPrinciple_Small.fastqsanger
1: CloudMap_TranscriptionFactors_wTF2.2.txt
HISTORY LISTS
Saved Histories
Histories Shared with Me
CURRENT HISTORY
Create New
Clone
Copy Datasets
Share or Publish
Extract Workflow
Dataset Security
Show Deleted Datasets
Show Hidden Datasets
Purge Deleted Datasets
Show Structure
Export to File
Delete
Delete Permanently
OTHER ACTIONS
Import from File

20) You may unhide any files that are hidden:



This dataset has been hidden. Click here to unhide.

12: Indel Realigner on data 3, data 10, and data 9 (BAM)

This dataset has been hidden. Click here to unhide.

11: Realigner Target Creator on data 3 and data 9 (log)

This dataset has been hidden. Click here to unhide.

10: Realigner Target Creator on data 3 and data 9 (GATK intervals)

This dataset has been hidden. Click here to unhide.

9: Add or Replace Groups on data 8: bam with read groups replaced

This dataset has been hidden. Click here to unhide.

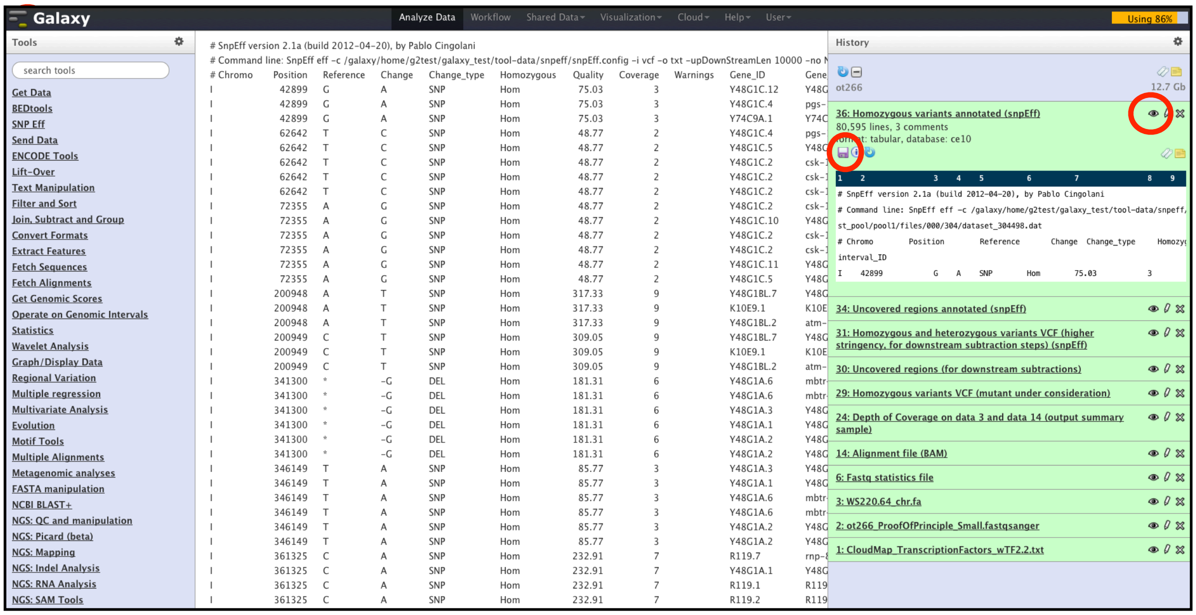
8: SAM-to-BAM on data 3 and data 7: converted BAM

This dataset has been hidden. Click here to unhide.

7: Filter SAM on data 5

6: Fastq statistics file

21) Click on a file to view more information on that file or to download the file:



```
# SnpEff version 2.1a (build 2012-04-20), by Pablo Cingolani
# Command line: SnpEff eff -c /galaxy/home/g2test/galaxy_test/tool-data/snpeff/snpEff.config -i vcf -o txt -upDownStreamLen 10000 -no N
# Chromo   Position  Reference  Change  Change_type  Homozygous  Quality  Coverage  Warnings  Gene_ID     Gene
I            42899   G          A       SNP          Hom            75.03        3            Y48G1C.12   Y48G
I            42899   G          A       SNP          Hom            75.03        3            Y48G1C.4    pgs-
I            42899   G          A       SNP          Hom            75.03        3            Y74C9A.1    Y74C
I            62642   T          C       SNP          Hom            48.77        2            Y48G1C.4    pgs-
I            62642   T          C       SNP          Hom            48.77        2            Y48G1C.5    Y48G
I            62642   T          C       SNP          Hom            48.77        2            Y48G1C.2    csk-1
I            62642   T          C       SNP          Hom            48.77        2            Y48G1C.2    csk-1
I            62642   T          C       SNP          Hom            48.77        2            Y48G1C.2    csk-1
I            72355   A          G       SNP          Hom            48.77        2            Y48G1C.2    csk-1
I            72355   A          G       SNP          Hom            48.77        2            Y48G1C.10   Y48G
I            72355   A          G       SNP          Hom            48.77        2            Y48G1C.2    csk-1
I            72355   A          G       SNP          Hom            48.77        2            Y48G1C.2    csk-1
I            72355   A          G       SNP          Hom            48.77        2            Y48G1C.11   Y48G
I            72355   A          G       SNP          Hom            48.77        2            Y48G1C.5    Y48G
I           200948   A          T       SNP          Hom           317.33        9            Y48G1BL.7   Y48G
I           200948   A          T       SNP          Hom           317.33        9            K10E9.1     K10E
I           200948   A          T       SNP          Hom           317.33        9            Y48G1BL.2   atm-
I           200949   C          T       SNP          Hom           309.05        9            Y48G1BL.7   Y48G
I           200949   C          T       SNP          Hom           309.05        9            K10E9.1     K10E
I           200949   C          T       SNP          Hom           309.05        9            Y48G1BL.2   atm-
I           341300   *          -G      DEL          Hom           181.31        6            Y48G1A.6    mbtr-
I           341300   *          -G      DEL          Hom           181.31        6            Y48G1A.6    mbtr-
I           341300   *          -G      DEL          Hom           181.31        6            Y48G1A.3    Y48G
I           341300   *          -G      DEL          Hom           181.31        6            Y48G1A.1    Y48G
I           341300   *          -G      DEL          Hom           181.31        6            Y48G1A.2    Y48G
I           341300   *          -G      DEL          Hom           181.31        6            Y48G1A.2    Y48G
I           346149   T          A       SNP          Hom            85.77        3            Y48G1A.3    Y48G
I           346149   T          A       SNP          Hom            85.77        3            Y48G1A.1    Y48G
I           346149   T          A       SNP          Hom            85.77        3            Y48G1A.6    mbtr-
I           346149   T          A       SNP          Hom            85.77        3            Y48G1A.6    mbtr-
I           346149   T          A       SNP          Hom            85.77        3            Y48G1A.2    Y48G
I           346149   T          A       SNP          Hom            85.77        3            Y48G1A.2    Y48G
I           361325   C          A       SNP          Hom           232.91        7            R119.7      rnp-
I           361325   C          A       SNP          Hom           232.91        7            Y48G1A.1    Y48G
I           361325   C          A       SNP          Hom           232.91        7            R119.1      R119
I           361325   C          A       SNP          Hom           232.91        7            R119.2      R119
```

ot266 12.7 Gb

36: Homozygous variants annotated (snpEff)
80,595 lines, 3 comments
format: tabular, database: ce10

```
1    2          3    4    5        6          7              8    9
# SnpEff version 2.1a (build 2012-04-20), by Pablo Cingolani
# Command line: SnpEff eff -c /galaxy/home/g2test/galaxy_test/tool-data/snpeff/
st_pool/pool1/files/000/304/dataset_304498.dat
# Chromo       Position       Reference      Change  Change_type     Homozyg
interval_ID
I    42899          G    A    SNP       Hom       75.03          3
```

34: Uncovered regions annotated (snpEff)

31: Homozygous and heterozygous variants VCF (higher stringency, for downstream subtraction steps) (snpEff)

30: Uncovered regions (for downstream subtractions)

29: Homozygous variants VCF (mutant under consideration)

24: Depth of Coverage on data 3 and data 14 (output summary sample)

14: Alignment file (BAM)

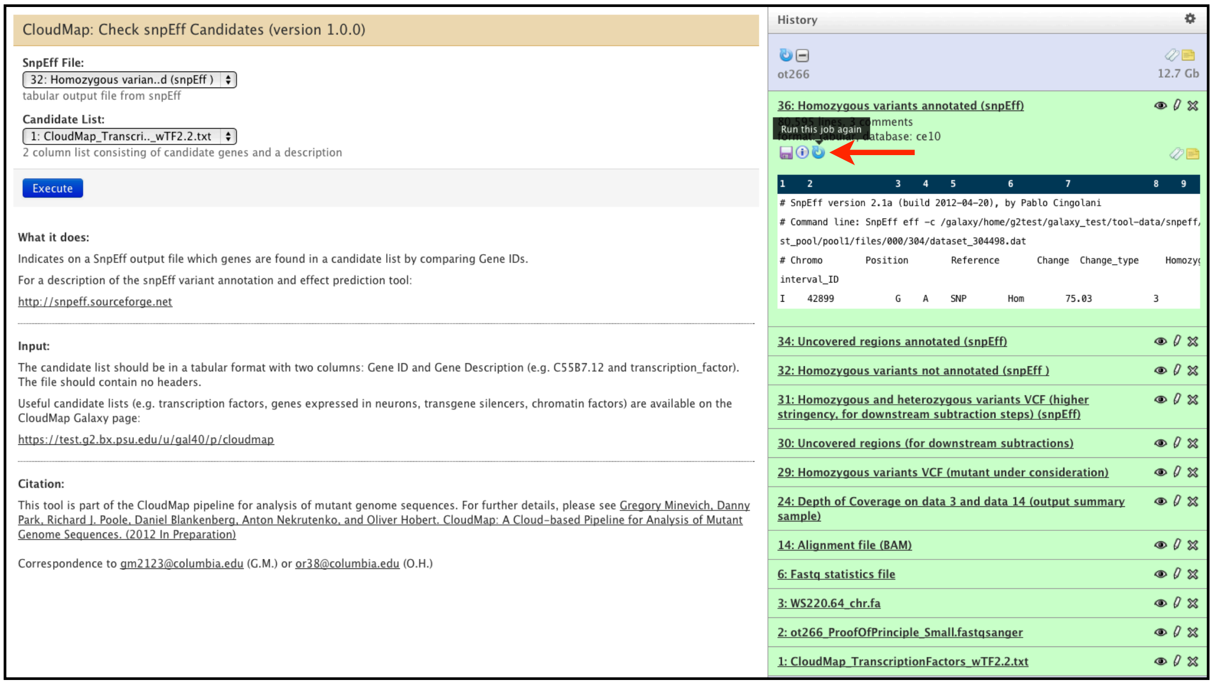
6: Fastq statistics file

3: WS220.64_chr.fa

2: ot266_ProofOfPrinciple_Small.fastqsanger

1: CloudMap_TranscriptionFactors_wTF2.2.txt

22) If you want to rerun a tool with different parameters, click the ***run this job again*** arrow. To rerun a tool on a hidden dataset, make sure to unhide the hidden dataset first. If a tool fails (it will turn red) for no apparent reason when it has previously worked successfully, try running it again before submitting a bug report to Galaxy.



CloudMap: Check snpEff Candidates (version 1.0.0)

SnpEff File:
32: Homozygous varian..d (snpEff)
tabular output file from snpEff

Candidate List:
1: CloudMap_Transcri..._wTF2.2.txt
2 column list consisting of candidate genes and a description

What it does:

Indicates on a SnpEff output file which genes are found in a candidate list by comparing Gene IDs.

For a description of the snpEff variant annotation and effect prediction tool:

http://snpeff.sourceforge.net

Input:

The candidate list should be in a tabular format with two columns: Gene ID and Gene Description (e.g. C55B7.12 and transcription_factor). The file should contain no headers.

Useful candidate lists (e.g. transcription factors, genes expressed in neurons, transgene silencers, chromatin factors) are available on the CloudMap Galaxy page:

https://test.g2.bx.psu.edu/u/gal40/p/cloudmap

Citation:

This tool is part of the CloudMap pipeline for analysis of mutant genome sequences. For further details, please see Gregory Minevich, Danny Park, Richard J. Poole, Daniel Blankenberg, Anton Nekrutenko, and Oliver Hobert. CloudMap: A Cloud-based Pipeline for Analysis of Mutant Genome Sequences. (2012 In Preparation)

Correspondence to gm2123@columbia.edu (G.M.) or or38@columbia.edu (O.H.)

ot266
12.7 Gb

36: Homozygous variants annotated (snpEff)
80,595 lines, 3 comments
format: tabular, database: ce10

```
1    2               3    4    5         6         7              8    9
# SnpEff version 2.1a (build 2012-04-20), by Pablo Cingolani
# Command line: SnpEff eff -c /galaxy/home/g2test/galaxy_test/tool-data/snpeff/
st_pool/pool1/files/000/304/dataset_304498.dat
# Chromo        Position        Reference       Change  Change_type     Homozyg
interval_ID
I    42899           G    A    SNP       Hom       75.03          3
```

34: Uncovered regions annotated (snpEff)
32: Homozygous variants not annotated (snpEff)
31: Homozygous and heterozygous variants VCF (higher stringency, for downstream subtraction steps) (snpEff)
30: Uncovered regions (for downstream subtractions)
29: Homozygous variants VCF (mutant under consideration)
24: Depth of Coverage on data 3 and data 14 (output summary sample)
14: Alignment file (BAM)
6: Fastq statistics file
3: WS220.64_chr.fa
2: ot266_ProofOfPrinciple_Small.fastqsanger
1: CloudMap_TranscriptionFactors_wTF2.2.txt

23) Several ***sample metric*** files are created as part of the workflow (more details on following pages):

1. A ***FASTQ quality statistics*** file summarizes the quality of all reads before they are aligned to the reference genome (*Galaxy's FASTQ manipulation tools*).

2. A ***Depth of Coverage*** file gives a summary of overall read depth in the BAM alignment file (*GATK*).

3. A ***graphical summary of all the variants*** in the sample (*snpEff)*. This file must be downloaded to be viewed properly. It will not appear correctly if viewed within Galaxy using the "peek" (eye) icon. (For more information on file format, see: http://snpeff.sourceforge.net/)

24) A ***primary set of files for analysis*** are created as part of the workflow:

1. An ***annotated set of homozygous variants*** in the entire sample (*snpEff*). (For more information on file format, see: http://snpeff.sourceforge.net/)

2. A ***BAM alignment file*** that can be viewed in your choice of alignment viewers (*SAMtools*). (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat)

3. A list of ***annotated uncovered regions*** (BED file) that may be putative deletions (*BEDtools* & *snpEff)*. (For more information on file format, see: http://snpeff.sourceforge.net/)

25) Additional files that can be used for ***downstream subtraction workflows*** are generated (for more details see the ***Subtract Variants*** and ***Uncovered Region Subtraction*** workflows):

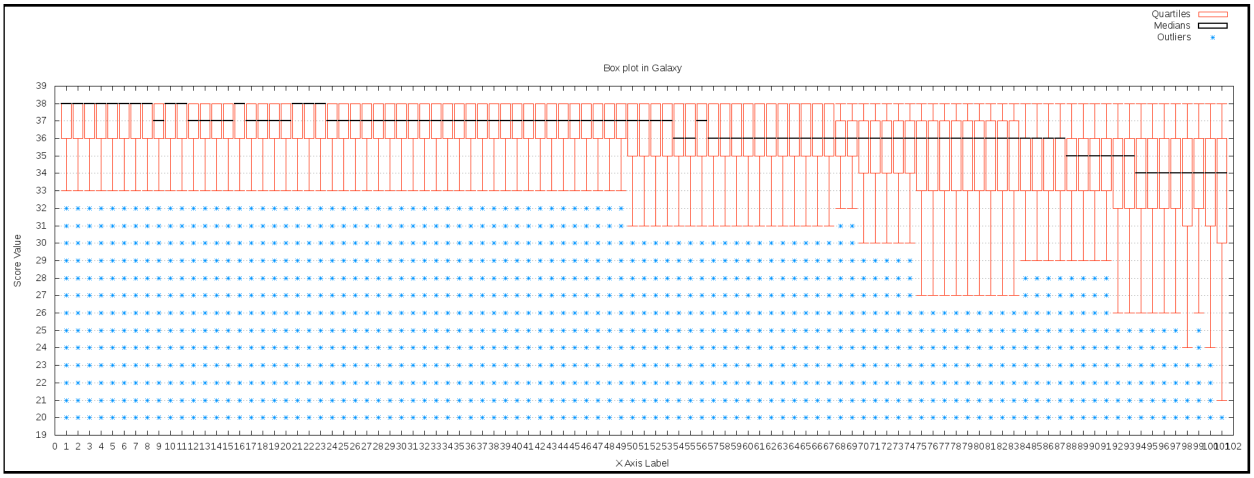
1. A ***set of homozygous variants*** (VCF file) in the entire sample that can be further filtered by subtracting variants present in other samples using the ***CloudMap Subtract Variants*** workflow (*GATK*). This VCF file is used as input into snpEff to generate the ***annotated list of homozygous variants*** mentioned in the section above. It has Hawaiian unfiltered variants subtracted and includes variants that pass a low quality filtering threshold. This file should be downloaded to be easily viewed in its entirety. The first several lines in any VCF file are header lines starting with "#" so users who wish to filter or sort these files in Excel are advised to remove the header lines. (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat.html)

2. A ***set of homozygous and heterozygous variants*** (VCF file) in the entire sample (run at higher quality stringency) that can be used as a set of variants to subtract from other samples (GATK). It has Hawaiian unfiltered variants subtracted and includes variants that pass a higher quality filtering threshold (read mapping quality ≥ 30 and coverage ≥ 3). In an effort to subtract as many variants as possible, users may subtract not only homozygous variants from other strains, but also heterozygous variants. Such a strategy assumes that phenotype-inducing homozygous mutant variants in the strain under analysis are unlikely to be heterozygous in strains that will be used for subtraction. It is especially important to apply this strategy when subtracting variant lists generated using the *Hawaiian Variant Mapping with WGS Data* approach (see section "**CloudMap *Hawaiian Variant Mapping with WGS Data*** tool"), since background variants will be present in a heterozygous state in these pooled samples as a consequence of the mapping cross. (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat.html)

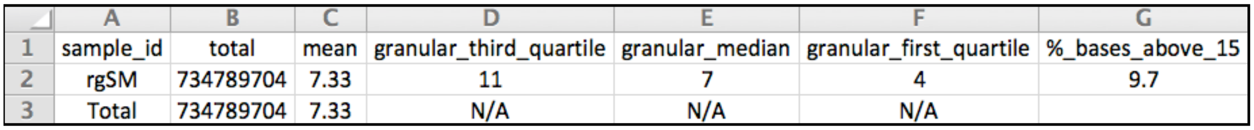
3. A set of ***uncovered regions*** (BED file) used to generate the annotated uncovered regions mentioned in the section above. This list of uncovered regions can be used in two ways. It can be further filtered by subtracting uncovered regions present in other samples using the ***CloudMap Uncovered Region Subtraction*** workflow to find uncovered regions unique to the sample under analysis. The resultant file can then be annotated using snpEff. Alternatively, these uncovered regions can be used to subtract from the set of uncovered regions in other samples (using *BEDtools*). (for more details see the ***Subtract Variants*** and ***Uncovered Region Subtraction*** workflows) (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat.html)

Examples of ***sample metric*** files (mentioned in section 22 above):

23.1) ***FASTQ quality statistics*** file (*Galaxy's FASTQ manipulation tools*)

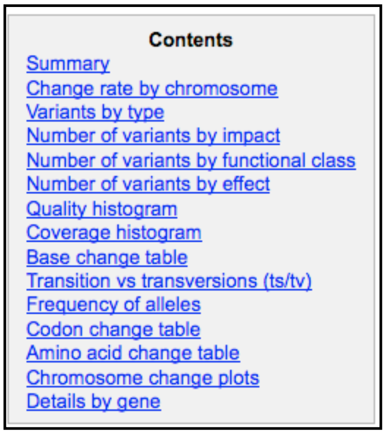


23.2) ***Depth of Coverage*** file (*GATK)*



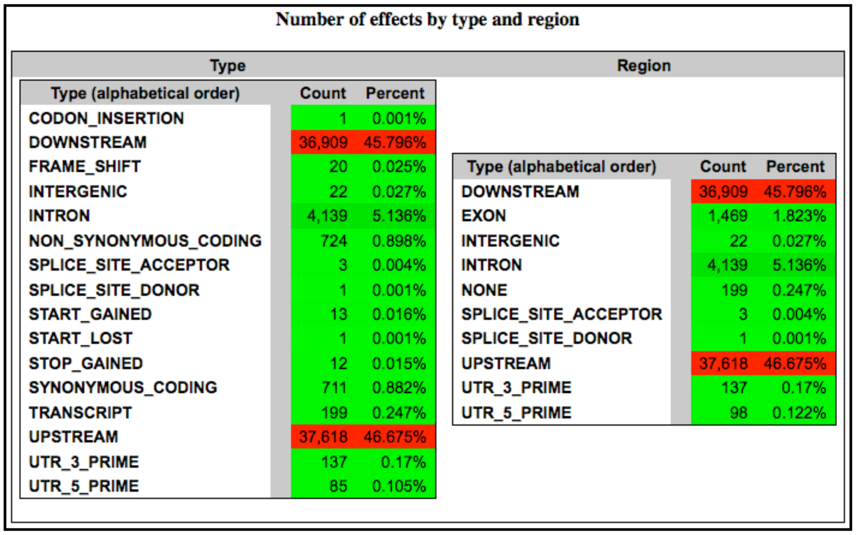
	A	B	C	D	E	F	G
1	sample_id	total	mean	granular_third_quartile	granular_median	granular_first_quartile	%_bases_above_15
2	rgSM	734789704	7.33	11	7	4	9.7
3	Total	734789704	7.33	N/A	N/A	N/A	

23.3) ***Graphical summary of all the variants*** in the sample (html file from *snpEff)*. Note: this file is very comprehensive and only excerpts of it are shown here:



Contents

Summary
Change rate by chromosome
Variants by type
Number of variants by impact
Number of variants by functional class
Number of variants by effect
Quality histogram
Coverage histogram
Base change table
Transition vs transversions (ts/tv)
Frequency of alleles
Codon change table
Amino acid change table
Chromosome change plots
Details by gene



Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
CODON_INSERTION	1	0.001%
DOWNSTREAM	36,909	45.796%
FRAME_SHIFT	20	0.025%
INTERGENIC	22	0.027%
INTRON	4,139	5.136%
NON_SYNONYMOUS_CODING	724	0.898%
SPLICE_SITE_ACCEPTOR	3	0.004%
SPLICE_SITE_DONOR	1	0.001%
START_GAINED	13	0.016%
START_LOST	1	0.001%
STOP_GAINED	12	0.015%
SYNONYMOUS_CODING	711	0.882%
TRANSCRIPT	199	0.247%
UPSTREAM	37,618	46.675%
UTR_3_PRIME	137	0.17%
UTR_5_PRIME	85	0.105%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	36,909	45.796%
EXON	1,469	1.823%
INTERGENIC	22	0.027%
INTRON	4,139	5.136%
NONE	199	0.247%
SPLICE_SITE_ACCEPTOR	3	0.004%
SPLICE_SITE_DONOR	1	0.001%
UPSTREAM	37,618	46.675%
UTR_3_PRIME	137	0.17%
UTR_5_PRIME	98	0.122%

Examples of ***primary set of files for analysis*** (mentioned in step 23 above):

24.1) ***Annotated set of homozygous variants*** (Fig.4) (*snpEff*)

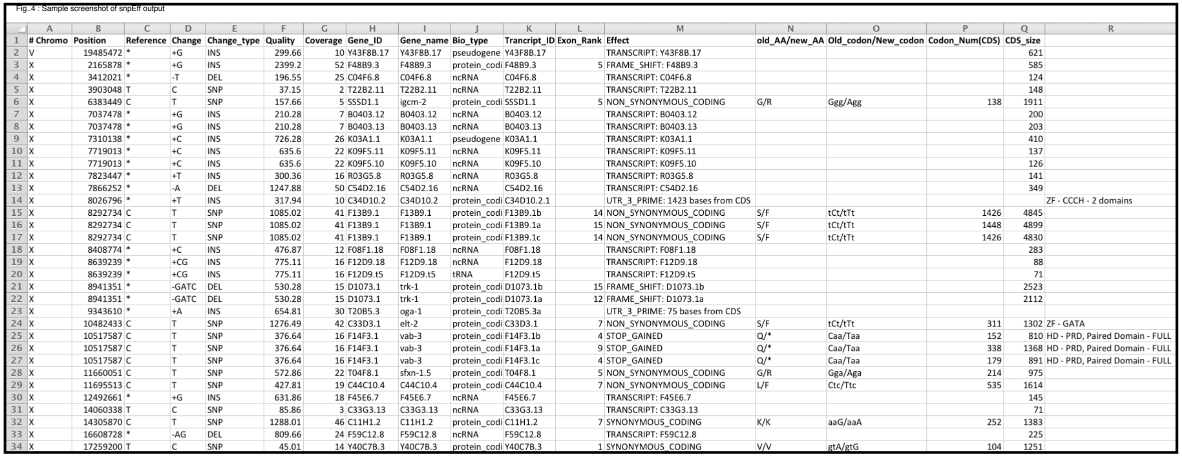
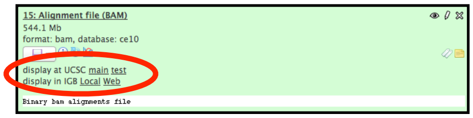


Fig..4 : Sample screenshot of snpEff output

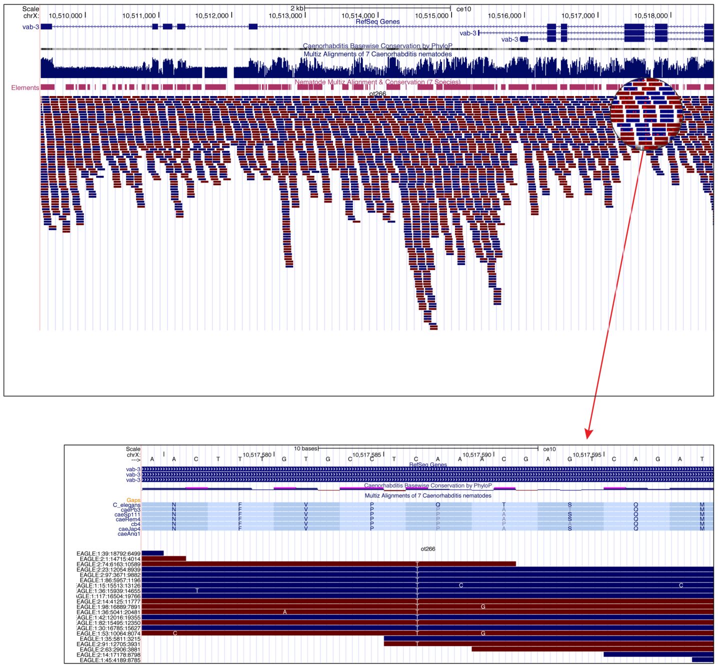
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	# Chromo	Position	Reference	Change	Change_type	Quality	Coverage	Gene_ID	Gene_name	Bio_type	Trancript_ID	Exon_Rank	Effect	old_AA/new_AA	Old_codon/New_codon	Codon_Num(CDS)	CDS_size	
2	V	19485472	*	+G	INS	299.66	10	Y43F8B.17	Y43F8B.17	pseudogene	Y43F8B.17		TRANSCRIPT: Y43F8B.17				621	
3	X	2165878	*	+G	INS	2399.2	52	F48B9.3	F48B9.3	protein_codi	F48B9.3	5	FRAME_SHIFT: F48B9.3				585	
4	X	3412021	*	-T	DEL	196.55	25	C04F6.8	C04F6.8	ncRNA	C04F6.8		TRANSCRIPT: C04F6.8				124	
5	X	3903048	T	C	SNP	37.15	2	T22B2.11	T22B2.11	ncRNA	T22B2.11		TRANSCRIPT: T22B2.11				148	
6	X	6383449	C	T	SNP	157.66	5	SSSD1.1	igcm-2	protein_codi	SSSD1.1	5	NON_SYNONYMOUS_CODING	G/R	Ggg/Agg	138	1911	
7	X	7037478	*	+G	INS	210.28	7	B0403.12	B0403.12	ncRNA	B0403.12		TRANSCRIPT: B0403.12				200	
8	X	7037478	*	+G	INS	210.28	7	B0403.13	B0403.13	ncRNA	B0403.13		TRANSCRIPT: B0403.13				203	
9	X	7310138	*	+C	INS	726.28	26	K03A1.1	K03A1.1	pseudogene	K03A1.1		TRANSCRIPT: K03A1.1				410	
10	X	7719013	*	+C	INS	635.6	22	K09F5.11	K09F5.11	ncRNA	K09F5.11		TRANSCRIPT: K09F5.11				137	
11	X	7719013	*	+C	INS	635.6	22	K09F5.10	K09F5.10	ncRNA	K09F5.10		TRANSCRIPT: K09F5.10				126	
12	X	7823447	*	+T	INS	300.36	16	R03G5.8	R03G5.8	ncRNA	R03G5.8		TRANSCRIPT: R03G5.8				141	
13	X	7866252	*	-A	DEL	1247.88	50	C54D2.16	C54D2.16	ncRNA	C54D2.16		TRANSCRIPT: C54D2.16				349	
14	X	8026796	*	+T	INS	317.94	10	C34D10.2	C34D10.2	protein_codi	C34D10.2.1		UTR_3_PRIME: 1423 bases from CDS					ZF - CCCH - 2 domains
15	X	8292734	C	T	SNP	1085.02	41	F13B9.1	F13B9.1	protein_codi	F13B9.1b	14	NON_SYNONYMOUS_CODING	S/F	tCt/tTt	1426	4845	
16	X	8292734	C	T	SNP	1085.02	41	F13B9.1	F13B9.1	protein_codi	F13B9.1a	15	NON_SYNONYMOUS_CODING	S/F	tCt/tTt	1448	4899	
17	X	8292734	C	T	SNP	1085.02	41	F13B9.1	F13B9.1	protein_codi	F13B9.1c	14	NON_SYNONYMOUS_CODING	S/F	tCt/tTt	1426	4830	
18	X	8408774	*	+C	INS	476.87	12	F08F1.18	F08F1.18	ncRNA	F08F1.18		TRANSCRIPT: F08F1.18				283	
19	X	8639239	*	+CG	INS	775.11	16	F12D9.18	F12D9.18	ncRNA	F12D9.18		TRANSCRIPT: F12D9.18				88	
20	X	8639239	*	+CG	INS	775.11	16	F12D9.t5	F12D9.t5	tRNA	F12D9.t5		TRANSCRIPT: F12D9.t5				71	
21	X	8941351	*	-GATC	DEL	530.28	15	D1073.1	trk-1	protein_codi	D1073.1b	15	FRAME_SHIFT: D1073.1b				2523	
22	X	8941351	*	-GATC	DEL	530.28	15	D1073.1	trk-1	protein_codi	D1073.1a	12	FRAME_SHIFT: D1073.1a				2112	
23	X	9343610	*	+A	INS	654.81	30	T20B5.3	oga-1	protein_codi	T20B5.3a		UTR_3_PRIME: 75 bases from CDS					
24	X	10482433	C	T	SNP	1276.49	42	C33D3.1	elt-2	protein_codi	C33D3.1	7	NON_SYNONYMOUS_CODING	S/F	tCt/tTt	311	1302	ZF - GATA
25	X	10517587	C	T	SNP	376.64	16	F14F3.1	vab-3	protein_codi	F14F3.1b	4	STOP_GAINED	Q/*	Caa/Taa	152	810	HD - PRD, Paired Domain - FULL
26	X	10517587	C	T	SNP	376.64	16	F14F3.1	vab-3	protein_codi	F14F3.1a	9	STOP_GAINED	Q/*	Caa/Taa	338	1368	HD - PRD, Paired Domain - FULL
27	X	10517587	C	T	SNP	376.64	16	F14F3.1	vab-3	protein_codi	F14F3.1c	4	STOP_GAINED	Q/*	Caa/Taa	179	891	HD - PRD, Paired Domain - FULL
28	X	11660051	C	T	SNP	572.86	22	T04F8.1	sfxn-1.5	protein_codi	T04F8.1	5	NON_SYNONYMOUS_CODING	G/R	Gga/Aga	214	975	
29	X	11695513	C	T	SNP	427.81	19	C44C10.4	C44C10.4	protein_codi	C44C10.4	7	NON_SYNONYMOUS_CODING	L/F	Ctc/Ttc	535	1614	
30	X	12492661	*	+G	INS	631.86	18	F45E6.7	F45E6.7	ncRNA	F45E6.7		TRANSCRIPT: F45E6.7				145	
31	X	14060338	T	C	SNP	85.86	3	C33G3.13	C33G3.13	ncRNA	C33G3.13		TRANSCRIPT: C33G3.13				71	
32	X	14305870	C	T	SNP	1288.01	46	C11H1.2	C11H1.2	protein_codi	C11H1.2	7	SYNONYMOUS_CODING	K/K	aaG/aaA	252	1383	
33	X	16608728	*	-AG	DEL	809.66	24	F59C12.8	F59C12.8	ncRNA	F59C12.8		TRANSCRIPT: F59C12.8				225	
34	X	17259200	T	C	SNP	45.01	14	Y40C7B.3	Y40C7B.3	protein_codi	Y40C7B.3	1	SYNONYMOUS_CODING	V/V	gtA/gtG	104	1251	

24.2) ***BAM alignment*** file (*SAMtools*) (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat.html). Click on the "***display in***" link in your history or download the BAM file to view it in your alignment viewer of choice:



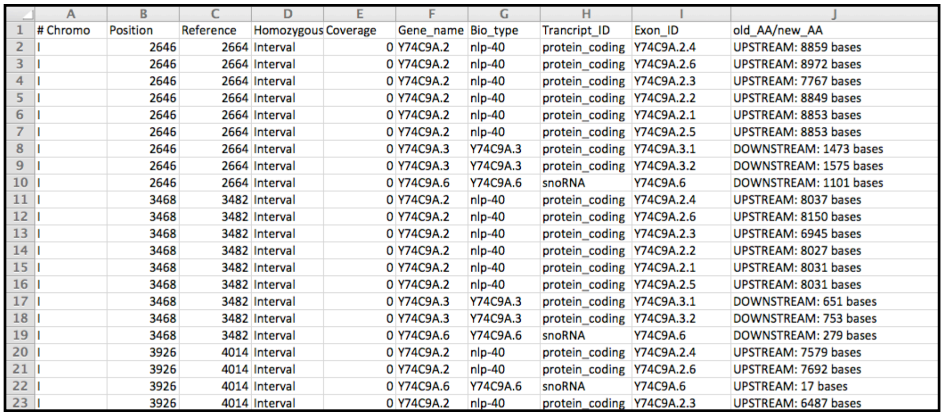
15: Alignment file (BAM)
544.1 Mb
format: bam, database: ce10
display at UCSC main test
display in IGB Local Web
Binary bam alignments file

(e.g. Fig.9 UCSC Genome Browser)



Note: Information displayed in alignment viewers often will not exactly match that in variant files (VCFs) or lists of annotated variants (snpEff). This is because read mapping qualities and base qualities are incorporated into which variants are ultimately called. Most alignment viewers have filter settings that can be used to only display reads with mapping quality scores above a certain value. Applying these filters should result in alignments that more closely approximate variant lists.

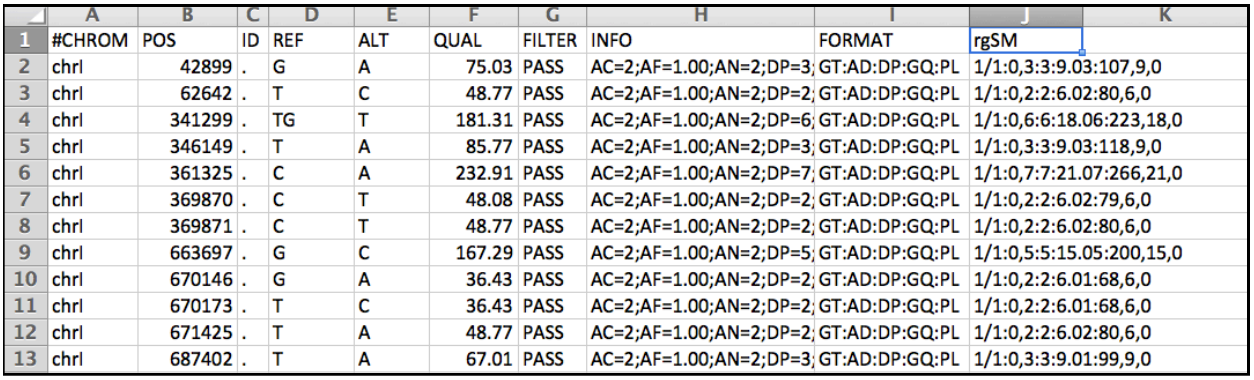
24.3) A list of ***annotated uncovered regions*** (BED file) (*BEDtools* & *snpEff)* (For more information on file format, see: http://snpeff.sourceforge.net/)



	A	B	C	D	E	F	G	H	I	J
1	# Chromo	Position	Reference	Homozygous	Coverage	Gene_name	Bio_type	Trancript_ID	Exon_ID	old_AA/new_AA
2	I	2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.4	UPSTREAM: 8859 bases
3	I	2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.6	UPSTREAM: 8972 bases
4	I	2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.3	UPSTREAM: 7767 bases
5	I	2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.2	UPSTREAM: 8849 bases
6	I	2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.1	UPSTREAM: 8853 bases
7	I	2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.5	UPSTREAM: 8853 bases
8	I	2646	2664	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.1	DOWNSTREAM: 1473 bases
9	I	2646	2664	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.2	DOWNSTREAM: 1575 bases
10	I	2646	2664	Interval	0	Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6	DOWNSTREAM: 1101 bases
11	I	3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.4	UPSTREAM: 8037 bases
12	I	3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.6	UPSTREAM: 8150 bases
13	I	3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.3	UPSTREAM: 6945 bases
14	I	3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.2	UPSTREAM: 8027 bases
15	I	3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.1	UPSTREAM: 8031 bases
16	I	3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.5	UPSTREAM: 8031 bases
17	I	3468	3482	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.1	DOWNSTREAM: 651 bases
18	I	3468	3482	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.2	DOWNSTREAM: 753 bases
19	I	3468	3482	Interval	0	Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6	DOWNSTREAM: 279 bases
20	I	3926	4014	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.4	UPSTREAM: 7579 bases
21	I	3926	4014	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.6	UPSTREAM: 7692 bases
22	I	3926	4014	Interval	0	Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6	UPSTREAM: 17 bases
23	I	3926	4014	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.3	UPSTREAM: 6487 bases

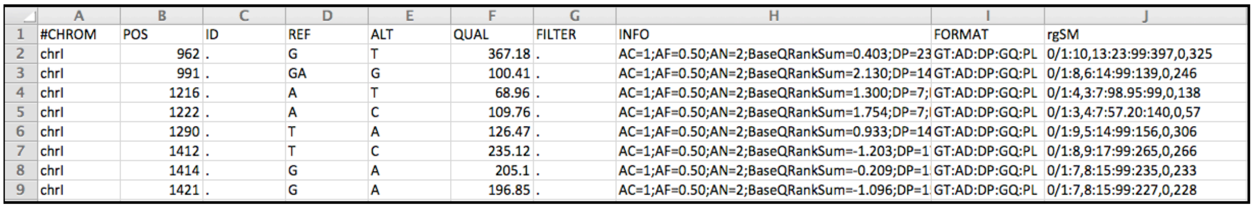
Additional files that can be used for ***downstream subtraction workflows*** (mentioned in step 25 above):

25.1) ***Set of homozygous variants*** (VCF file generated by *GATK*). Header lines starting with "#" have been removed in Excel. (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat)



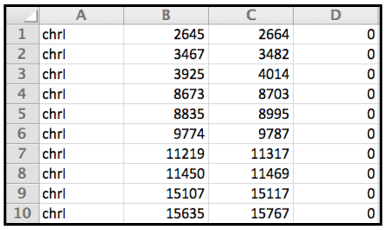
	A	B	C	D	E	F	G	H	I	J	K
1	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	rgSM	
2	chrI	42899	.	G	A	75.03	PASS	AC=2;AF=1.00;AN=2;DP=3;	GT:AD:DP:GQ:PL	1/1:0,3:3:9.03:107,9,0	
3	chrI	62642	.	T	C	48.77	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:80,6,0	
4	chrI	341299	.	TG	T	181.31	PASS	AC=2;AF=1.00;AN=2;DP=6;	GT:AD:DP:GQ:PL	1/1:0,6:6:18.06:223,18,0	
5	chrI	346149	.	T	A	85.77	PASS	AC=2;AF=1.00;AN=2;DP=3;	GT:AD:DP:GQ:PL	1/1:0,3:3:9.03:118,9,0	
6	chrI	361325	.	C	A	232.91	PASS	AC=2;AF=1.00;AN=2;DP=7;	GT:AD:DP:GQ:PL	1/1:0,7:7:21.07:266,21,0	
7	chrI	369870	.	C	T	48.08	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:79,6,0	
8	chrI	369871	.	C	T	48.77	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:80,6,0	
9	chrI	663697	.	G	C	167.29	PASS	AC=2;AF=1.00;AN=2;DP=5;	GT:AD:DP:GQ:PL	1/1:0,5:5:15.05:200,15,0	
10	chrI	670146	.	G	A	36.43	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.01:68,6,0	
11	chrI	670173	.	T	C	36.43	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.01:68,6,0	
12	chrI	671425	.	T	A	48.77	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:80,6,0	
13	chrI	687402	.	T	A	67.01	PASS	AC=2;AF=1.00;AN=2;DP=3;	GT:AD:DP:GQ:PL	1/1:0,3:3:9.01:99,9,0	

25.2) ***Set of homozygous and heterozygous variants*** (VCF file generated by *GATK*). Header lines starting with "#" have been removed in Excel. (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat)



	A	B	C	D	E	F	G	H	I	J
1	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	rgSM
2	chrI	962	.	G	T	367.18	.	AC=1;AF=0.50;AN=2;BaseQRankSum=0.403;DP=23	GT:AD:DP:GQ:PL	0/1:10,13:23:99:397,0,325
3	chrI	991	.	GA	G	100.41	.	AC=1;AF=0.50;AN=2;BaseQRankSum=2.130;DP=14	GT:AD:DP:GQ:PL	0/1:8,6:14:99:139,0,246
4	chrI	1216	.	A	T	68.96	.	AC=1;AF=0.50;AN=2;BaseQRankSum=1.300;DP=7;	GT:AD:DP:GQ:PL	0/1:4,3:7:98.95:99,0,138
5	chrI	1222	.	A	C	109.76	.	AC=1;AF=0.50;AN=2;BaseQRankSum=1.754;DP=7;	GT:AD:DP:GQ:PL	0/1:3,4:7:57.20:140,0,57
6	chrI	1290	.	T	A	126.47	.	AC=1;AF=0.50;AN=2;BaseQRankSum=0.933;DP=14	GT:AD:DP:GQ:PL	0/1:9,5:14:99:156,0,306
7	chrI	1412	.	T	C	235.12	.	AC=1;AF=0.50;AN=2;BaseQRankSum=-1.203;DP=1	GT:AD:DP:GQ:PL	0/1:8,9:17:99:265,0,266
8	chrI	1414	.	G	A	205.1	.	AC=1;AF=0.50;AN=2;BaseQRankSum=-0.209;DP=1	GT:AD:DP:GQ:PL	0/1:7,8:15:99:235,0,233
9	chrI	1421	.	G	A	196.85	.	AC=1;AF=0.50;AN=2;BaseQRankSum=-1.096;DP=1	GT:AD:DP:GQ:PL	0/1:7,8:15:99:227,0,228

25.3) ***Set of uncovered regions (**BED file**)*** (*BEDtools*). (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat)



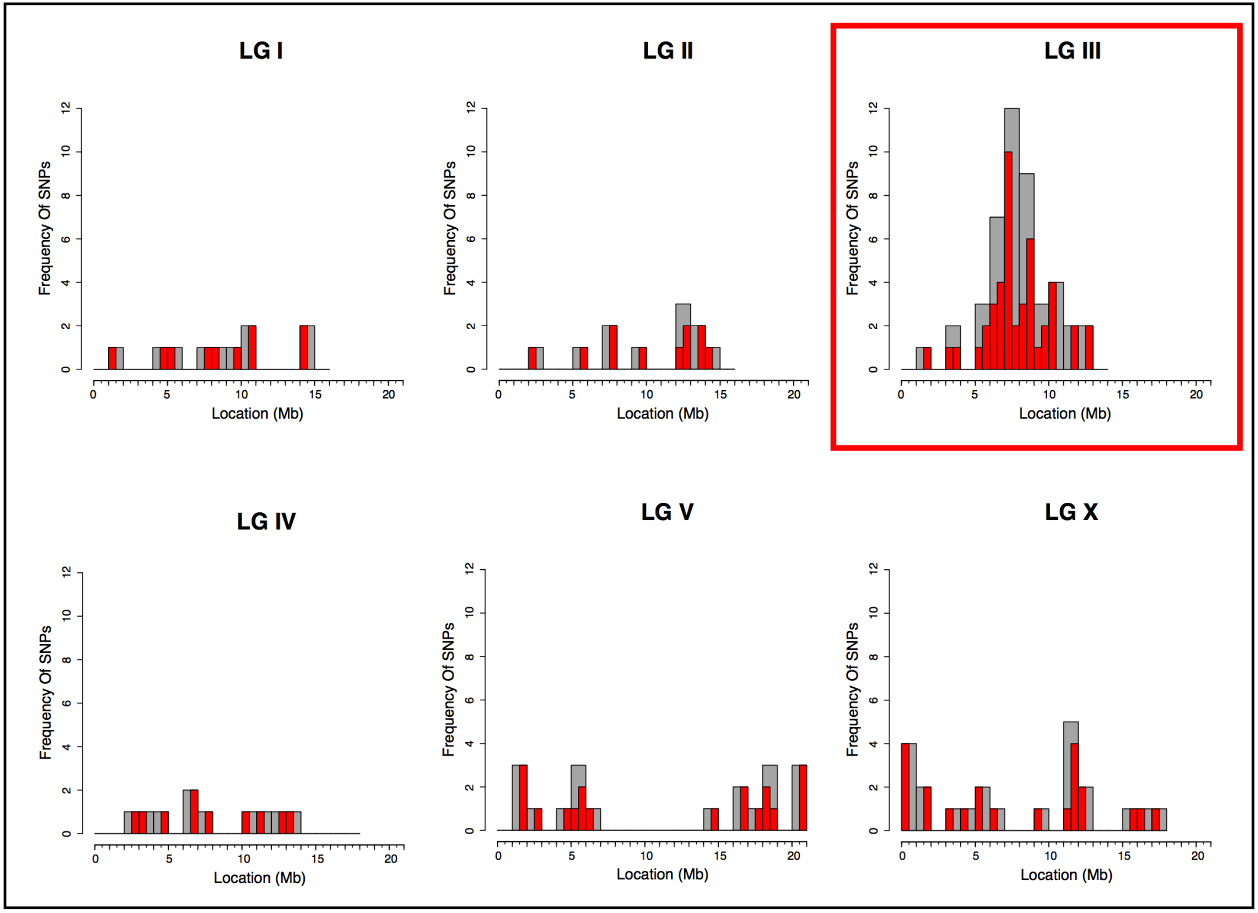
	A	B	C	D
1	chrI	2645	2664	0
2	chrI	3467	3482	0
3	chrI	3925	4014	0
4	chrI	8673	8703	0
5	chrI	8835	8995	0
6	chrI	9774	9787	0
7	chrI	11219	11317	0
8	chrI	11450	11469	0
9	chrI	15107	15117	0
10	chrI	15635	15767	0

Note: We strongly suggest that users employ the ***Subtract Variants*** and ***Uncovered Region Subtraction*** workflows if additional strains are available for this purpose. The general concept is shown in **Fig.5** of the CloudMap paper.

CloudMap EMS Variant Density Mapping Workflow

The ***EMS Variant Density Mapping*** workflow consists of the ***Unmapped Mutant*** workflow followed by the ***Subtract Variants*** workflow. The final VCF output is then plotted using the CloudMap ***EMS Variant Density Mapping*** tool. Readers are directed to the sections of this user guide that describe these workflows.

Fig.S3 from the CloudMap paper:



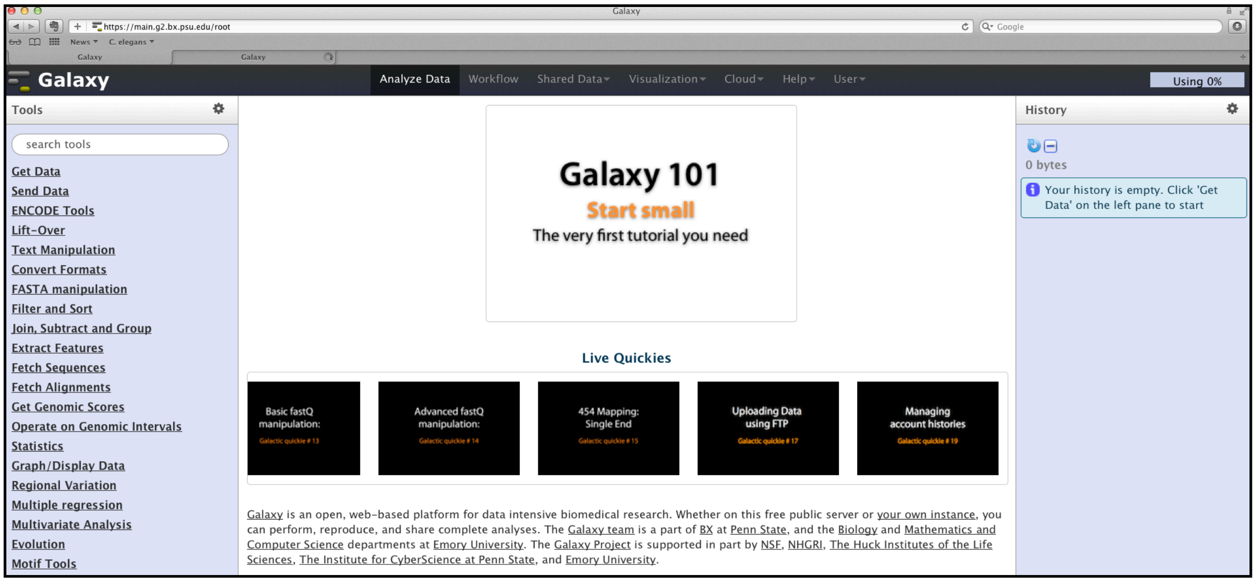
CloudMap | Cloud-based Pipeline for Analysis of Mutant Genome Sequences

CloudMap Subtract Variants workflow (using *ot266* Proof of Principle example from the CloudMap paper). A video version of this user guide is available at: http://usegalaxy.org/cloudmap.

This workflow should be used downstream of either of the following workflows: ***Hawaiian Variant Mapping with WGS data and Variant Calling , EMS Density Mapping, or Unmapped Mutant workflows.*** Here we demonstrate the workflow using the *ot266* example from the Cloudmap paper (**Fig.8**). Users may apply this workflow to their own data by substituting the datasets in this example with their own datasets.

These workflows provide default function parameters, ensuring that users follow best practices, and allow for automated execution of sequential operations. We provide these workflows as helpful guides, but experienced users may execute functions in any meaningful order they please and may also create and share their own workflows to take advantage of the automation feature. More CloudMap documentation is available at http://usegalaxy.org/cloudmap.

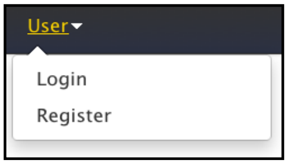
1) Navigate to http://usegalaxy.org



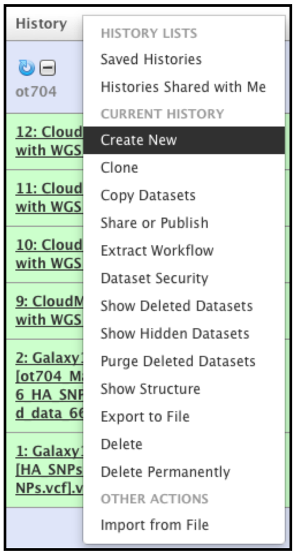
Your history is empty. Click 'Get Data' on the left pane to start

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

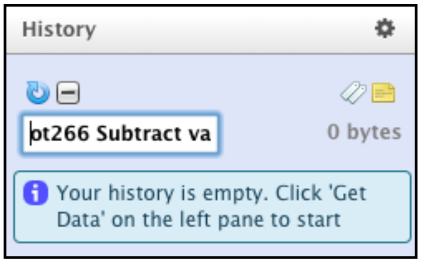
2) You should already have a Galaxy account at this point because you have run earlier workflows:



3) Once you are logged in using your email address, create a new history:



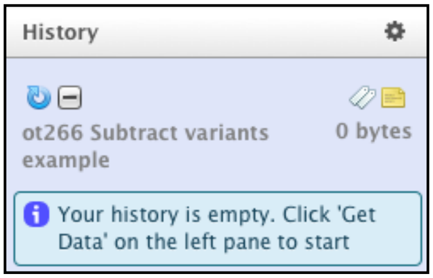
4) Now name that history “ot266 Subtract variants example”:



ot266 Subtract va

Your history is empty. Click 'Get Data' on the left pane to start

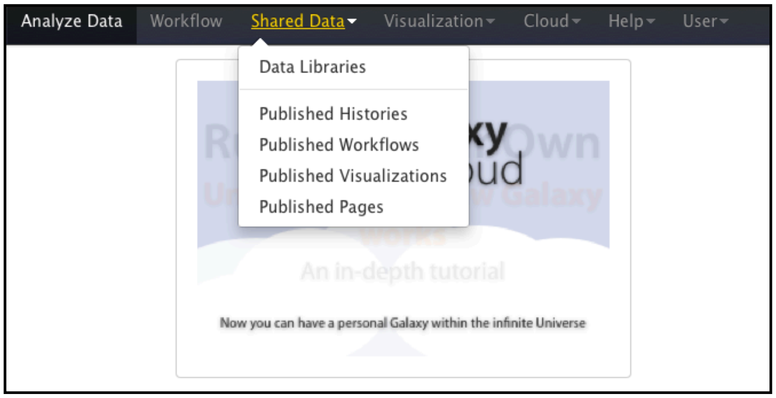
5) You now need to import the ***ot266 Proof of principle*** files or your own files to run the workflow:



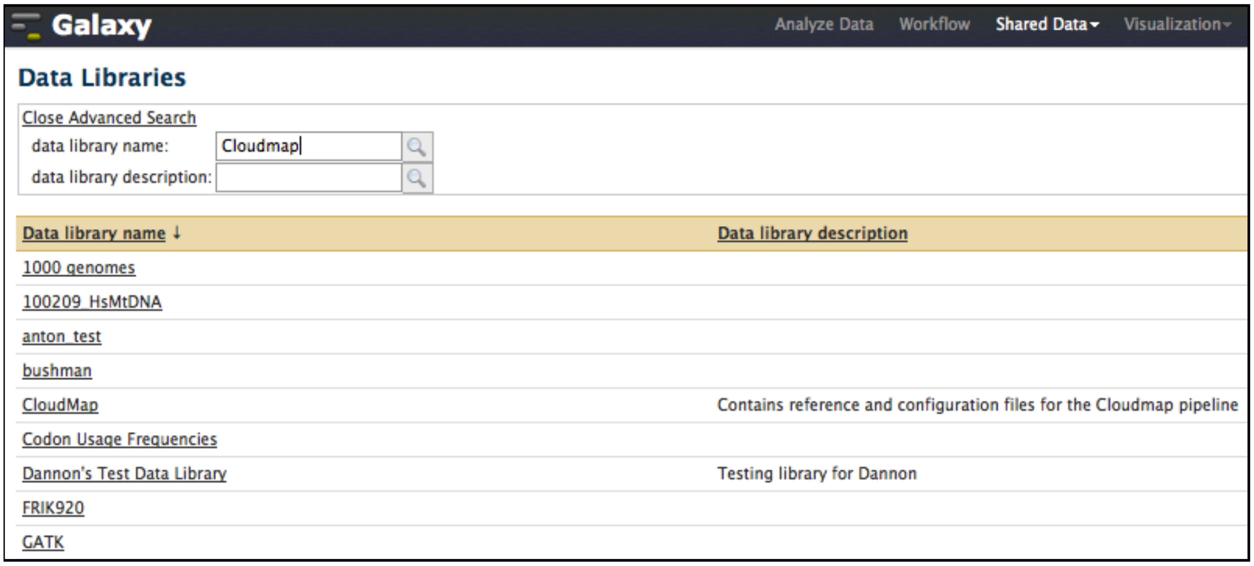
ot266 Subtract variants example

Your history is empty. Click 'Get Data' on the left pane to start

6) Click on the ***Shared Data*** link at the top of the page:



7) Click on ***Data Libraries*** to view the CloudMap data library:



Data Libraries

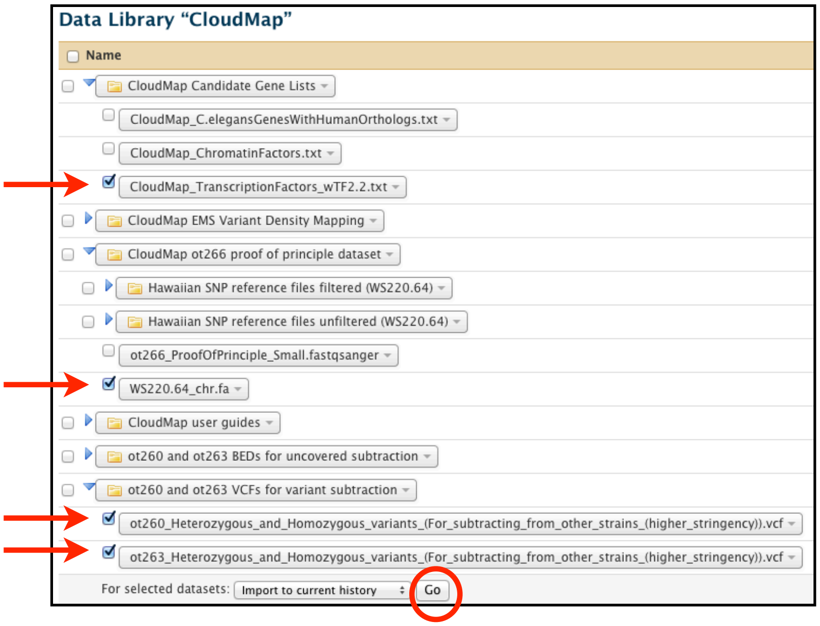
Close Advanced Search

data library name: Cloudmap

data library description:

Data library name ↓	Data library description
1000 genomes	
100209_HsMtDNA	
anton_test	
bushman	
CloudMap	Contains reference and configuration files for the Cloudmap pipeline
Codon Usage Frequencies	
Dannon's Test Data Library	Testing library for Dannon
FRIK920	
GATK	

8) Click on the ***CloudMap*** library and select the 4 data files below for the *ot266* example. Then click "Go" to import these files into your history.



Data Library "CloudMap"

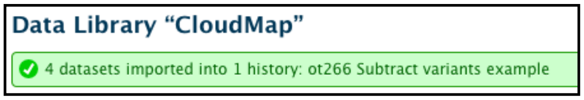
☐ Name

- ☐ CloudMap Candidate Gene Lists
  - ☐ CloudMap_C.elegansGenesWithHumanOrthologs.txt
  - ☐ CloudMap_ChromatinFactors.txt
  - ☑ CloudMap_TranscriptionFactors_wTF2.2.txt
- ☐ CloudMap EMS Variant Density Mapping
- ☐ CloudMap ot266 proof of principle dataset
  - ☐ Hawaiian SNP reference files filtered (WS220.64)
  - ☐ Hawaiian SNP reference files unfiltered (WS220.64)
  - ☐ ot266_ProofOfPrinciple_Small.fastqsanger
  - ☑ WS220.64_chr.fa
- ☐ CloudMap user guides
- ☐ ot260 and ot263 BEDs for uncovered subtraction
- ☐ ot260 and ot263 VCFs for variant subtraction
  - ☑ ot260_Heterozygous_and_Homozygous_variants_(For_subtracting_from_other_strains_(higher_stringency)).vcf
  - ☑ ot263_Heterozygous_and_Homozygous_variants_(For_subtracting_from_other_strains_(higher_stringency)).vcf

For selected datasets: Import to current history Go

In an effort to subtract as many variants as possible, we subtract not only homozygous variants from other strains, but also heterozygous variants (*ot260* and *ot263* in this example). Such a strategy assumes that phenotype-inducing homozygous mutant variants in the strain under analysis are unlikely to be heterozygous in strains that will be used for subtraction. It is especially important to apply this strategy when subtracting variant lists generated using the *Hawaiian Variant Mapping with WGS Data* approach (see section "**CloudMap *Hawaiian Variant Mapping with WGS Data*** tool"), since background variants will be present in a heterozygous state in these pooled samples as a consequence of the mapping cross. We also subtract Hawaiian SNPs in this workflow.

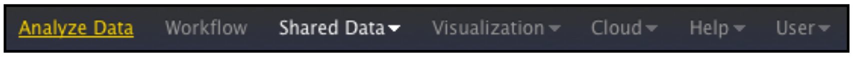
9) You will see that the files have been imported successfully:



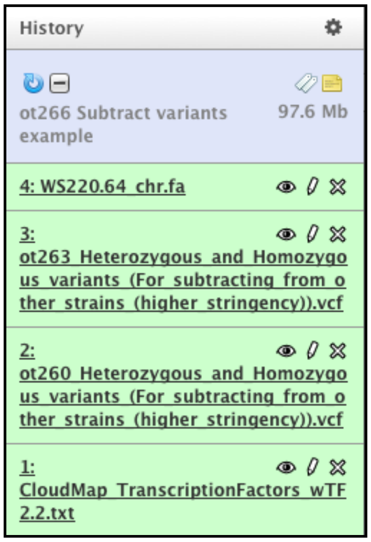
Data Library "CloudMap"

✓ 4 datasets imported into 1 history: ot266 Subtract variants example

10) Click on ***Analyze Data*** to see the files in your history:



11) You will now see these files in your history:



History

ot266 Subtract variants example 97.6 Mb

4: WS220.64_chr.fa

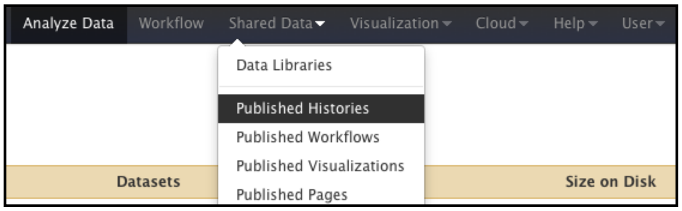
3: ot263_Heterozygous_and_Homozygous_variants_(For_subtracting_from_other_strains_(higher_stringency)).vcf

2: ot260_Heterozygous_and_Homozygous_variants_(For_subtracting_from_other_strains_(higher_stringency)).vcf

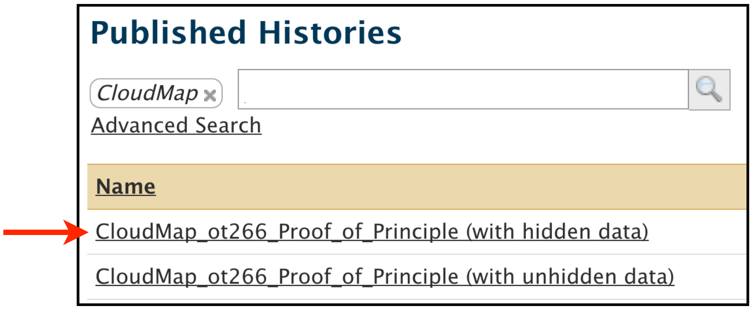
1: CloudMap_TranscriptionFactors_wTF2.2.txt

12) You will also need to import homozygous variants (VCF file) from the workflow you performed earlier. In this example, we will use the *ot266* homozygous variants from running the ***Hawaiian Variant Mapping with WGS Data and Variant Calling*** workflow. The *ot266* example history is shared so we will import the homozygous variants from that history. Note: the *ot260* and *ot263* variants that we use for data subtraction in this example come from strains that were not mapped with Hawaiian, while the *ot266* sample was mapped with Hawaiian.

Click on ***Shared Data*—> *Published Histories***:



13) Click on the history ***CloudMap: ot266 Proof of Principle (with hidden data)***:



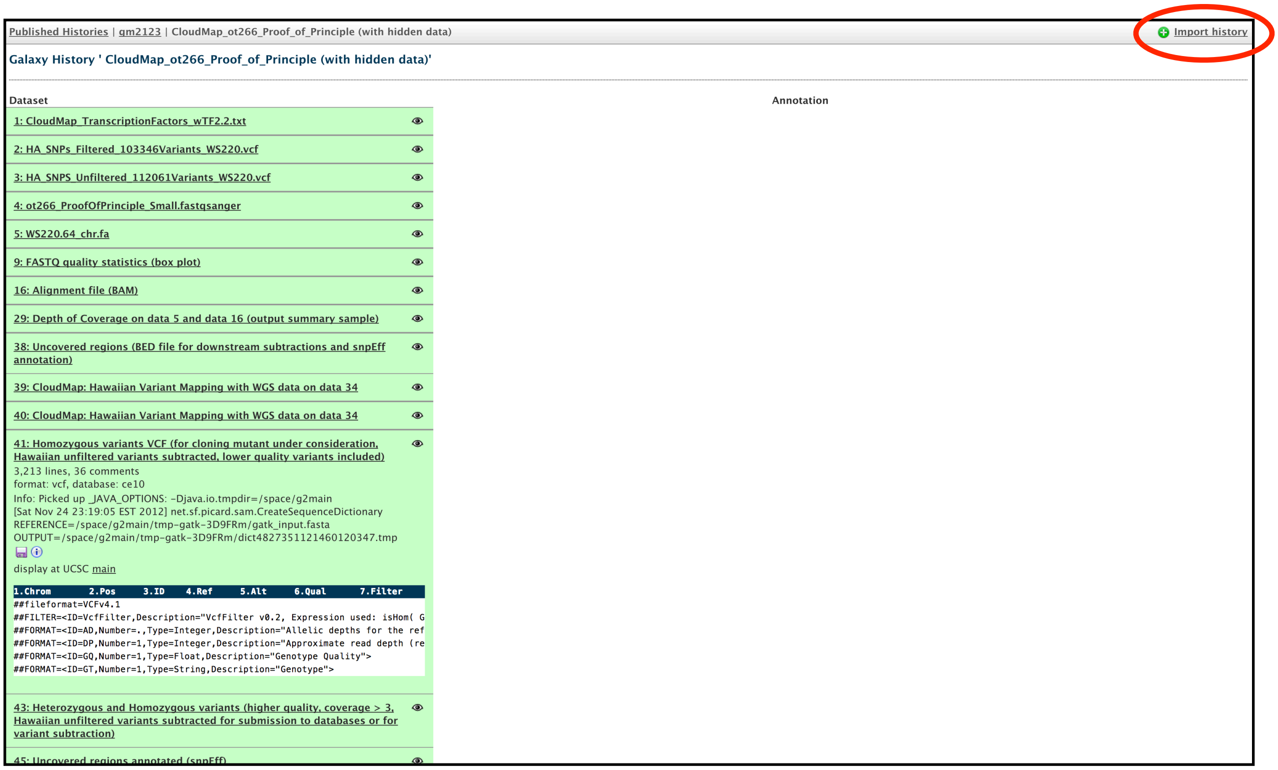
Published Histories

CloudMap

Advanced Search

Name
CloudMap_ot266_Proof_of_Principle (with hidden data)
CloudMap_ot266_Proof_of_Principle (with unhidden data)

14) Import the *ot266* history. The homozygous variants VCF we will subtract ot260 and ot263 variants from is expanded in this screenshot.



Published Histories | gm2123 | CloudMap_ot266_Proof_of_Principle (with hidden data)

Galaxy History ' CloudMap_ot266_Proof_of_Principle (with hidden data)'

1: CloudMap_TranscriptionFactors_wTF2.2.txt

2: HA_SNPs_Filtered_103346Variants_WS220.vcf

3: HA_SNPS_Unfiltered_112061Variants_WS220.vcf

4: ot266_ProofOfPrinciple_Small.fastqsanger

5: WS220.64_chr.fa

9: FASTQ quality statistics (box plot)

16: Alignment file (BAM)

29: Depth of Coverage on data 5 and data 16 (output summary sample)

38: Uncovered regions (BED file for downstream subtractions and snpEff annotation)

39: CloudMap: Hawaiian Variant Mapping with WGS data on data 34

40: CloudMap: Hawaiian Variant Mapping with WGS data on data 34

41: Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)
3,213 lines, 36 comments
format: vcf, database: ce10

```
Info: Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/space/g2main
[Sat Nov 24 23:19:05 EST 2012] net.sf.picard.sam.CreateSequenceDictionary
REFERENCE=/space/g2main/tmp-gatk-3D9FRm/gatk_input.fasta
OUTPUT=/space/g2main/tmp-gatk-3D9FRm/dict4827351121460120347.tmp
```

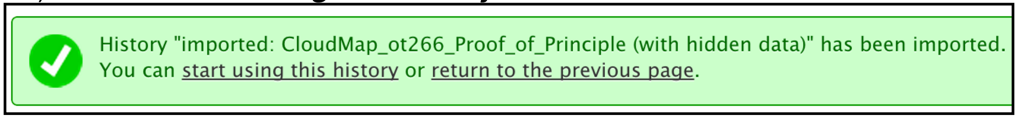
display at UCSC main

```
1.Chrom      2.Pos     3.ID    4.Ref     5.Alt     6.Qual      7.Filter
##fileformat=VCFv4.1
##FILTER=<ID=VcfFilter,Description="VcfFilter v0.2, Expression used: isHom( G
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (re
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
```

43: Heterozygous and Homozygous variants (higher quality, coverage > 3, Hawaiian unfiltered variants subtracted for submission to databases or for variant subtraction)

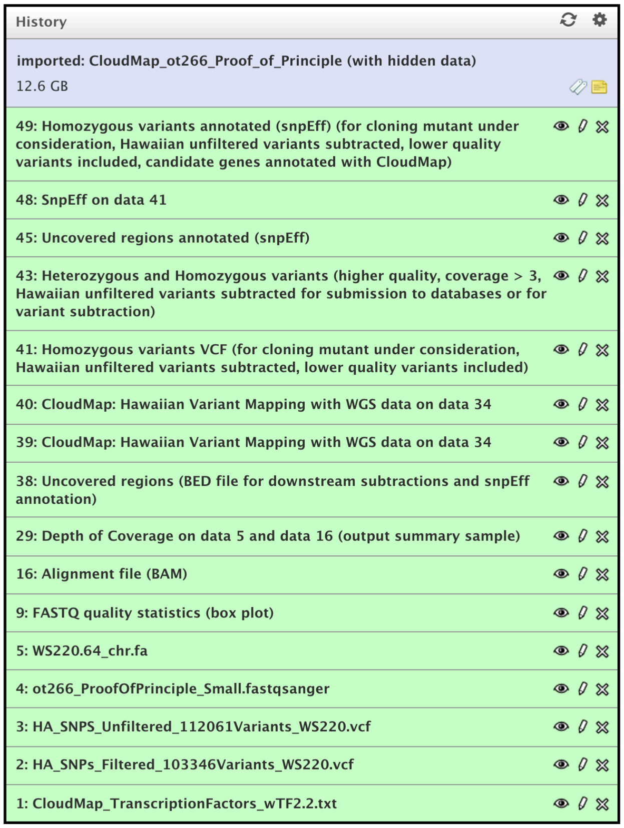
45: Uncovered regions annotated (snpEff)

15) Click the ***Start using this history*** link.



History "imported: CloudMap_ot266_Proof_of_Principle (with hidden data)" has been imported.
You can start using this history or return to the previous page.

16) You now can view all the files in the *ot266* history.



History

imported: CloudMap_ot266_Proof_of_Principle (with hidden data)
12.6 GB

49: Homozygous variants annotated (snpEff) (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included, candidate genes annotated with CloudMap)

48: SnpEff on data 41

45: Uncovered regions annotated (snpEff)

43: Heterozygous and Homozygous variants (higher quality, coverage > 3, Hawaiian unfiltered variants subtracted for submission to databases or for variant subtraction)

41: Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)

40: CloudMap: Hawaiian Variant Mapping with WGS data on data 34

39: CloudMap: Hawaiian Variant Mapping with WGS data on data 34

38: Uncovered regions (BED file for downstream subtractions and snpEff annotation)

29: Depth of Coverage on data 5 and data 16 (output summary sample)

16: Alignment file (BAM)

9: FASTQ quality statistics (box plot)

5: WS220.64_chr.fa

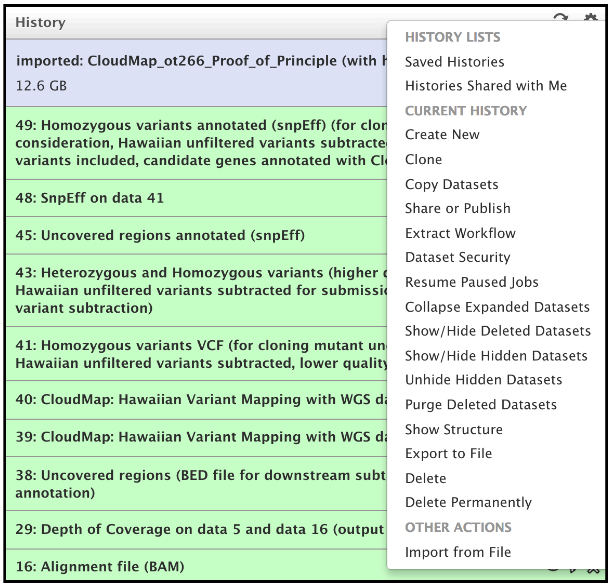
4: ot266_ProofOfPrinciple_Small.fastqsanger

3: HA_SNPS_Unfiltered_112061Variants_WS220.vcf

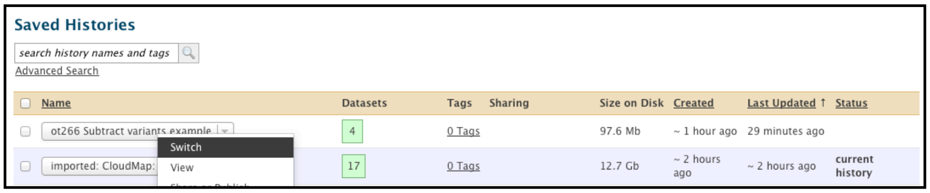
2: HA_SNPs_Filtered_103346Variants_WS220.vcf

1: CloudMap_TranscriptionFactors_wTF2.2.txt

17) Switch back to the ***ot266 Subtract Variants example*** history you created earlier by clicking ***Saved HIstories*** in your history options.

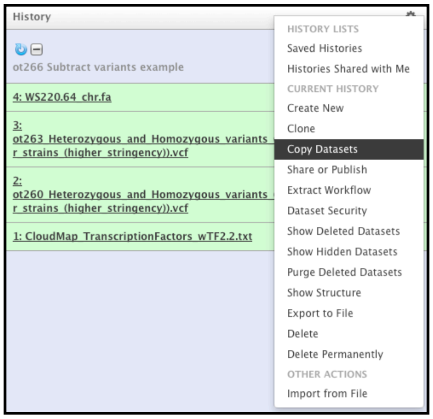

History
imported: CloudMap_ot266_Proof_of_Principle (with h
12.6 GB
49: Homozygous variants annotated (snpEff) (for clon
consideration, Hawaiian unfiltered variants subtracte
variants included, candidate genes annotated with Cl
48: SnpEff on data 41
45: Uncovered regions annotated (snpEff)
43: Heterozygous and Homozygous variants (higher c
Hawaiian unfiltered variants subtracted for submissi
variant subtraction)
41: Homozygous variants VCF (for cloning mutant un
Hawaiian unfiltered variants subtracted, lower quality
40: CloudMap: Hawaiian Variant Mapping with WGS d
39: CloudMap: Hawaiian Variant Mapping with WGS d
38: Uncovered regions (BED file for downstream subt
annotation)
29: Depth of Coverage on data 5 and data 16 (output
16: Alignment file (BAM)
HISTORY LISTS
Saved Histories
Histories Shared with Me
CURRENT HISTORY
Create New
Clone
Copy Datasets
Share or Publish
Extract Workflow
Dataset Security
Resume Paused Jobs
Collapse Expanded Datasets
Show/Hide Deleted Datasets
Show/Hide Hidden Datasets
Unhide Hidden Datasets
Purge Deleted Datasets
Show Structure
Export to File
Delete
Delete Permanently
OTHER ACTIONS
Import from File

18) Click on the ***ot266 Subtract Variants example*** history and click ***Switch*** to return to that history:



□	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
□	ot266 Subtract variants example	4	0 Tags		97.6 Mb	~ 1 hour ago	29 minutes ago	
□	imported: CloudMap:	17	0 Tags		12.7 Gb	~ 2 hours ago	~ 2 hours ago	current history

19) To copy the *ot266* homozygous variants into this history, click ***Copy Datasets*** in your history options:



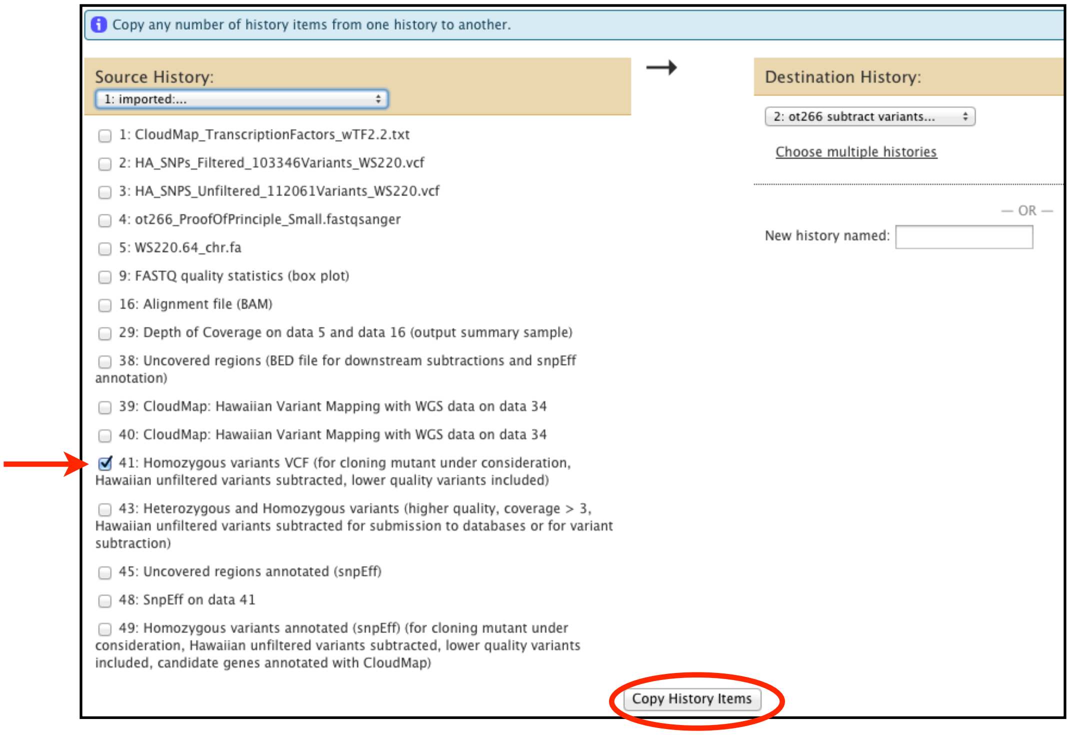
4: WS220.64_chr.fa

3: ot263_Heterozygous_and_Homozygous_variants_ r_strains_(higher_stringency)).vcf

2: ot260_Heterozygous_and_Homozygous_variants_ r_strains_(higher_stringency)).vcf

1: CloudMap_TranscriptionFactors_wTF2.2.txt

20) Copy the *ot266* ***Homozygous variants VCF*** from the newly imported *ot266* history:



Copy any number of history items from one history to another.

Source History: 1: imported:...

→

Destination History: 2: ot266 subtract variants...

Choose multiple histories

— OR —

New history named:

□ 1: CloudMap_TranscriptionFactors_wTF2.2.txt

□ 2: HA_SNPs_Filtered_103346Variants_WS220.vcf

□ 3: HA_SNPS_Unfiltered_112061Variants_WS220.vcf

□ 4: ot266_ProofOfPrinciple_Small.fastqsanger

□ 5: WS220.64_chr.fa

□ 9: FASTQ quality statistics (box plot)

□ 16: Alignment file (BAM)

□ 29: Depth of Coverage on data 5 and data 16 (output summary sample)

□ 38: Uncovered regions (BED file for downstream subtractions and snpEff annotation)

□ 39: CloudMap: Hawaiian Variant Mapping with WGS data on data 34

□ 40: CloudMap: Hawaiian Variant Mapping with WGS data on data 34

☑ 41: Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)

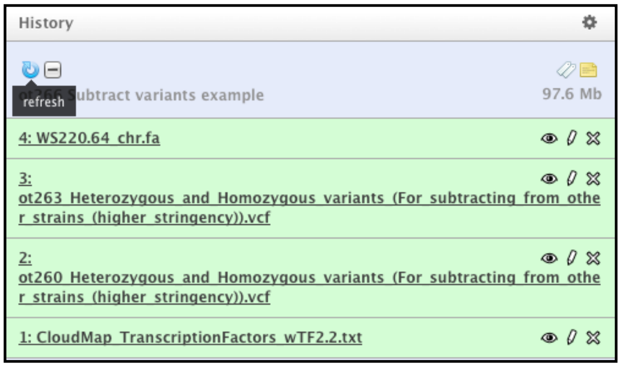
□ 43: Heterozygous and Homozygous variants (higher quality, coverage > 3, Hawaiian unfiltered variants subtracted for submission to databases or for variant subtraction)

□ 45: Uncovered regions annotated (snpEff)

□ 48: SnpEff on data 41

□ 49: Homozygous variants annotated (snpEff) (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included, candidate genes annotated with CloudMap)

21) Hit refresh in your history:



ot266 subtract variants example 97.6 Mb

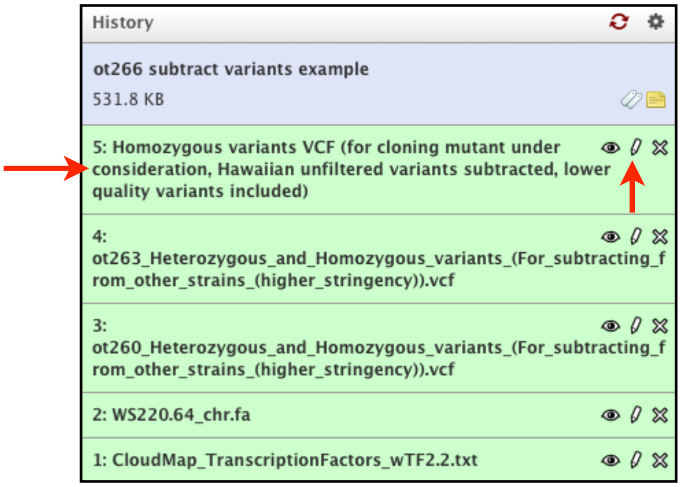
4: WS220.64_chr.fa

3: ot263_Heterozygous_and_Homozygous_variants_(For_subtracting_from_other_strains_(higher_stringency)).vcf

2: ot260_Heterozygous_and_Homozygous_variants_(For_subtracting_from_other_strains_(higher_stringency)).vcf

1: CloudMap_TranscriptionFactors_wTF2.2.txt

22) You will now see the ***ot266 Homozygous Variants*** (VCF) in your history. Click on the pencil icon to change the name of the file to add the *ot266* prefix.



ot266 subtract variants example

531.8 KB

5: Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)

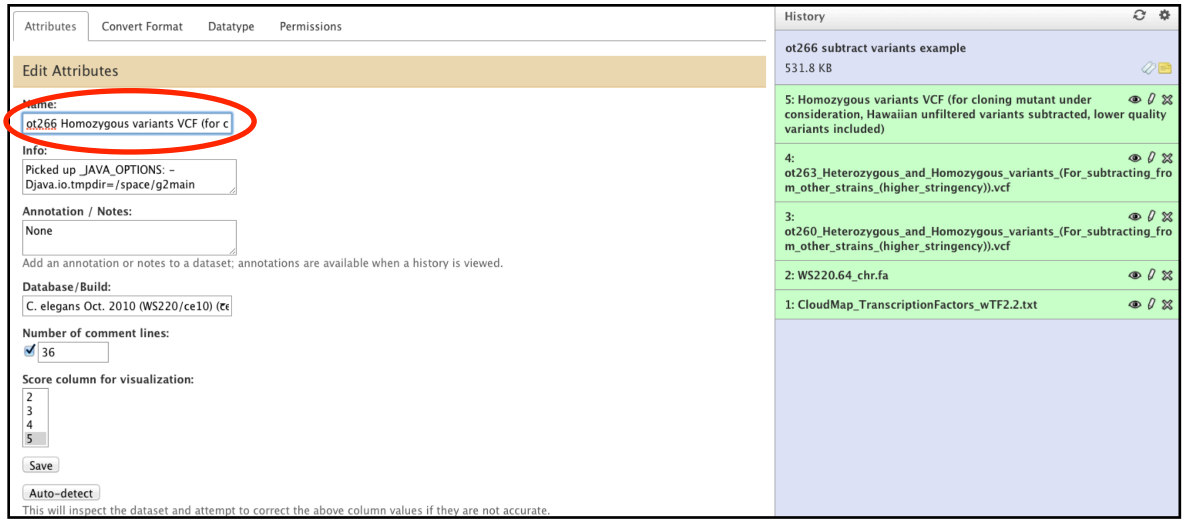
4: ot263_Heterozygous_and_Homozygous_variants_(For_subtracting_from_other_strains_(higher_stringency)).vcf

3: ot260_Heterozygous_and_Homozygous_variants_(For_subtracting_from_other_strains_(higher_stringency)).vcf

2: WS220.64_chr.fa

1: CloudMap_TranscriptionFactors_wTF2.2.txt

23) Add the *ot266* prefix to the file name:



Edit Attributes

Name: ot266 Homozygous variants VCF (for c

Info: Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/space/g2main

Annotation / Notes: None

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

Database/Build: C. elegans Oct. 2010 (WS220/ce10) (ce

Number of comment lines: ☑ 36

Score column for visualization: 2 3 4 5

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

ot266 subtract variants example

531.8 KB

5: Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)

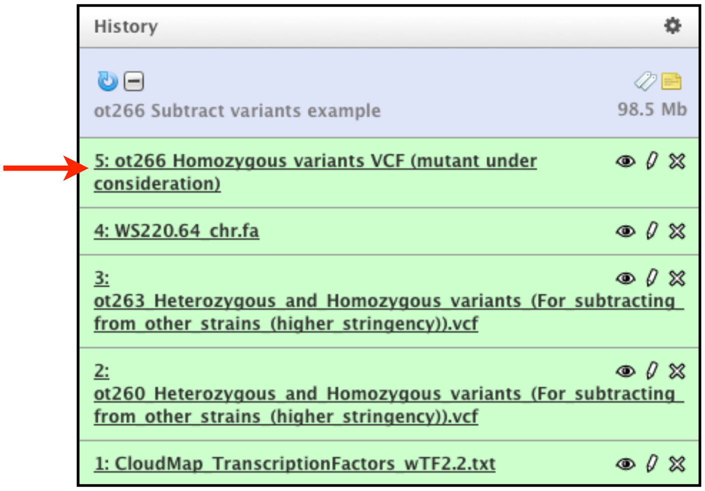
4: ot263_Heterozygous_and_Homozygous_variants_(For_subtracting_from_other_strains_(higher_stringency)).vcf

3: ot260_Heterozygous_and_Homozygous_variants_(For_subtracting_from_other_strains_(higher_stringency)).vcf

2: WS220.64_chr.fa

1: CloudMap_TranscriptionFactors_wTF2.2.txt

24) You will see that the file name has been updated:



History

ot266 Subtract variants example

98.5 Mb

5: ot266 Homozygous variants VCF (mutant under consideration)

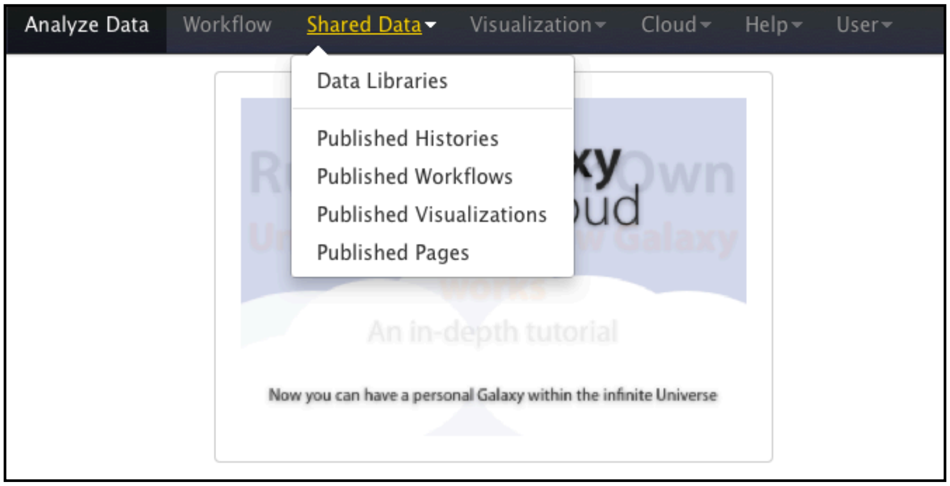
4: WS220.64 chr.fa

3: ot263 Heterozygous and Homozygous variants (For subtracting from other strains (higher stringency)).vcf

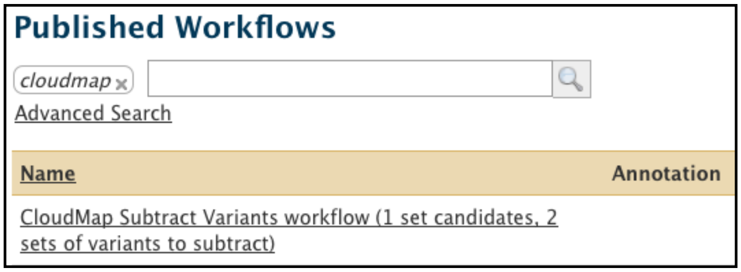
2: ot260 Heterozygous and Homozygous variants (For subtracting from other strains (higher stringency)).vcf

1: CloudMap TranscriptionFactors wTF2.2.txt

25) Now you have all the files ready to run the ***Subtract Variants*** workflow. Click on the ***Shared Data—>Published Workflows*** link at the top of the page:



26) Select the ***CloudMap Subtract Variants*** workflow:



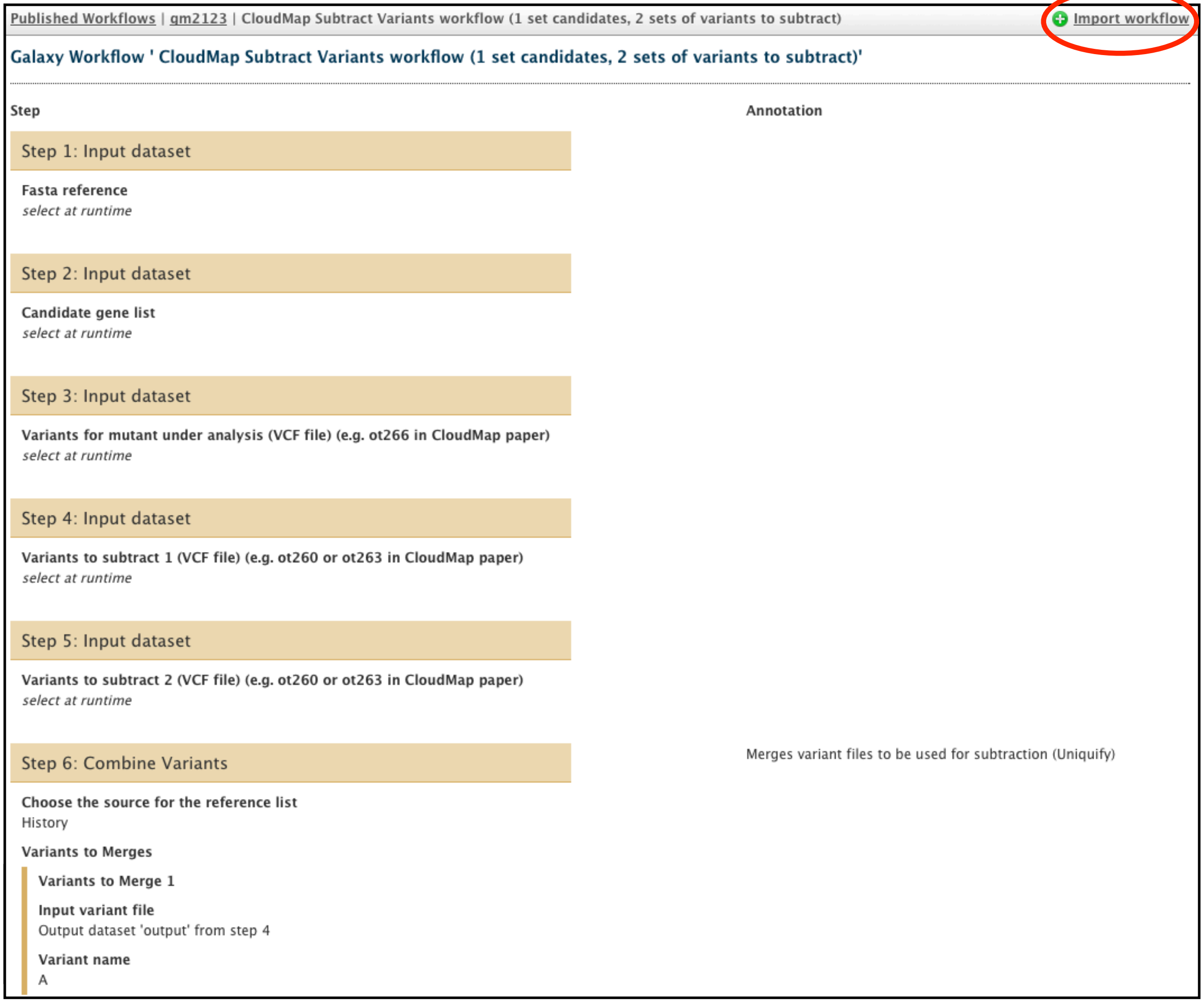
Published Workflows

cloudmap ×

Advanced Search

Name	Annotation
CloudMap Subtract Variants workflow (1 set candidates, 2 sets of variants to subtract)	

27) You will now have the option to ***Import workflow***:

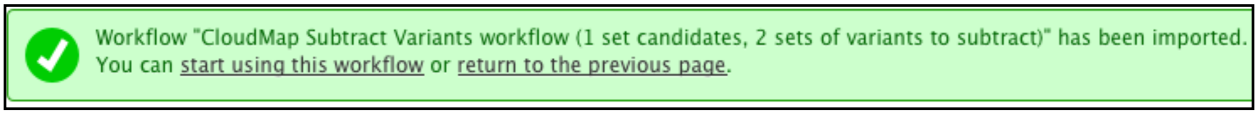


Published Workflows | gm2123 | CloudMap Subtract Variants workflow (1 set candidates, 2 sets of variants to subtract)

Galaxy Workflow ' CloudMap Subtract Variants workflow (1 set candidates, 2 sets of variants to subtract)'

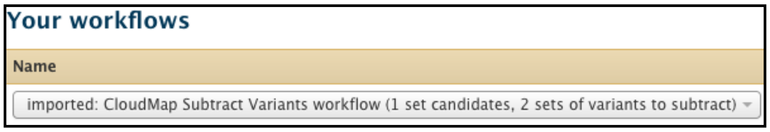
Step	Annotation
Step 1: Input dataset Fasta reference *select at runtime*	
Step 2: Input dataset Candidate gene list *select at runtime*	
Step 3: Input dataset Variants for mutant under analysis (VCF file) (e.g. ot266 in CloudMap paper) *select at runtime*	
Step 4: Input dataset Variants to subtract 1 (VCF file) (e.g. ot260 or ot263 in CloudMap paper) *select at runtime*	
Step 5: Input dataset Variants to subtract 2 (VCF file) (e.g. ot260 or ot263 in CloudMap paper) *select at runtime*	
Step 6: Combine Variants Choose the source for the reference list History Variants to Merges Variants to Merge 1 Input variant file Output dataset 'output' from step 4 Variant name A	Merges variant files to be used for subtraction (Uniquify)

28) You will see a message indicating that the workflow has been imported:



Workflow "CloudMap Subtract Variants workflow (1 set candidates, 2 sets of variants to subtract)" has been imported. You can start using this workflow or return to the previous page.

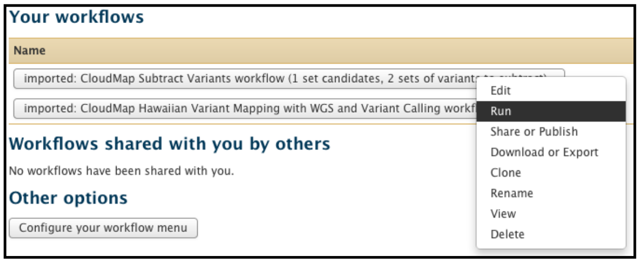
29) Click ***Start using this workflow*** and you will see that the workflow has been imported. From now on, you can easily access this workflow under the ***Workflow*** tab.



Your workflows

Name
imported: CloudMap Subtract Variants workflow (1 set candidates, 2 sets of variants to subtract)

30) Click on the workflow and select ***Run***:



Your workflows

Name

imported: CloudMap Subtract Variants workflow (1 set candidates, 2 sets of variant

imported: CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workf

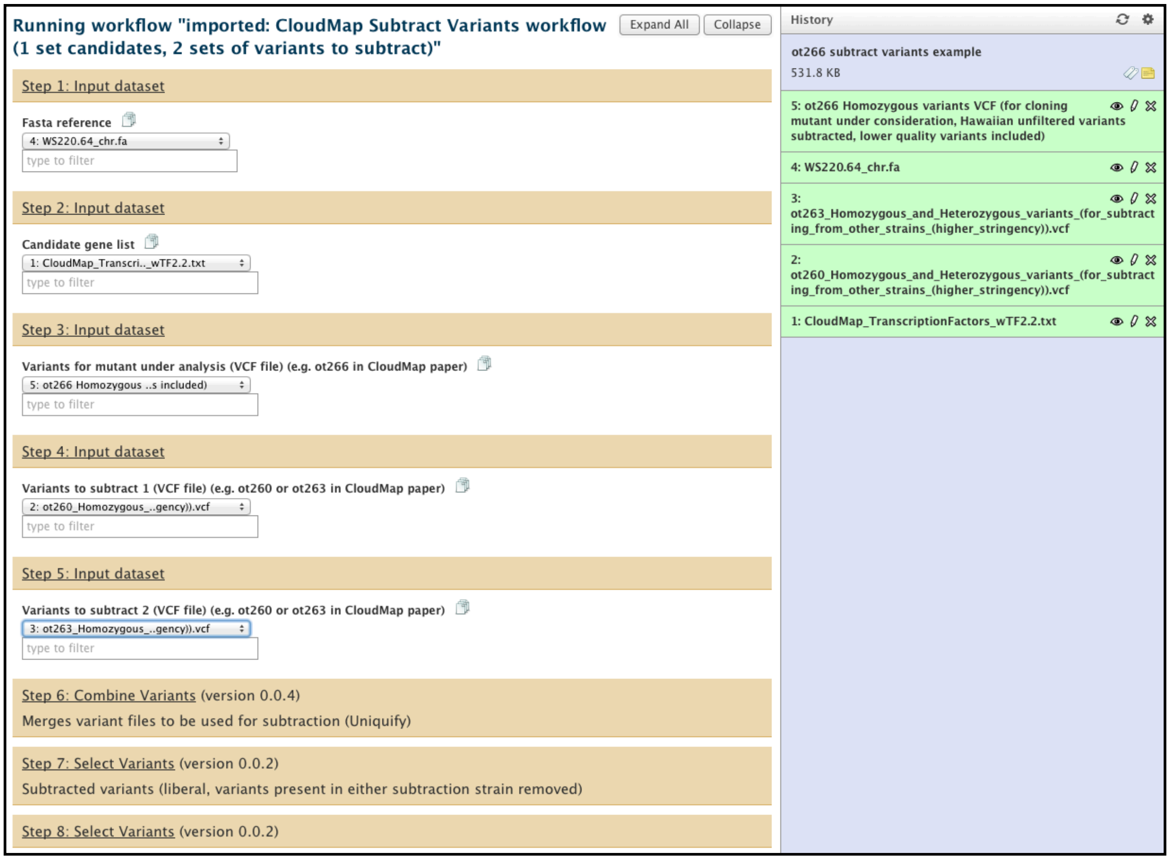
Workflows shared with you by others

No workflows have been shared with you.

Other options

Configure your workflow menu

31) You will see all the steps in the workflow prior to running it. Make sure that each of the input fields corresponds to the appropriate file in your history. Click ***Run Workflow*** when ready.



Running workflow "imported: CloudMap Subtract Variants workflow (1 set candidates, 2 sets of variants to subtract)"

Step 1: Input dataset

Fasta reference: 4: WS220.64_chr.fa

Step 2: Input dataset

Candidate gene list: 1: CloudMap_Transcri..._wTF2.2.txt

Step 3: Input dataset

Variants for mutant under analysis (VCF file) (e.g. ot266 in CloudMap paper): 5: ot266 Homozygous ..s included)

Step 4: Input dataset

Variants to subtract 1 (VCF file) (e.g. ot260 or ot263 in CloudMap paper): 2: ot260_Homozygous_..gency)).vcf

Step 5: Input dataset

Variants to subtract 2 (VCF file) (e.g. ot260 or ot263 in CloudMap paper): 3: ot263_Homozygous_..gency)).vcf

Step 6: Combine Variants (version 0.0.4)
Merges variant files to be used for subtraction (Uniquify)

Step 7: Select Variants (version 0.0.2)
Subtracted variants (liberal, variants present in either subtraction strain removed)

Step 8: Select Variants (version 0.0.2)

History

ot266 subtract variants example
531.8 KB

5: ot266 Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)

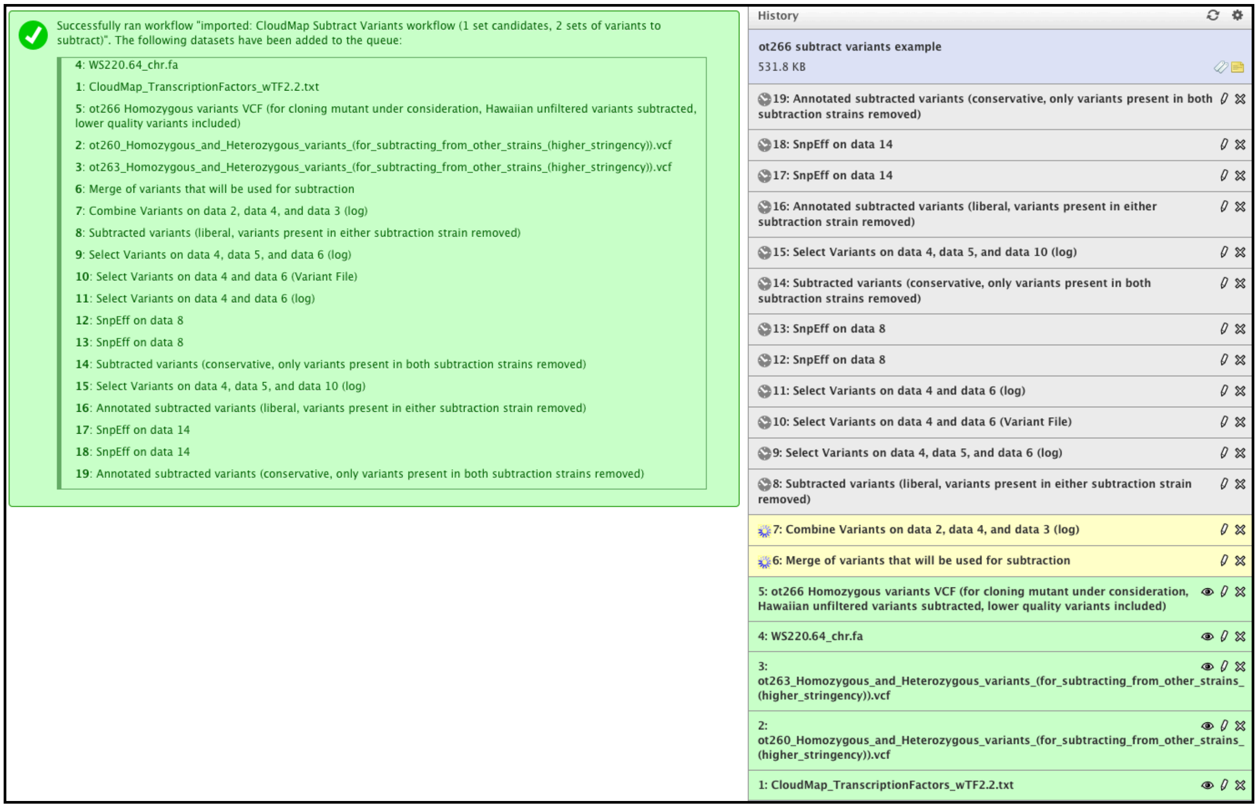
4: WS220.64_chr.fa

3: ot263_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strains_(higher_stringency)).vcf

2: ot260_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strains_(higher_stringency)).vcf

1: CloudMap_TranscriptionFactors_wTF2.2.txt

32) All of the automated functions have the appropriate default parameters configured, although experienced users may want to modify these prior to running. Once you are ready to run the workflow, press ***Run Workflow*** and the workflow will start (this step takes a minute or two to begin, be patient and don't hit the ***Run Workflow*** button repeatedly). You will receive an email when the workflow is completed:



Successfully ran workflow "imported: CloudMap Subtract Variants workflow (1 set candidates, 2 sets of variants to subtract)". The following datasets have been added to the queue:

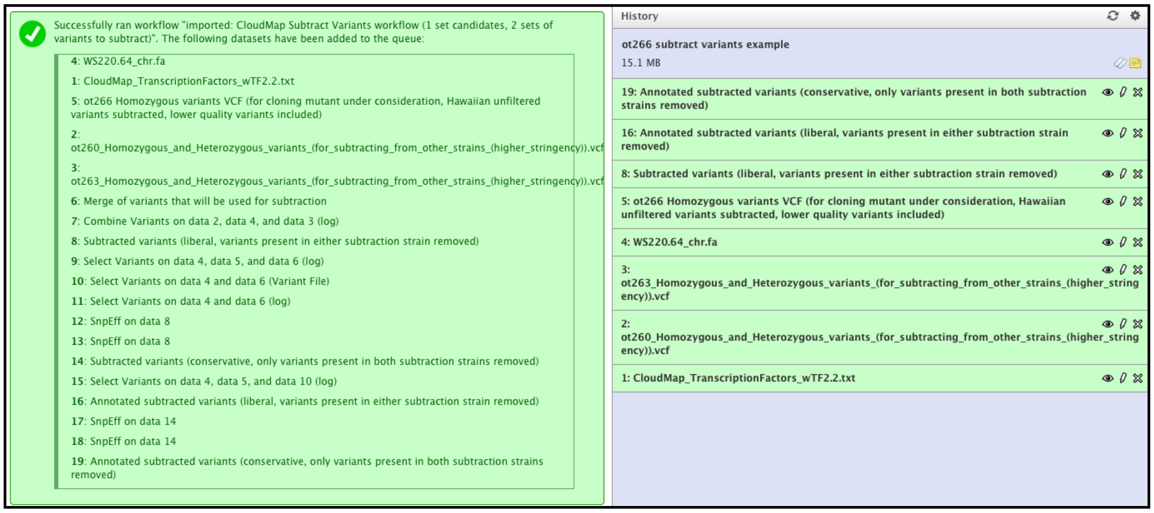
4: WS220.64_chr.fa
1: CloudMap_TranscriptionFactors_wTF2.2.txt
5: ot266 Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)
2: ot260_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strains_(higher_stringency)).vcf
3: ot263_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strains_(higher_stringency)).vcf
6: Merge of variants that will be used for subtraction
7: Combine Variants on data 2, data 4, and data 3 (log)
8: Subtracted variants (liberal, variants present in either subtraction strain removed)
9: Select Variants on data 4, data 5, and data 6 (log)
10: Select Variants on data 4 and data 6 (Variant File)
11: Select Variants on data 4 and data 6 (log)
12: SnpEff on data 8
13: SnpEff on data 8
14: Subtracted variants (conservative, only variants present in both subtraction strains removed)
15: Select Variants on data 4, data 5, and data 10 (log)
16: Annotated subtracted variants (liberal, variants present in either subtraction strain removed)
17: SnpEff on data 14
18: SnpEff on data 14
19: Annotated subtracted variants (conservative, only variants present in both subtraction strains removed)

History

ot266 subtract variants example
531.8 KB

19: Annotated subtracted variants (conservative, only variants present in both subtraction strains removed)
18: SnpEff on data 14
17: SnpEff on data 14
16: Annotated subtracted variants (liberal, variants present in either subtraction strain removed)
15: Select Variants on data 4, data 5, and data 10 (log)
14: Subtracted variants (conservative, only variants present in both subtraction strains removed)
13: SnpEff on data 8
12: SnpEff on data 8
11: Select Variants on data 4 and data 6 (log)
10: Select Variants on data 4 and data 6 (Variant File)
9: Select Variants on data 4, data 5, and data 6 (log)
8: Subtracted variants (liberal, variants present in either subtraction strain removed)
7: Combine Variants on data 2, data 4, and data 3 (log)
6: Merge of variants that will be used for subtraction
5: ot266 Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)
4: WS220.64_chr.fa
3: ot263_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strains_(higher_stringency)).vcf
2: ot260_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strains_(higher_stringency)).vcf
1: CloudMap_TranscriptionFactors_wTF2.2.txt

33) The workflow has finished running and you can view the resulting output:



Successfully ran workflow "imported: CloudMap Subtract Variants workflow (1 set candidates, 2 sets of variants to subtract)". The following datasets have been added to the queue:

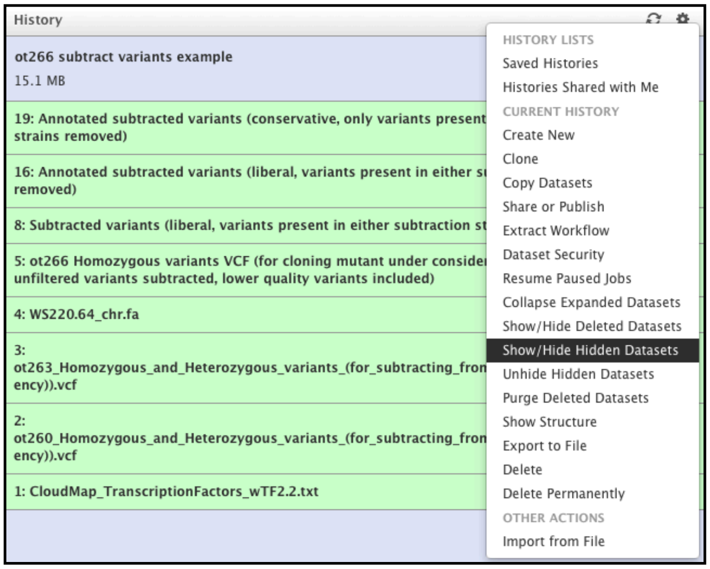
4: WS220.64_chr.fa
1: CloudMap_TranscriptionFactors_wTF2.2.txt
5: ot266 Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)
2: ot260_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strains_(higher_stringency)).vcf
3: ot263_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strains_(higher_stringency)).vcf
6: Merge of variants that will be used for subtraction
7: Combine Variants on data 2, data 4, and data 3 (log)
8: Subtracted variants (liberal, variants present in either subtraction strain removed)
9: Select Variants on data 4, data 5, and data 6 (log)
10: Select Variants on data 4 and data 6 (Variant File)
11: Select Variants on data 4 and data 6 (log)
12: SnpEff on data 8
13: SnpEff on data 8
14: Subtracted variants (conservative, only variants present in both subtraction strains removed)
15: Select Variants on data 4, data 5, and data 10 (log)
16: Annotated subtracted variants (liberal, variants present in either subtraction strain removed)
17: SnpEff on data 14
18: SnpEff on data 14
19: Annotated subtracted variants (conservative, only variants present in both subtraction strains removed)

History

ot266 subtract variants example
15.1 MB

19: Annotated subtracted variants (conservative, only variants present in both subtraction strains removed)
16: Annotated subtracted variants (liberal, variants present in either subtraction strain removed)
8: Subtracted variants (liberal, variants present in either subtraction strain removed)
5: ot266 Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)
4: WS220.64_chr.fa
3: ot263_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strains_(higher_stringency)).vcf
2: ot260_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strains_(higher_stringency)).vcf
1: CloudMap_TranscriptionFactors_wTF2.2.txt

34) You will notice that while approximately 20 output files were generated during the course of the workflow (output files are sequentially numbered), only some output files remain visible while others are hidden. The visible files are most important for analysis of the mutant under consideration or downstream analysis. In order to view hidden files, click ***Show Hidden Datasets*** in the History menu:



History

ot266 subtract variants example
15.1 MB

19: Annotated subtracted variants (conservative, only variants present
strains removed)

16: Annotated subtracted variants (liberal, variants present in either su
removed)

8: Subtracted variants (liberal, variants present in either subtraction st

5: ot266 Homozygous variants VCF (for cloning mutant under consider
unfiltered variants subtracted, lower quality variants included)

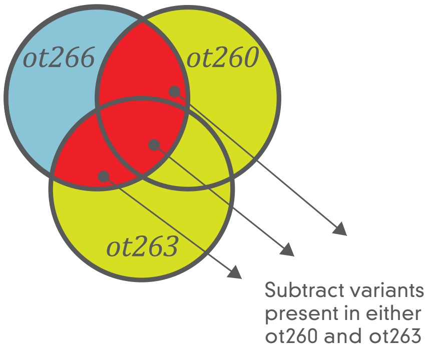
4: WS220.64_chr.fa

3:
ot263_Homozygous_and_Heterozygous_variants_(for_subtracting_fron
ency)).vcf

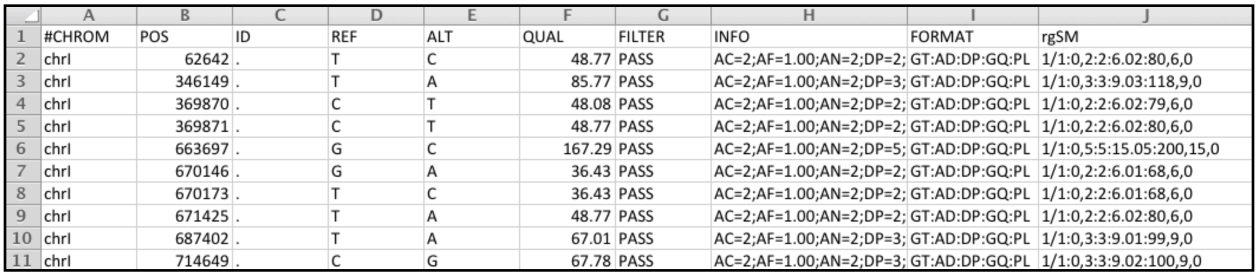
2:
ot260_Homozygous_and_Heterozygous_variants_(for_subtracting_fron
ency)).vcf

1: CloudMap_TranscriptionFactors_wTF2.2.txt

35) There are 3 main output files. The first, named ***Subtracted variants (liberal, variants present in either subtraction strain removed)*** is a VCF file generated by *GATK* that corresponds to the variant subtraction described in **Fig.8** of the CloudMap paper.

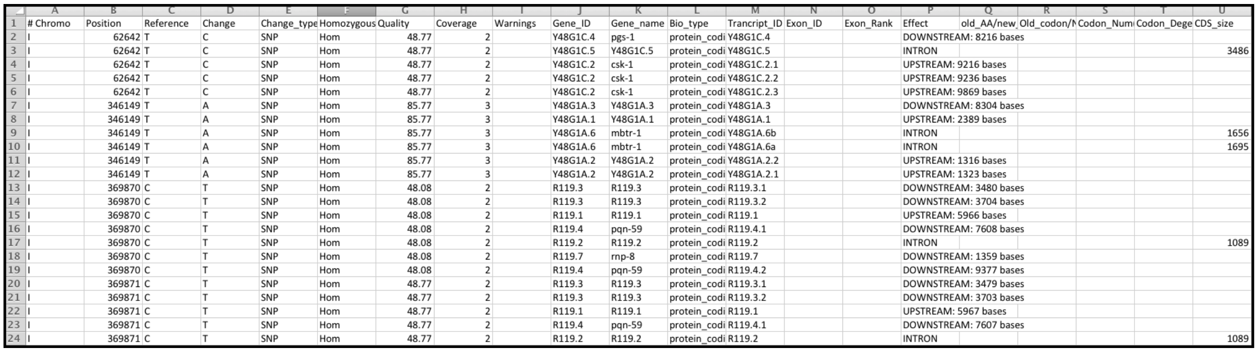


This file contains *ot266* homozygous variants after both homozygous and heterozygous variants present in **<u>either</u>** *ot260* **<u>or</u>** *ot263* have been subtracted. This file should be downloaded to be easily viewed in its entirety. The first several lines in any VCF file are header lines starting with "#" so users who wish to filter or sort these files in Excel are advised to remove the header lines. (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat.html). Below you can see a snippet of the file after header lines have been removed:



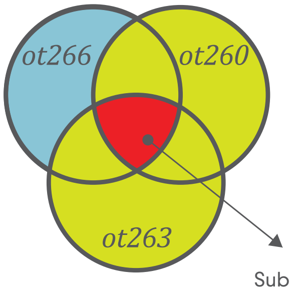
	A	B	C	D	E	F	G	H	I	J
1	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	rgSM
2	chrI	62642	.	T	C	48.77	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:80,6,0
3	chrI	346149	.	T	A	85.77	PASS	AC=2;AF=1.00;AN=2;DP=3;	GT:AD:DP:GQ:PL	1/1:0,3:3:9.03:118,9,0
4	chrI	369870	.	C	T	48.08	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:79,6,0
5	chrI	369871	.	C	T	48.77	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:80,6,0
6	chrI	663697	.	G	C	167.29	PASS	AC=2;AF=1.00;AN=2;DP=5;	GT:AD:DP:GQ:PL	1/1:0,5:5:15.05:200,15,0
7	chrI	670146	.	G	A	36.43	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.01:68,6,0
8	chrI	670173	.	T	C	36.43	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.01:68,6,0
9	chrI	671425	.	T	A	48.77	PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:80,6,0
10	chrI	687402	.	T	A	67.01	PASS	AC=2;AF=1.00;AN=2;DP=3;	GT:AD:DP:GQ:PL	1/1:0,3:3:9.01:99,9,0
11	chrI	714649	.	C	G	67.78	PASS	AC=2;AF=1.00;AN=2;DP=3;	GT:AD:DP:GQ:PL	1/1:0,3:3:9.02:100,9,0

36) The file ***Annotated subtracted variants (liberal, variants present in either subtraction strain removed)*** is simply the VCF described in the previous step which has now had its variants annotated for their predicted effect on genes with *snpEff*. The ***CloudMap Candidate Checker*** has also annotated any candidate genes that appear in the *snpEff* output.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	# Chromo	Position	Reference	Change	Change_type	Homozygous	Quality	Coverage	Warnings	Gene_ID	Gene_name	Bio_type	Trancript_ID	Exon_ID	Exon_Rank	Effect	old_AA/new_	Old_codon/N	Codon_Num	Codon_Dege	CDS_size
2	I	62642	T	C	SNP	Hom	48.77	2		Y48G1C.4	pgs-1	protein_codi	Y48G1C.4			DOWNSTREAM: 8216 bases					
3	I	62642	T	C	SNP	Hom	48.77	2		Y48G1C.5	Y48G1C.5	protein_codi	Y48G1C.5			INTRON					3486
4	I	62642	T	C	SNP	Hom	48.77	2		Y48G1C.2	csk-1	protein_codi	Y48G1C.2.1			UPSTREAM: 9216 bases					
5	I	62642	T	C	SNP	Hom	48.77	2		Y48G1C.2	csk-1	protein_codi	Y48G1C.2.2			UPSTREAM: 9236 bases					
6	I	62642	T	C	SNP	Hom	48.77	2		Y48G1C.2	csk-1	protein_codi	Y48G1C.2.3			UPSTREAM: 9869 bases					
7	I	346149	T	A	SNP	Hom	85.77	3		Y48G1A.3	Y48G1A.3	protein_codi	Y48G1A.3			DOWNSTREAM: 8304 bases					
8	I	346149	T	A	SNP	Hom	85.77	3		Y48G1A.1	Y48G1A.1	protein_codi	Y48G1A.1			UPSTREAM: 2389 bases					
9	I	346149	T	A	SNP	Hom	85.77	3		Y48G1A.6	mbtr-1	protein_codi	Y48G1A.6b			INTRON					1656
10	I	346149	T	A	SNP	Hom	85.77	3		Y48G1A.6	mbtr-1	protein_codi	Y48G1A.6a			INTRON					1695
11	I	346149	T	A	SNP	Hom	85.77	3		Y48G1A.2	Y48G1A.2	protein_codi	Y48G1A.2.2			UPSTREAM: 1316 bases					
12	I	346149	T	A	SNP	Hom	85.77	3		Y48G1A.2	Y48G1A.2	protein_codi	Y48G1A.2.1			UPSTREAM: 1323 bases					
13	I	369870	C	T	SNP	Hom	48.08	2		R119.3	R119.3	protein_codi	R119.3.1			DOWNSTREAM: 3480 bases					
14	I	369870	C	T	SNP	Hom	48.08	2		R119.3	R119.3	protein_codi	R119.3.2			DOWNSTREAM: 3704 bases					
15	I	369870	C	T	SNP	Hom	48.08	2		R119.1	R119.1	protein_codi	R119.1			UPSTREAM: 5966 bases					
16	I	369870	C	T	SNP	Hom	48.08	2		R119.4	pqn-59	protein_codi	R119.4.1			DOWNSTREAM: 7608 bases					
17	I	369870	C	T	SNP	Hom	48.08	2		R119.2	R119.2	protein_codi	R119.2			INTRON					1089
18	I	369870	C	T	SNP	Hom	48.08	2		R119.7	rnp-8	protein_codi	R119.7			DOWNSTREAM: 1359 bases					
19	I	369870	C	T	SNP	Hom	48.08	2		R119.4	pqn-59	protein_codi	R119.4.2			DOWNSTREAM: 9377 bases					
20	I	369871	C	T	SNP	Hom	48.77	2		R119.3	R119.3	protein_codi	R119.3.1			DOWNSTREAM: 3479 bases					
21	I	369871	C	T	SNP	Hom	48.77	2		R119.3	R119.3	protein_codi	R119.3.2			DOWNSTREAM: 3703 bases					
22	I	369871	C	T	SNP	Hom	48.77	2		R119.1	R119.1	protein_codi	R119.1			UPSTREAM: 5967 bases					
23	I	369871	C	T	SNP	Hom	48.77	2		R119.4	pqn-59	protein_codi	R119.4.1			DOWNSTREAM: 7607 bases					
24	I	369871	C	T	SNP	Hom	48.77	2		R119.2	R119.2	protein_codi	R119.2			INTRON					1089

37) The final file, ***Annotated subtracted variants (conservative, only variants present in both subtraction strains removed)*** is exactly the same as the file in step #36 with the only exception being that only variants present in **<u>both</u>** *ot260* and *ot263* were subtracted from *ot266*. We label this file "conservative" because it is less likely that a causal variant in ot266 will be incorrectly subtracted since that same causal variant would have to be present in **<u>both</u>** *ot260* and *ot263*.



Subtract variants present in both ot260 and ot263

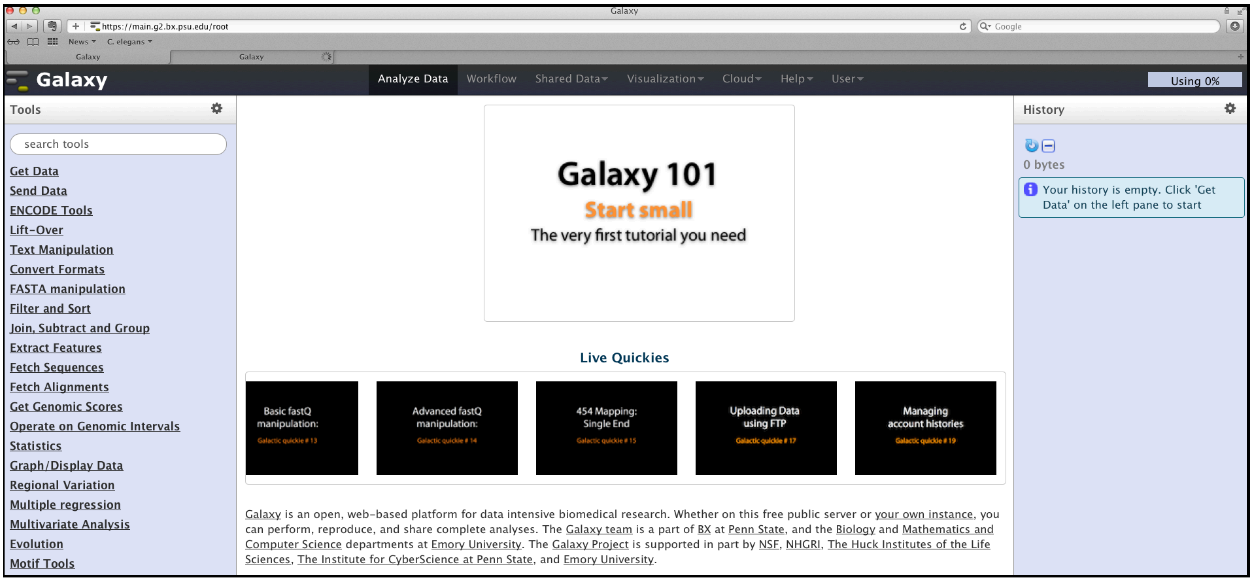
Note: We strongly suggest that users employ the ***Uncovered Region Subtraction*** workflow using the same strains (from their own screens) used in this workflow for variant subtraction. The general concept is shown in **Fig.5** of the CloudMap paper and is the same as used in this ***Subtract Variants*** workflow.

Also, please note that the number of variants per sample in this example do not match that in **Fig.8** of the CloudMap paper because the ot266 dataset used is a small subset of the full FASTQ file for that sample.

CloudMap Uncovered Region Subtraction workflow (using *ot266* Proof of Principle example from the CloudMap paper). A video version of this user guide is available at: http://usegalaxy.org/cloudmap. This workflow should be used downstream of either of the following workflows: ***Hawaiian Variant Mapping with WGS data and Variant Calling , EMS Density Mapping, or Unmapped Mutant workflows.*** Here we demonstrate the workflow using the *ot266* example from the Cloudmap paper (**Fig.8**). The goal is to subtract uncovered regions present in both *ot260* and *ot263* from uncovered regions in *ot266* (all from the same starting strain) and then to annotate the resulting uncovered regions for whether they intersect with functional genomic units (genes, ncRNAs, etc). Users may apply this workflow to their own data by substituting the datasets in this example with their own datasets.

These workflows provide default function parameters, ensuring that users follow best practices, and allow for automated execution of sequential operations. We provide these workflows as helpful guides, but experienced users may execute functions in any meaningful order they please and may also create and share their own workflows to take advantage of the automation feature. More CloudMap documentation is available at http://usegalaxy.org/cloudmap.

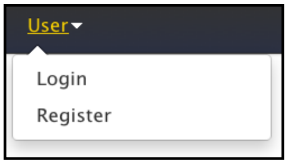
1) Navigate to http://usegalaxy.org



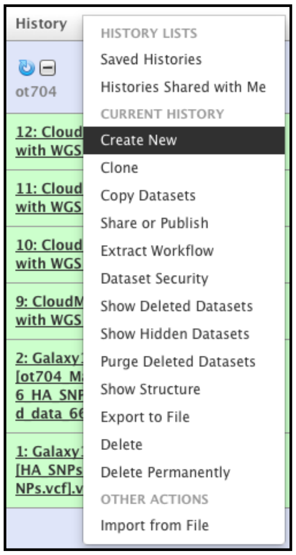
Your history is empty. Click 'Get Data' on the left pane to start

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

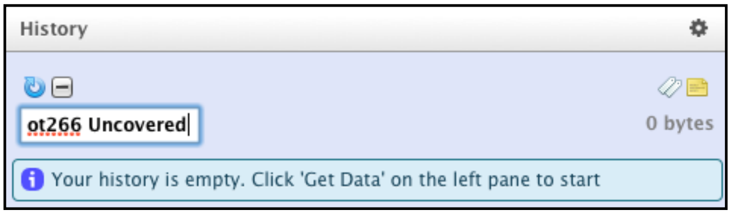
2) You should already have a Galaxy account at this point because you have run earlier workflows:



3) Once you are logged in using your email address, create a new history:



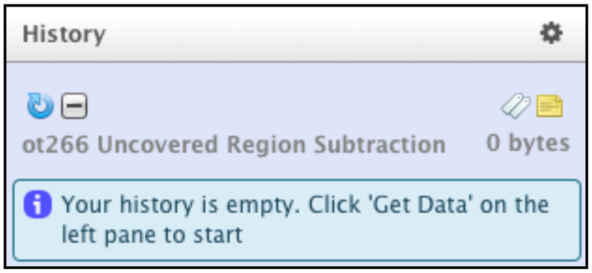
4) Now name that history:



ot266 Uncovered

Your history is empty. Click 'Get Data' on the left pane to start

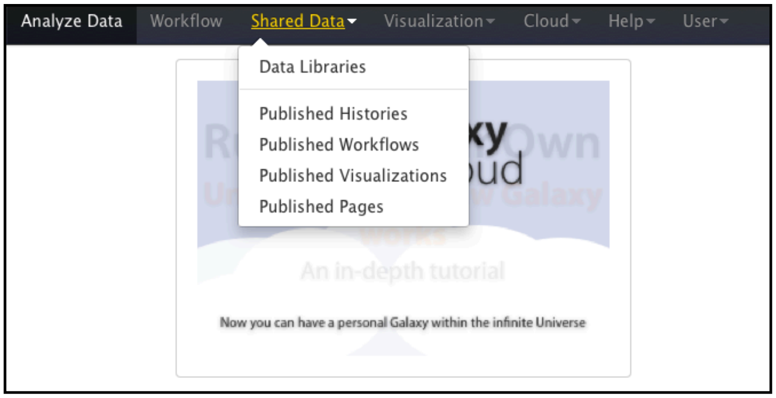
5) The history has been renamed.



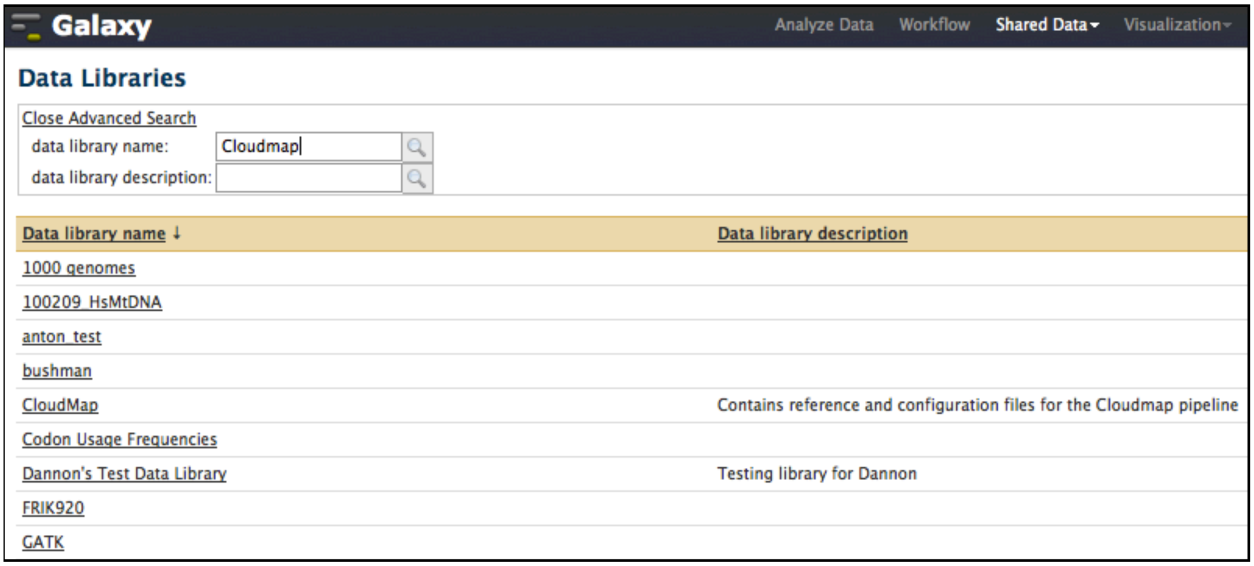
ot266 Uncovered Region Subtraction

Your history is empty. Click 'Get Data' on the left pane to start

6) You now need to import the ***ot266 Proof of principle*** files (from the CloudMap Shared Data library) or your own files to run the workflow (***See the Analyze Your Own Data Using CloudMap Workflows*** section of this user guide).



7) Click on ***Data Libraries*** to view the CloudMap data library:



Data Libraries

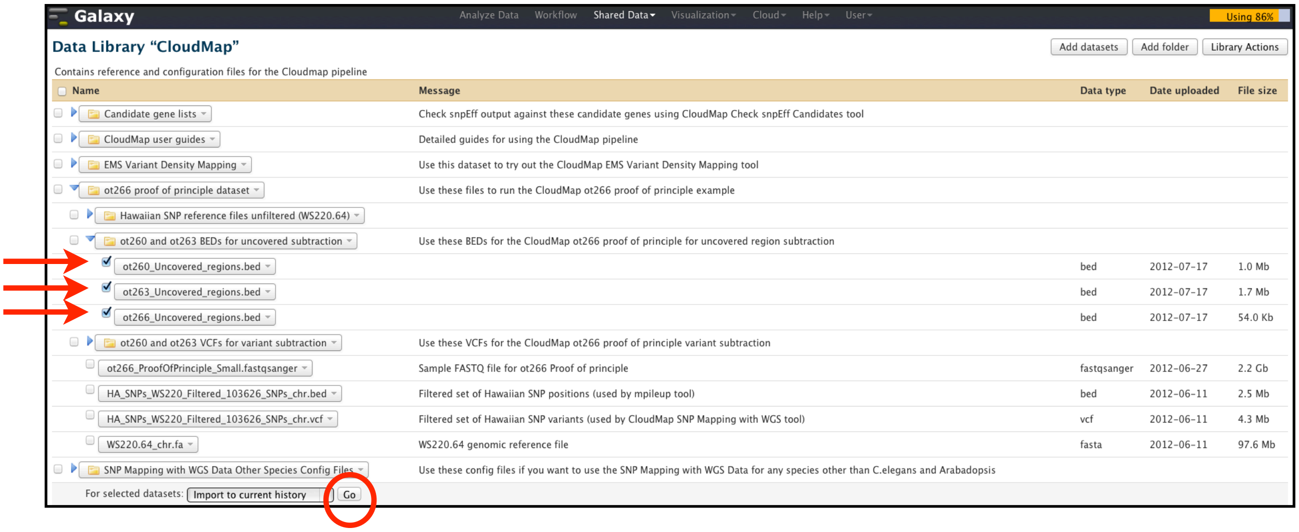
Close Advanced Search

data library name: Cloudmap

data library description:

Data library name ↓	Data library description
1000 genomes	
100209_HsMtDNA	
anton_test	
bushman	
CloudMap	Contains reference and configuration files for the Cloudmap pipeline
Codon Usage Frequencies	
Dannon's Test Data Library	Testing library for Dannon
FRIK920	
GATK	

8) Click on the ***CloudMap*** library and select the 3 data files below for the *ot266* example. Then click “Go” to import these files into your history.



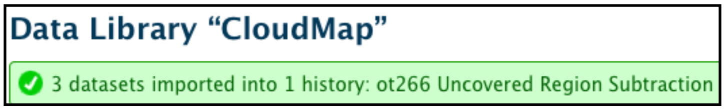
Data Library “CloudMap”

Contains reference and configuration files for the Cloudmap pipeline

Name	Message	Data type	Date uploaded	File size
□ Candidate gene lists	Check snpEff output against these candidate genes using CloudMap Check snpEff Candidates tool			
□ CloudMap user guides	Detailed guides for using the CloudMap pipeline			
□ EMS Variant Density Mapping	Use this dataset to try out the CloudMap EMS Variant Density Mapping tool			
□ ot266 proof of principle dataset	Use these files to run the CloudMap ot266 proof of principle example			
□ Hawaiian SNP reference files unfiltered (WS220.64)				
□ ot260 and ot263 BEDs for uncovered subtraction	Use these BEDs for the CloudMap ot266 proof of principle for uncovered region subtraction			
☑ ot260_Uncovered_regions.bed		bed	2012-07-17	1.0 Mb
☑ ot263_Uncovered_regions.bed		bed	2012-07-17	1.7 Mb
☑ ot266_Uncovered_regions.bed		bed	2012-07-17	54.0 Kb
□ ot260 and ot263 VCFs for variant subtraction	Use these VCFs for the CloudMap ot266 proof of principle variant subtraction			
□ ot266_ProofOfPrinciple_Small.fastqsanger	Sample FASTQ file for ot266 Proof of principle	fastqsanger	2012-06-27	2.2 Gb
□ HA_SNPs_WS220_Filtered_103626_SNPs_chr.bed	Filtered set of Hawaiian SNP positions (used by mpileup tool)	bed	2012-06-11	2.5 Mb
□ HA_SNPs_WS220_Filtered_103626_SNPs_chr.vcf	Filtered set of Hawaiian SNP variants (used by CloudMap SNP Mapping with WGS tool)	vcf	2012-06-11	4.3 Mb
□ WS220.64_chr.fa	WS220.64 genomic reference file	fasta	2012-06-11	97.6 Mb
□ SNP Mapping with WGS Data Other Species Config Files	Use these config files if you want to use the SNP Mapping with WGS Data for any species other than C.elegans and Arabadopsis			

For selected datasets: Import to current history | Go

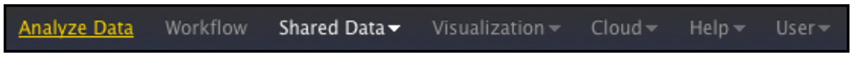
9) You will see that the files have been imported successfully:



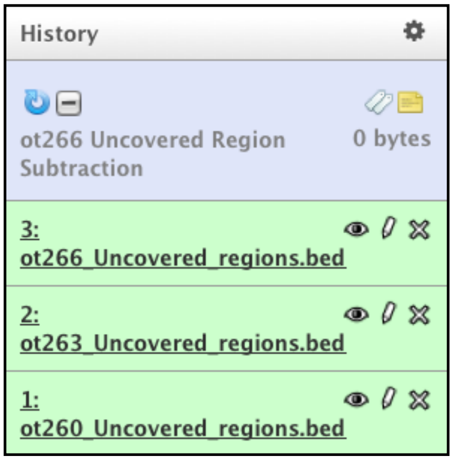
Data Library “CloudMap”

✓ 3 datasets imported into 1 history: ot266 Uncovered Region Subtraction

10) Click on ***Analyze Data*** to see the files in your history:



11) You will now see these files in your history:



History

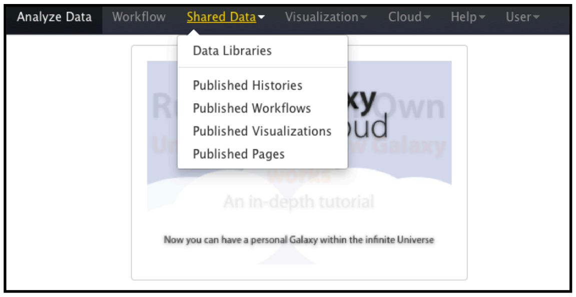
ot266 Uncovered Region Subtraction 0 bytes

3: ot266_Uncovered_regions.bed

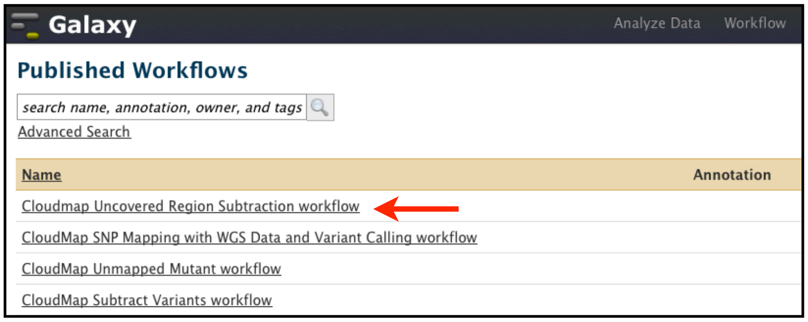
2: ot263_Uncovered_regions.bed

1: ot260_Uncovered_regions.bed

12) Now you have all the files ready to run the ***Uncovered Region Subtraction*** workflow. Click on the ***Shared Data—>Published Workflows*** link at the top of the page:



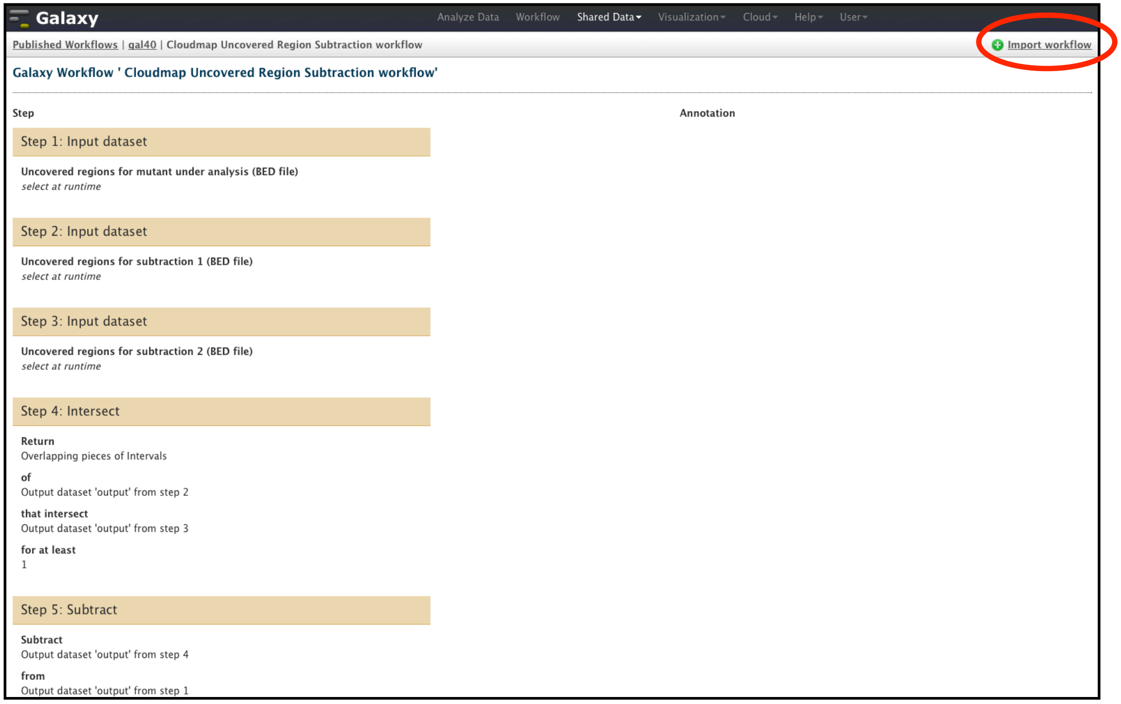
13) Select the ***Uncovered Region Subtraction*** workflow:



Published Workflows

Name	Annotation
Cloudmap Uncovered Region Subtraction workflow	
CloudMap SNP Mapping with WGS Data and Variant Calling workflow	
CloudMap Unmapped Mutant workflow	
CloudMap Subtract Variants workflow	

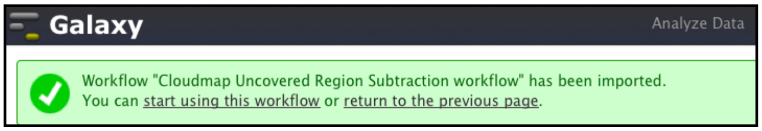
14) You will now have the option to ***Import workflow***:



Galaxy Workflow ' Cloudmap Uncovered Region Subtraction workflow'

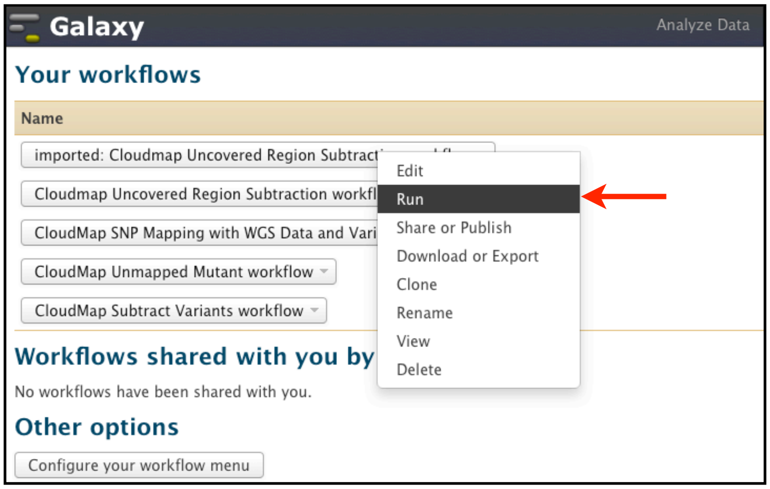
Step	Annotation
Step 1: Input dataset	
Uncovered regions for mutant under analysis (BED file) *select at runtime*	
Step 2: Input dataset	
Uncovered regions for subtraction 1 (BED file) *select at runtime*	
Step 3: Input dataset	
Uncovered regions for subtraction 2 (BED file) *select at runtime*	
Step 4: Intersect	
Return Overlapping pieces of Intervals **of** Output dataset 'output' from step 2 **that intersect** Output dataset 'output' from step 3 **for at least** 1	
Step 5: Subtract	
Subtract Output dataset 'output' from step 4 **from** Output dataset 'output' from step 1	

15) You will see a message indicating that the workflow has been imported:



Workflow "Cloudmap Uncovered Region Subtraction workflow" has been imported.
You can start using this workflow or return to the previous page.

16) Click ***Start using this workflow*** and you will see that the workflow has been imported. From now on, you can easily access this workflow under the ***Workflow*** tab. Click on the workflow and select ***Run***:



Your workflows

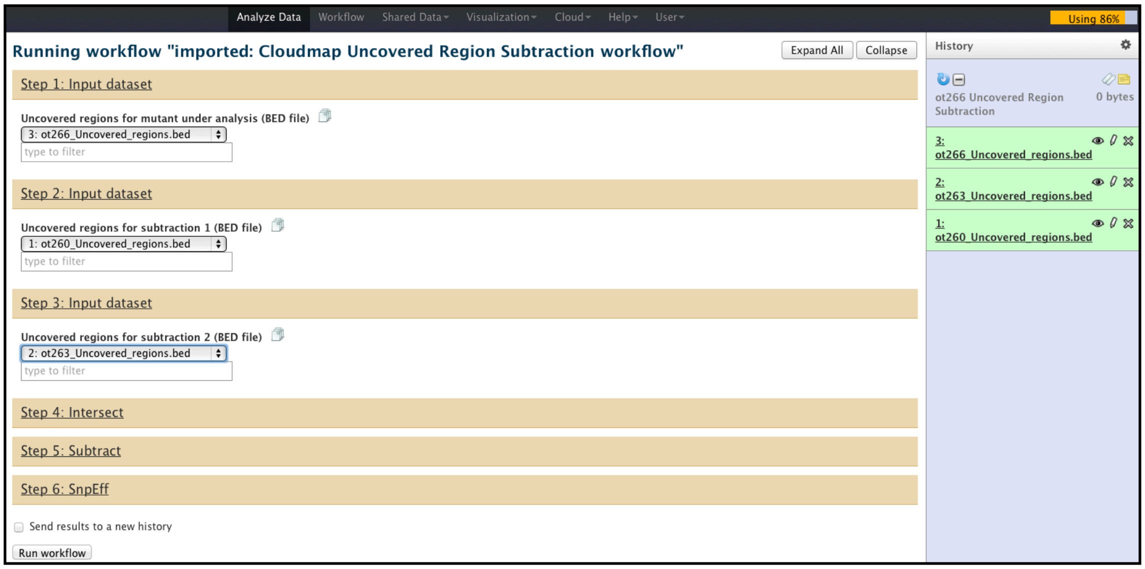
Name
imported: Cloudmap Uncovered Region Subtract
Cloudmap Uncovered Region Subtraction workfl
CloudMap SNP Mapping with WGS Data and Vari
CloudMap Unmapped Mutant workflow
CloudMap Subtract Variants workflow

Workflows shared with you by

No workflows have been shared with you.

Other options

17) You will see all the steps in the workflow prior to running it. Make sure that each of the input fields corresponds to the appropriate file in your history. In our example, we want to subtract uncovered regions present in both *ot260* and *ot263* from the uncovered regions in *ot266*. Click ***Run Workflow*** when ready.



Running workflow "imported: Cloudmap Uncovered Region Subtraction workflow"

Step 1: Input dataset

Uncovered regions for mutant under analysis (BED file): 3: ot266_Uncovered_regions.bed

Step 2: Input dataset

Uncovered regions for subtraction 1 (BED file): 1: ot260_Uncovered_regions.bed

Step 3: Input dataset

Uncovered regions for subtraction 2 (BED file): 2: ot263_Uncovered_regions.bed

Step 4: Intersect

Step 5: Subtract

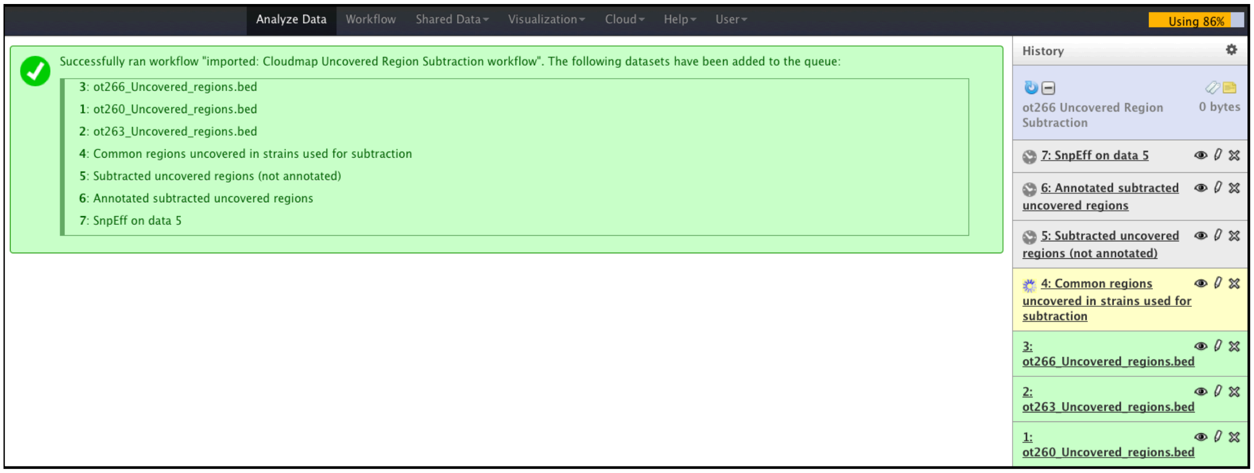
Step 6: SnpEff

□ Send results to a new history

History: ot266 Uncovered Region Subtraction

3: ot266_Uncovered_regions.bed
2: ot263_Uncovered_regions.bed
1: ot260_Uncovered_regions.bed

18) All of the automated functions have the appropriate default parameters configured, although experienced users may want to modify these prior to running. Once you are ready to run the workflow, press ***Run Workflow*** and the workflow will start (this step takes a minute or two to begin, be patient and don't hit the ***Run Workflow*** button repeatedly). You will receive an email when the workflow is completed:



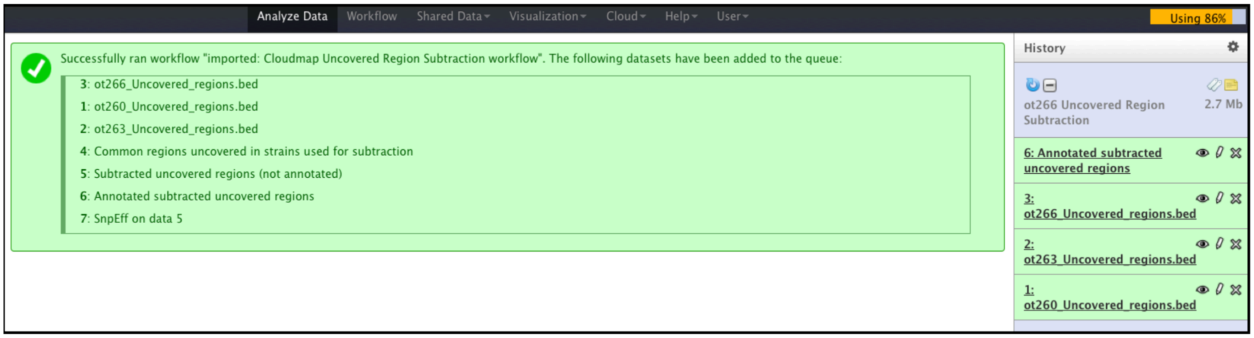
Successfully ran workflow "imported: Cloudmap Uncovered Region Subtraction workflow". The following datasets have been added to the queue:

3: ot266_Uncovered_regions.bed
1: ot260_Uncovered_regions.bed
2: ot263_Uncovered_regions.bed
4: Common regions uncovered in strains used for subtraction
5: Subtracted uncovered regions (not annotated)
6: Annotated subtracted uncovered regions
7: SnpEff on data 5

ot266 Uncovered Region Subtraction 0 bytes

7: SnpEff on data 5
6: Annotated subtracted uncovered regions
5: Subtracted uncovered regions (not annotated)
4: Common regions uncovered in strains used for subtraction
3: ot266_Uncovered_regions.bed
2: ot263_Uncovered_regions.bed
1: ot260_Uncovered_regions.bed

19) The workflow has finished running and you can view the resulting output:



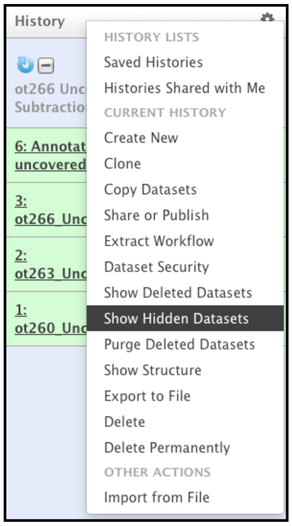
Successfully ran workflow "imported: Cloudmap Uncovered Region Subtraction workflow". The following datasets have been added to the queue:

3: ot266_Uncovered_regions.bed
1: ot260_Uncovered_regions.bed
2: ot263_Uncovered_regions.bed
4: Common regions uncovered in strains used for subtraction
5: Subtracted uncovered regions (not annotated)
6: Annotated subtracted uncovered regions
7: SnpEff on data 5

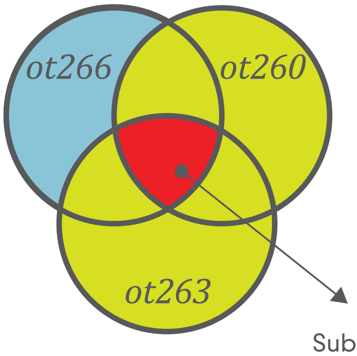
ot266 Uncovered Region Subtraction 2.7 Mb

6: Annotated subtracted uncovered regions
3: ot266_Uncovered_regions.bed
2: ot263_Uncovered_regions.bed
1: ot260_Uncovered_regions.bed

20) You will notice that while 4 output files were generated during the course of the workflow (output files are sequentially numbered), only one output file remains visible while others are hidden. The one visible file (***Annotated subtracted uncovered regions***) is the most important for analysis of the mutant under consideration. In order to view hidden files, click ***Show Hidden Datasets*** in the History menu:

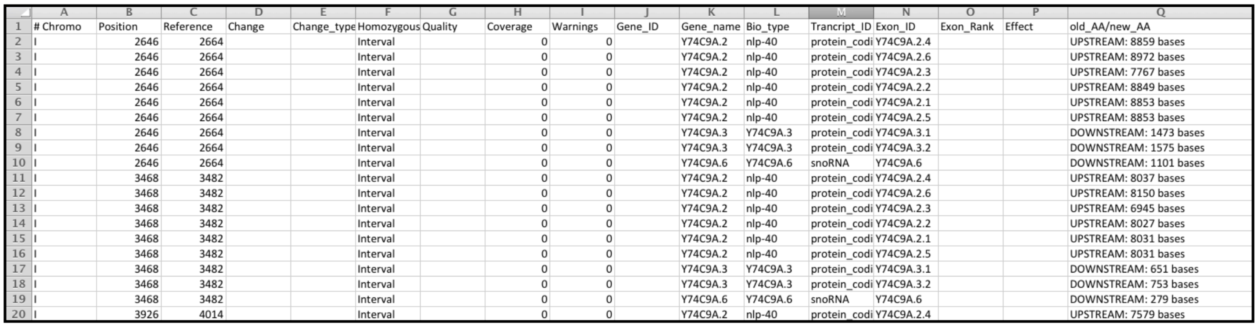


21) The ***Annotated subtracted uncovered regions*** output file conceptually corresponds to the ***Annotated subtracted variants (conservative, only variants present in both subtraction strains removed)*** file generated by the ***Subtract Variants*** workflow. This conservative strategy, as shown below, aims to only subtract uncovered regions that are present in both *ot260* and *ot263*. By selecting uncovered regions that only appear in more than one sample, we hope to err on the side of subtracting true deletions as opposed to subtracting regions that are simply uncovered in a given sample.



Subtract uncovered regions present in both ot260 and ot263

22) The ***Annotated subtracted uncovered regions*** output file (snpEff) is shown below:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	# Chromo	Position	Reference	Change	Change_type	Homozygous	Quality	Coverage	Warnings	Gene_ID	Gene_name	Bio_type	Trancript_ID	Exon_ID	Exon_Rank	Effect	old_AA/new_AA
2	I	2646	2664			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.4			UPSTREAM: 8859 bases
3	I	2646	2664			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.6			UPSTREAM: 8972 bases
4	I	2646	2664			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.3			UPSTREAM: 7767 bases
5	I	2646	2664			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.2			UPSTREAM: 8849 bases
6	I	2646	2664			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.1			UPSTREAM: 8853 bases
7	I	2646	2664			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.5			UPSTREAM: 8853 bases
8	I	2646	2664			Interval		0	0		Y74C9A.3	Y74C9A.3	protein_codi	Y74C9A.3.1			DOWNSTREAM: 1473 bases
9	I	2646	2664			Interval		0	0		Y74C9A.3	Y74C9A.3	protein_codi	Y74C9A.3.2			DOWNSTREAM: 1575 bases
10	I	2646	2664			Interval		0	0		Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6			DOWNSTREAM: 1101 bases
11	I	3468	3482			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.4			UPSTREAM: 8037 bases
12	I	3468	3482			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.6			UPSTREAM: 8150 bases
13	I	3468	3482			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.3			UPSTREAM: 6945 bases
14	I	3468	3482			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.2			UPSTREAM: 8027 bases
15	I	3468	3482			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.1			UPSTREAM: 8031 bases
16	I	3468	3482			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.5			UPSTREAM: 8031 bases
17	I	3468	3482			Interval		0	0		Y74C9A.3	Y74C9A.3	protein_codi	Y74C9A.3.1			DOWNSTREAM: 651 bases
18	I	3468	3482			Interval		0	0		Y74C9A.3	Y74C9A.3	protein_codi	Y74C9A.3.2			DOWNSTREAM: 753 bases
19	I	3468	3482			Interval		0	0		Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6			DOWNSTREAM: 279 bases
20	I	3926	4014			Interval		0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.4			UPSTREAM: 7579 bases

Analyzing your own data with CloudMap and Galaxy:

The various sections of this user guide detail how to analyze sample datasets from the CloudMap paper. In order to analyze your own sequencing data (in the form of FASTQ files), a few quick steps need to be performed prior to running the workflows detailed in this user guide.

For more details, please see the CloudMap paper or visit the CloudMap website at: http://usegalaxy.org/cloudmap. Video versions of these user guides are also available at this website.

Useful Galaxy screencasts are available here: http://wiki.g2.bx.psu.edu/Learn/Screencasts

SECTIONS OF THIS DOCUMENT:

1) UPLOADING FASTQ FILES (or any other type of file)

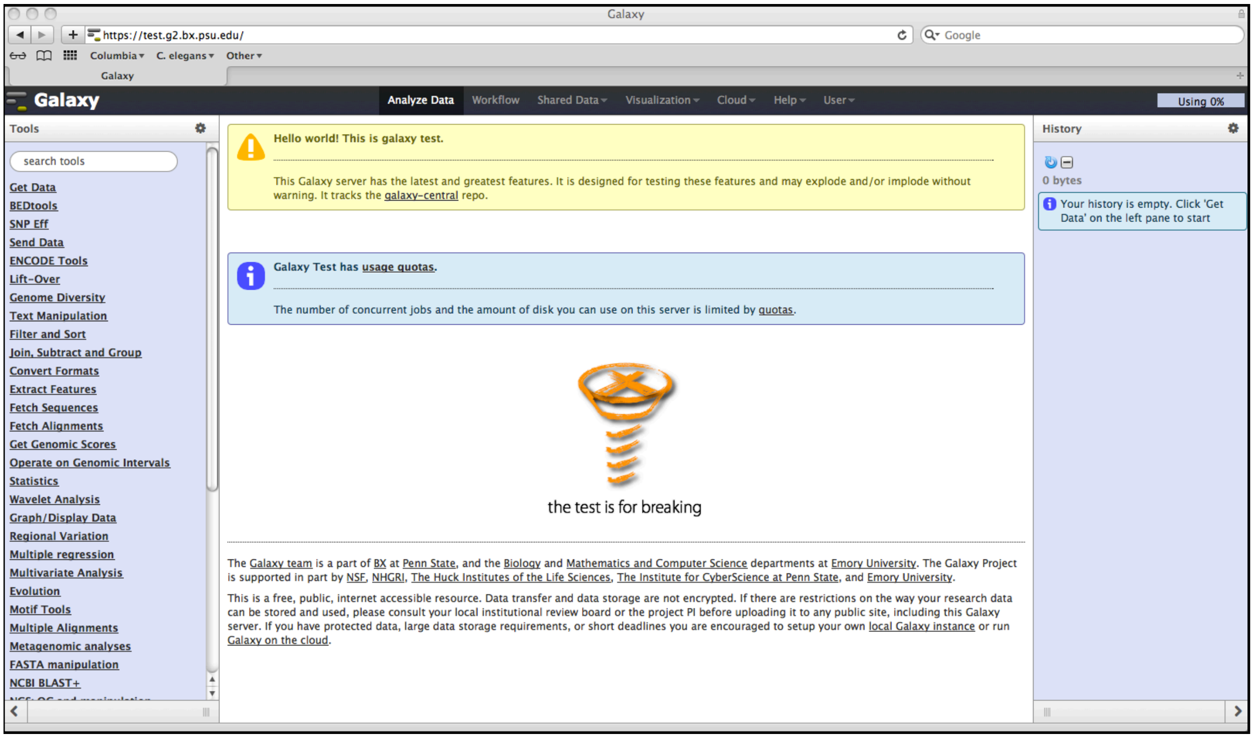
2) CONCATENATING FILES

3) MODIFYING WORKFLOWS & CHANGING TOOL PARAMETERS (single-end vs paired-end data as an example):

4) CONFIGURING THE *SNP MAPPING WITH WGS DATA* WORKFLOW TO SUPPORT SPECIES OTHER THAN *C.ELEGANS* AND *ARABIDOPSIS*:

UPLOADING FASTQ FILES (or any other type of file):

1) Navigate to the Galaxy site (http://usegalaxy.org)



Hello world! This is galaxy test.

This Galaxy server has the latest and greatest features. It is designed for testing these features and may explode and/or implode without warning. It tracks the galaxy-central repo.

Galaxy Test has usage quotas.

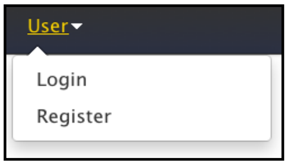
The number of concurrent jobs and the amount of disk you can use on this server is limited by quotas.

The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

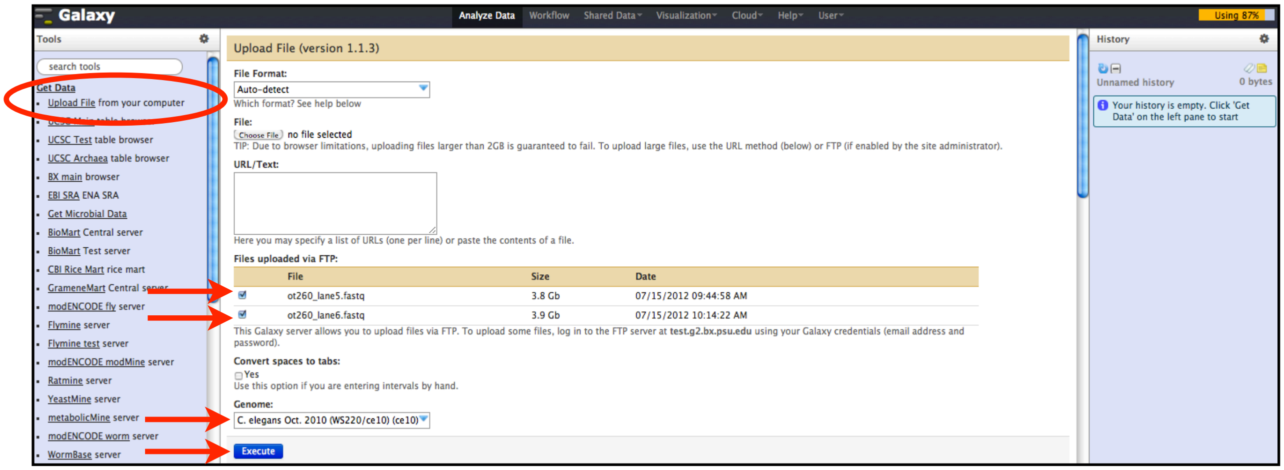
This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own local Galaxy instance or run Galaxy on the cloud.

Your history is empty. Click 'Get Data' on the left pane to start

2) Register for an account or login if you already have an account:



3) Once you are logged in using your email address, click on the ***Get Data*** link in the tools section on the left side of the screen. If the file you want to upload is < 2Gb, you can select the file through the ***Choose file*** link in the browser. Otherwise, you will need to upload your files via FTP (http://wiki.g2.bx.psu.edu/FTPUpload). If you upload your files via FTP, you will see the uploaded files in the ***Upload File*** browser window. Once the files have finished uploading via FTP, select them and the appropriate reference genome (***ce10*** for most of the examples in this user guide) and click ***Execute*** in order to add them to your history.



Upload File (version 1.1.3)

File Format:
Auto-detect
Which format? See help below

File:
no file selected
TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

	File	Size	Date
☑	ot260_lane5.fastq	3.8 Gb	07/15/2012 09:44:58 AM
☑	ot260_lane6.fastq	3.9 Gb	07/15/2012 10:14:22 AM

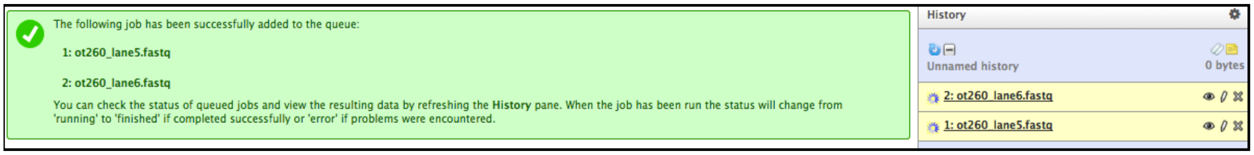
This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at **test.g2.bx.psu.edu** using your Galaxy credentials (email address and password).

Convert spaces to tabs:
☐ Yes
Use this option if you are entering intervals by hand.

Genome:
C. elegans Oct. 2010 (WS220/ce10) (ce10)

Your history is empty. Click 'Get Data' on the left pane to start

4) The files   will be added to your history:



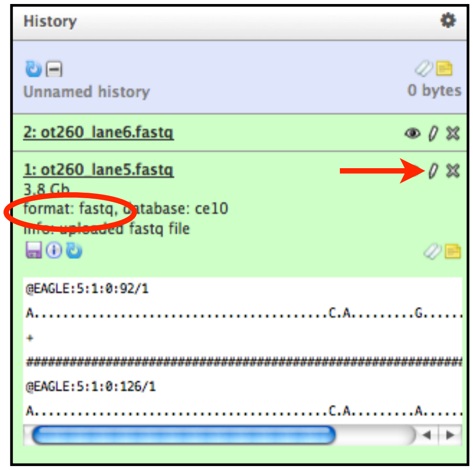
The following job has been successfully added to the queue:

1: ot260_lane5.fastq

2: ot260_lane6.fastq

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

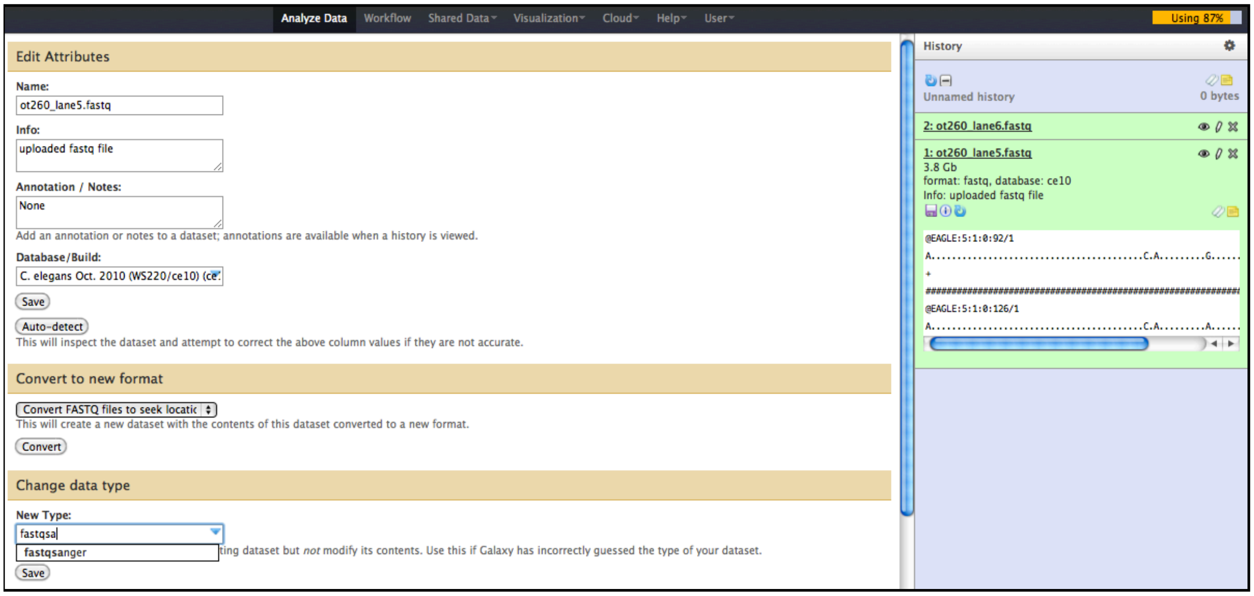
5) Once the FASTQ files are in your history, you will need to specify their data type (i.e. the base quality encoding scheme) by clicking on the file and then on the pencil icon:



3.8 Gb
format: fastq, database: ce10
info: uploaded fastq file

```
@EAGLE:5:1:0:92/1
A.......................................C.A.........G......
+
###########################################################
@EAGLE:5:1:0:126/1
A.......................................C.A.........A......
```

6) The aligners in Galaxy accept the major FASTQ encoding schemes (fastqsanger and illumina) and FASTQ files can be converted from one format to another using the ***FASTQ Groomer*** tool. To read more about FASTQ encoding schemes, see the ***FASTQ Groomer*** tool or http://en.wikipedia.org/wiki/FASTQ_format



Edit Attributes

Name: ot260_lane5.fastq

Info: uploaded fastq file

Annotation / Notes: None

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

Database/Build: C. elegans Oct. 2010 (WS220/ce10) (ce

Save

Auto-detect

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

Convert to new format

Convert FASTQ files to seek locatic

This will create a new dataset with the contents of this dataset converted to a new format.

Convert

Change data type

New Type: fastqsa

fastqsanger

ting dataset but *not* modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.

Save

History

Unnamed history 0 bytes

2: ot260_lane6.fastq

1: ot260_lane5.fastq

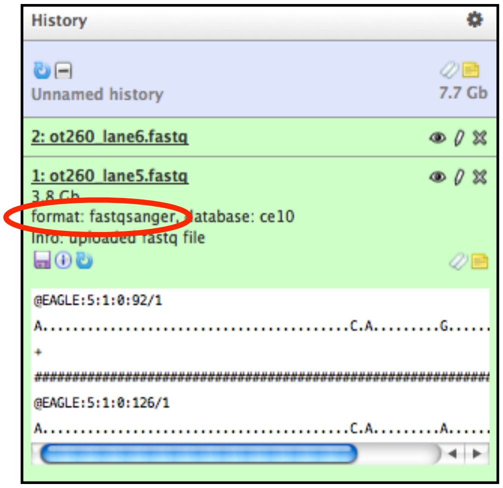
3.8 Gb

format: fastq, database: ce10

Info: uploaded fastq file

```
@EAGLE:5:1:0:92/1
A.......................................C.A.........G......
+
############################################################
@EAGLE:5:1:0:126/1
A.......................................C.A.........A......
```

7) Your FASTQ file will now reflect the change. You can now proceed to import the various reference and configuration files required for the CloudMap workflows detailed elsewhere in this user guide.



History

Unnamed history 7.7 Gb

2: ot260_lane6.fastq

1: ot260_lane5.fastq

3.8 Gb

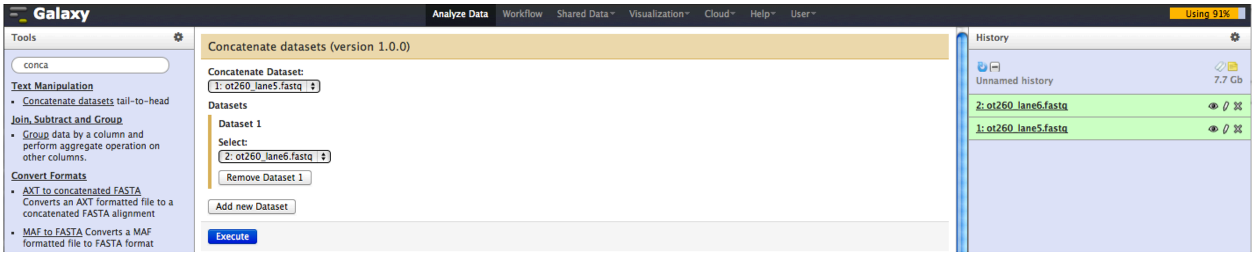
format: fastqsanger, database: ce10

Info: uploaded fastq file

```
@EAGLE:5:1:0:92/1
A.......................................C.A.........G......
+
############################################################
@EAGLE:5:1:0:126/1
A.......................................C.A.........A......
```

CONCATENATING MULTIPLE FILES:

On occasion, your sample may be split up among multiple FASTQ files. In this case, you will need to concatenate your FASTQ files using the Galaxy ***Concatenate datasets*** tool:



conca

Text Manipulation

- Concatenate datasets tail-to-head

Join, Subtract and Group

- Group data by a column and perform aggregate operation on other columns.

Convert Formats

- AXT to concatenated FASTA Converts an AXT formatted file to a concatenated FASTA alignment
- MAF to FASTA Converts a MAF formatted file to FASTA format

Concatenate datasets (version 1.0.0)

Concatenate Dataset: 1: ot260_lane5.fastq

Datasets

Dataset 1

Select: 2: ot260_lane6.fastq

Unnamed history 7.7 Gb

2: ot260_lane6.fastq

1: ot260_lane5.fastq

You can now proceed to import the various reference and configuration files required for the CloudMap workflows detailed in this user guide.

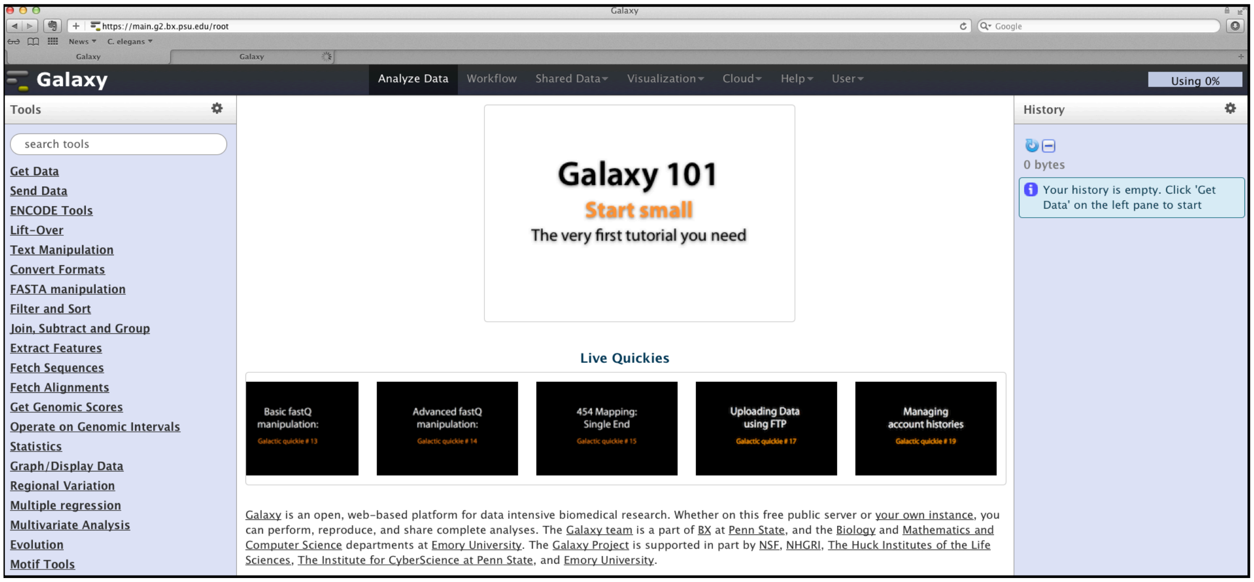
MODIFYING WORKFLOWS & CHANGING TOOL PARAMETERS (single-end vs paired-end data as an example):

The CloudMap workflows discussed in this user guide primarily describe how to run the ***ot266 Proof of principle***. However, these workflows can easily be edited to run any appropriate dataset. Here we will show you how to edit the ***CloudMap Hawaiian Variant Mapping with WGS Data and Variant Calling*** workflow to accept paired-end FASTQ data instead of single-end data. You can edit workflows to change parameters for each tool or to add new tools to your workflows.

Useful workflow-related screencasts from Galaxy are available here:

Create workflow from a history
Create workflow from scratch
Import workflow
Edit workflow
Convert workflow in a tool

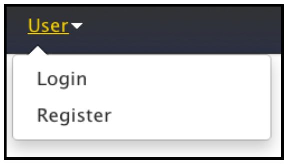
1) Let's assume that you haven't yet imported any CloudMap workflows. Navigate to http://usegalaxy.org/



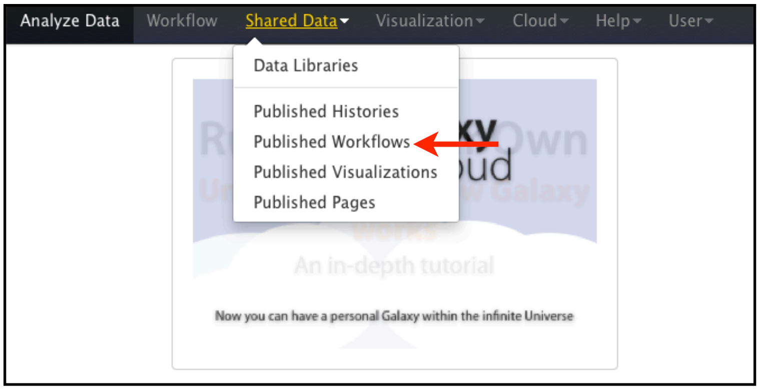
Your history is empty. Click 'Get Data' on the left pane to start

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

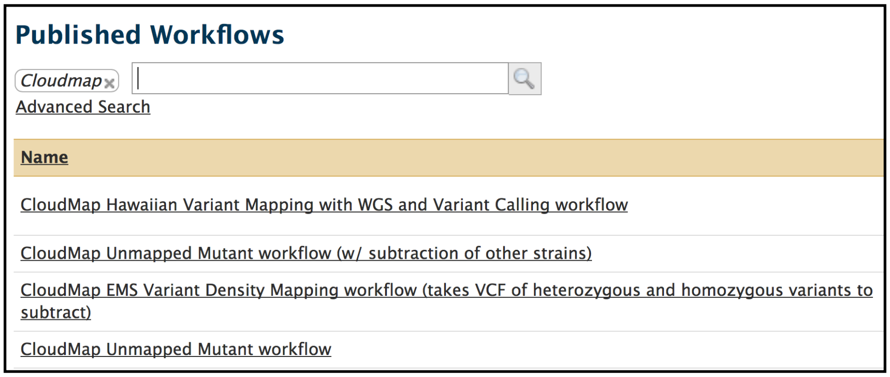
2) Register for an account or login if you already have an account:



3) Click on the ***Shared Data*** link at the top of the page:



4) Click ***Published Workflows*** on the menu bar to access the automated workflow. Select the ***CloudMap Hawaiian Variant Mapping with WGS Data and Variant Calling workflow***.



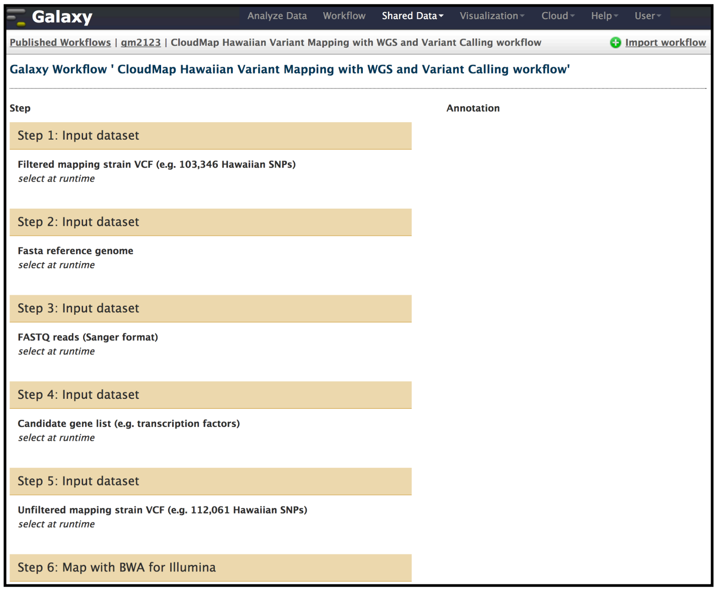
Published Workflows

Cloudmap ×

Advanced Search

Name
CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow
CloudMap Unmapped Mutant workflow (w/ subtraction of other strains)
CloudMap EMS Variant Density Mapping workflow (takes VCF of heterozygous and homozygous variants to subtract)
CloudMap Unmapped Mutant workflow

5) You will now have the option to ***Import workflow***

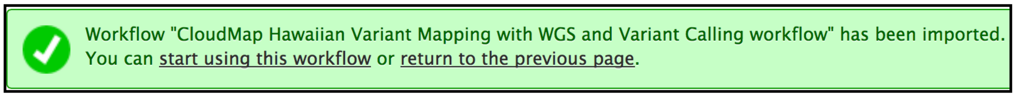


Published Workflows | gm2123 | CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow

Galaxy Workflow ' CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow'

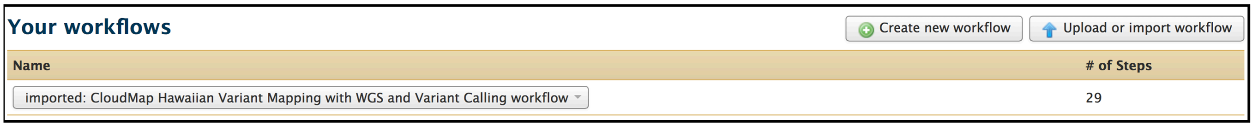
Step	Annotation
Step 1: Input dataset	
Filtered mapping strain VCF (e.g. 103,346 Hawaiian SNPs) *select at runtime*	
Step 2: Input dataset	
Fasta reference genome *select at runtime*	
Step 3: Input dataset	
FASTQ reads (Sanger format) *select at runtime*	
Step 4: Input dataset	
Candidate gene list (e.g. transcription factors) *select at runtime*	
Step 5: Input dataset	
Unfiltered mapping strain VCF (e.g. 112,061 Hawaiian SNPs) *select at runtime*	
Step 6: Map with BWA for Illumina	

6) You will see this message:



Workflow "CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow" has been imported.
You can start using this workflow or return to the previous page.

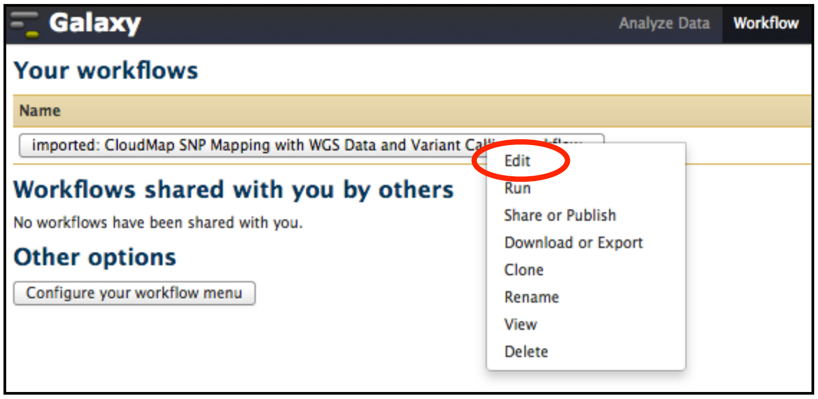
7) Click ***Start using this workflow*** and you will see that the workflow has been imported. From now on, you can easily access this workflow under the ***Workflow*** tab or in the Galaxy tools section (left frame of the browser window) under ***Workflows*** .



Your workflows

Name	# of Steps
imported: CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow	29

8) Click on the workflow and select ***Edit***:



Your workflows

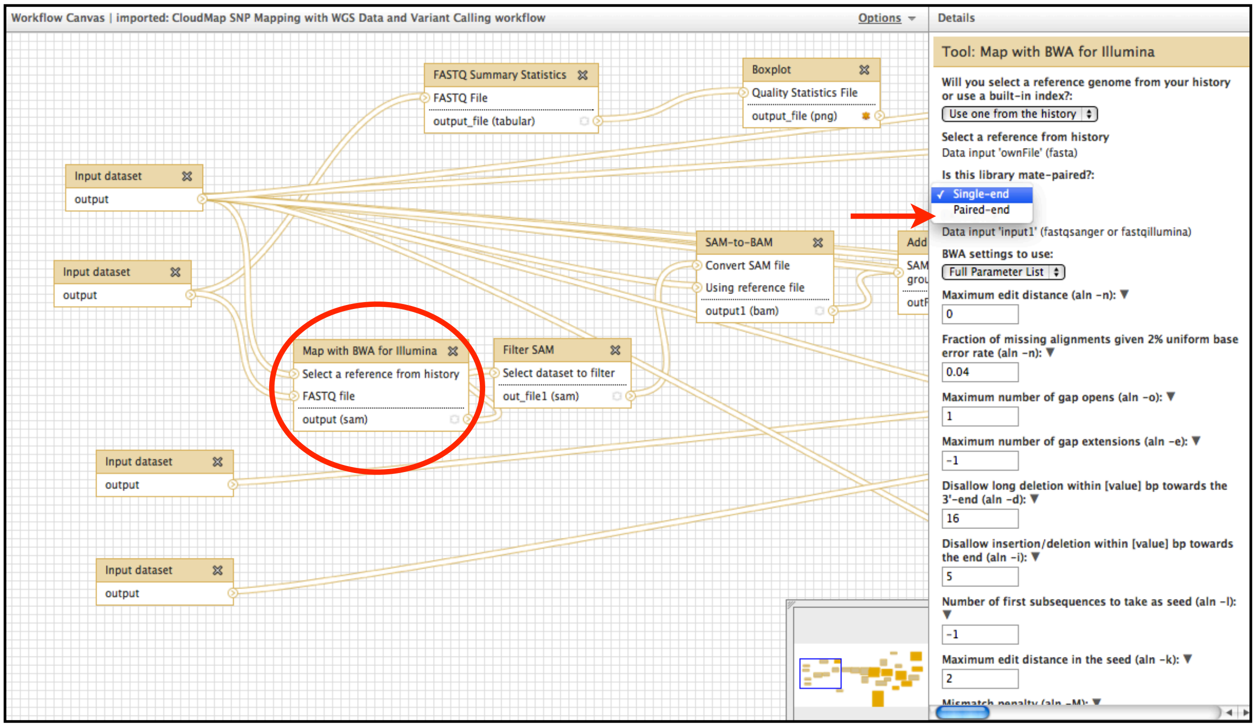
Name
imported: CloudMap SNP Mapping with WGS Data and Variant Calli...

Workflows shared with you by others

No workflows have been shared with you.

Other options

9) You will now see the workflow canvas that displays all the tools and input datasets in the workflow. By clicking on a given tool, you can change its parameters in the right frame of your browser window. We want to change the BWA mapping tool to accept paired-end data so we select the mapping tool and change the data input to paired-end:



Tool: Map with BWA for Illumina

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
Data input 'ownFile' (fasta)

Is this library mate-paired?: ✓ Single-end / Paired-end

Data input 'input1' (fastqsanger or fastqillumina)

BWA settings to use: Full Parameter List

Maximum edit distance (aln -n): 0

Fraction of missing alignments given 2% uniform base error rate (aln -n): 0.04

Maximum number of gap opens (aln -o): 1

Maximum number of gap extensions (aln -e): -1

Disallow long deletion within [value] bp towards the 3'-end (aln -d): 16

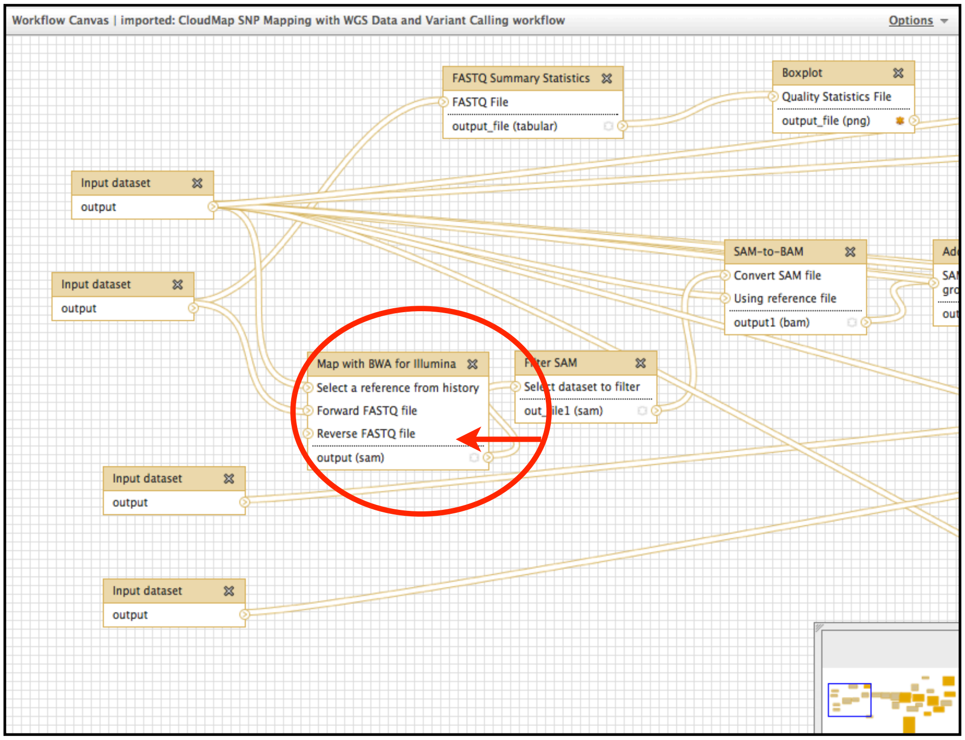
Disallow insertion/deletion within [value] bp towards the end (aln -i): 5

Number of first subsequences to take as seed (aln -l): -1

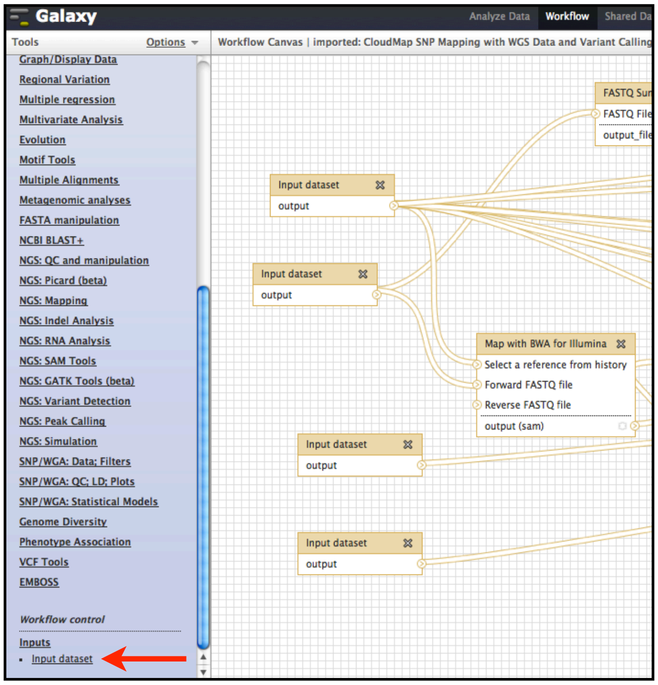
Maximum edit distance in the seed (aln -k): 2

Mismatch penalty (aln -M):

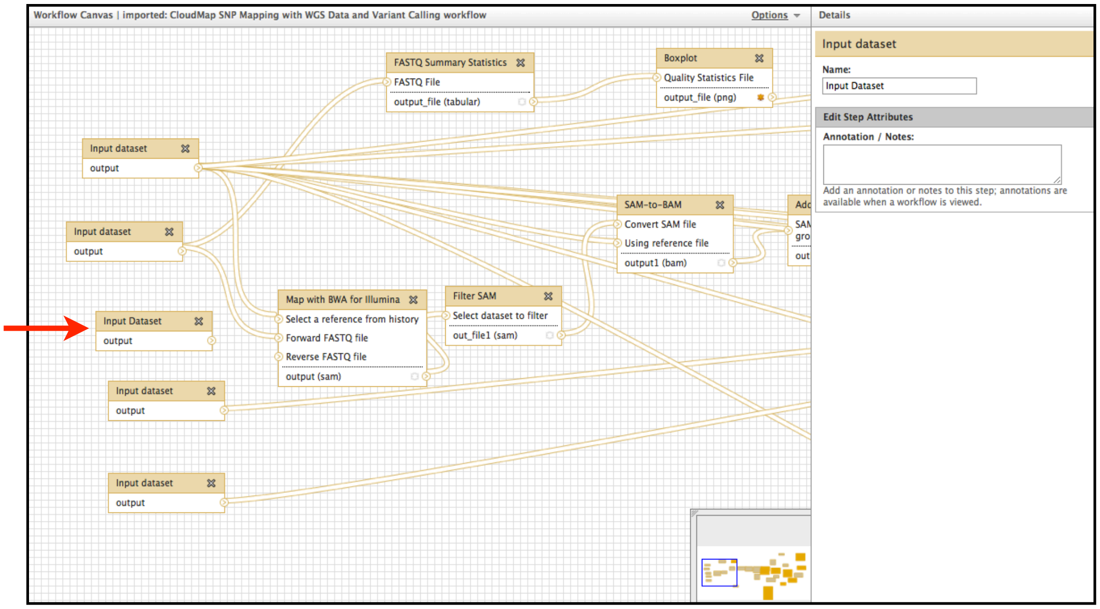
10) Once you select ***paired-end*** as the data type, the BWA mapping tool will now expect another input dataset.



11) To add another input dataset, click ***input dataset*** under Galaxy tools:



12) A new input dataset will appear in your workflow canvas. Attach the input dataset to the arrow next to ***Reverse FASTQ file*** in the ***Map with BWA for Illumina*** tool. If you don't have Illumina data, you can swap out the ***MAP with BWA for Illumina*** tool with one of the other aligners available within Galaxy. Make sure you give a name to your input dataset so you will know what data from your history should be matched to the input when you run the workflow:



Workflow Canvas | imported: CloudMap SNP Mapping with WGS Data and Variant Calling workflow

Details

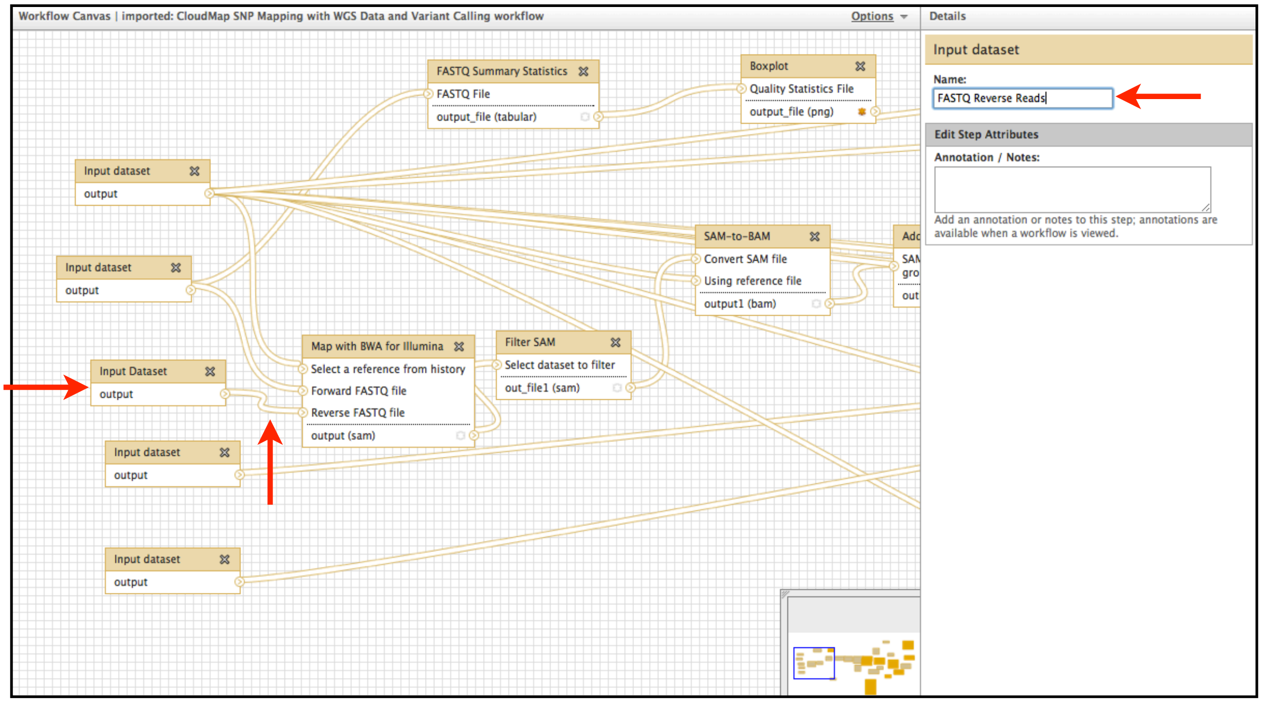
Input dataset

Name: Input Dataset

Edit Step Attributes

Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.



Workflow Canvas | imported: CloudMap SNP Mapping with WGS Data and Variant Calling workflow

Details

Input dataset

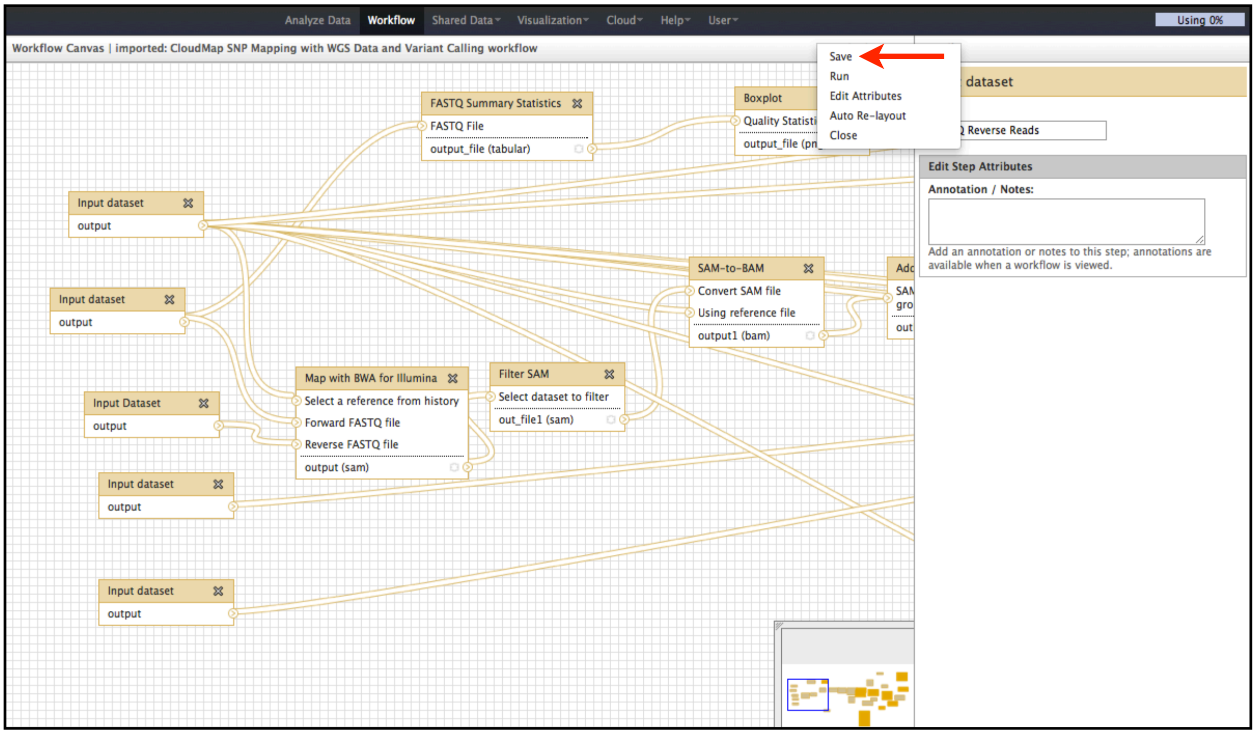
Name: FASTQ Reverse Reads

Edit Step Attributes

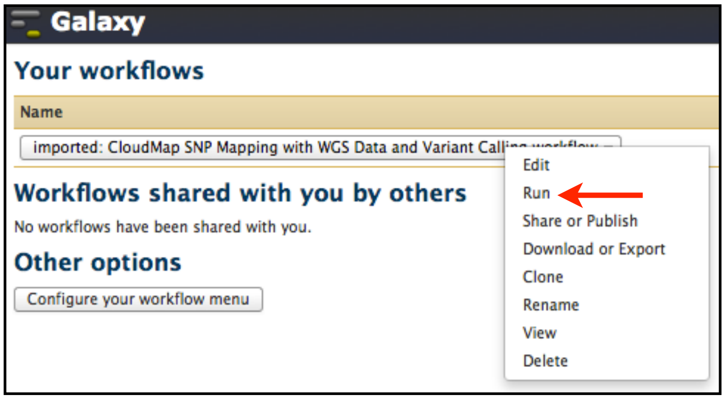
Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

13) Now ***save*** the workflow and ***close***.



14) You can now run the modified workflow:


Galaxy
Your workflows
Name
imported: CloudMap SNP Mapping with WGS Data and Variant Calling workflow
Edit
Run
Share or Publish
Download or Export
Clone
Rename
View
Delete
Workflows shared with you by others
No workflows have been shared with you.
Other options
Configure your workflow menu

CONFIGURING THE *HAWAIIAN VARIANT MAPPING WITH WGS DATA WORKFLOW* TO SUPPORT SPECIES OTHER THAN *C.ELEGANS* AND *ARABIDOPSIS*:

1) Upload the Fasta reference file for the species you wish to analyze and a configuration file for the ***Hawaiian Variant Mapping with WGS Data*** tool. Refer to the ***UPLOADING FASTQ FILES (or any other type of file)*** section of this user guide for details on how to upload your own data. The configuration file is simply a two column, tab delimited list composed of the chromosome number and length in megabases. The numbering scheme of the chromosome should match that of the FASTA reference used for the analysis. Make sure that the FASTA headers (lines starting with >) contain only the chromosome name in one of the following formats:

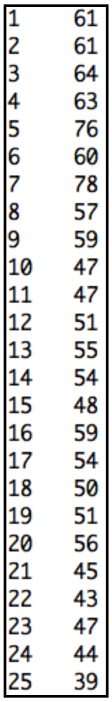
>CHROMOSOME_<number>

>CHROM_<number>

><number>

i.e.:
>CHROMOSOME_1
>CHROM_1
>1

Sample *D.rerio* configuration file:

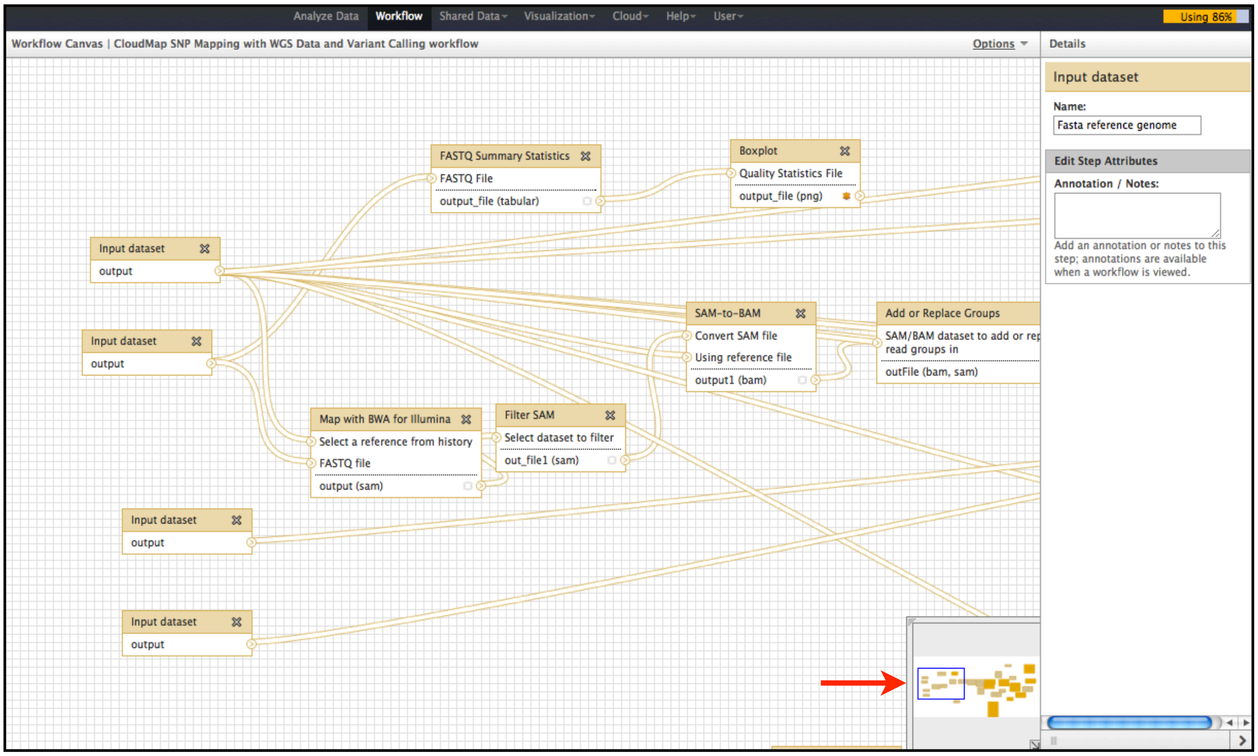


```
1	61
2	61
3	64
4	63
5	76
6	60
7	78
8	57
9	59
10	47
11	47
12	51
13	55
14	54
15	48
16	59
17	54
18	50
19	51
20	56
21	45
22	43
23	47
24	44
25	39
```

Please see more sample ***Other species*** configuration files in the CloudMap data library in the ***Hawaiian Variant Mapping with WGS Data Other Species Config Files*** folder.

2) Now refer to steps 1-8 of the ***MODIFYING WORKFLOWS & CHANGING TOOL PARAMETERS*** section of this user guide to see how to edit the ***Hawaiian Variant Mapping with WGS Data and Variant Calling*** workflow. Step 3 below continues after step 8 of that workflow.

3) You should now see the workflow canvas that displays all the tools and input datasets in the workflow. Scroll across the window displaying all of the tools in the workflow by dragging the small square at the bottom right of your window.



Workflow Canvas | CloudMap SNP Mapping with WGS Data and Variant Calling workflow

Details

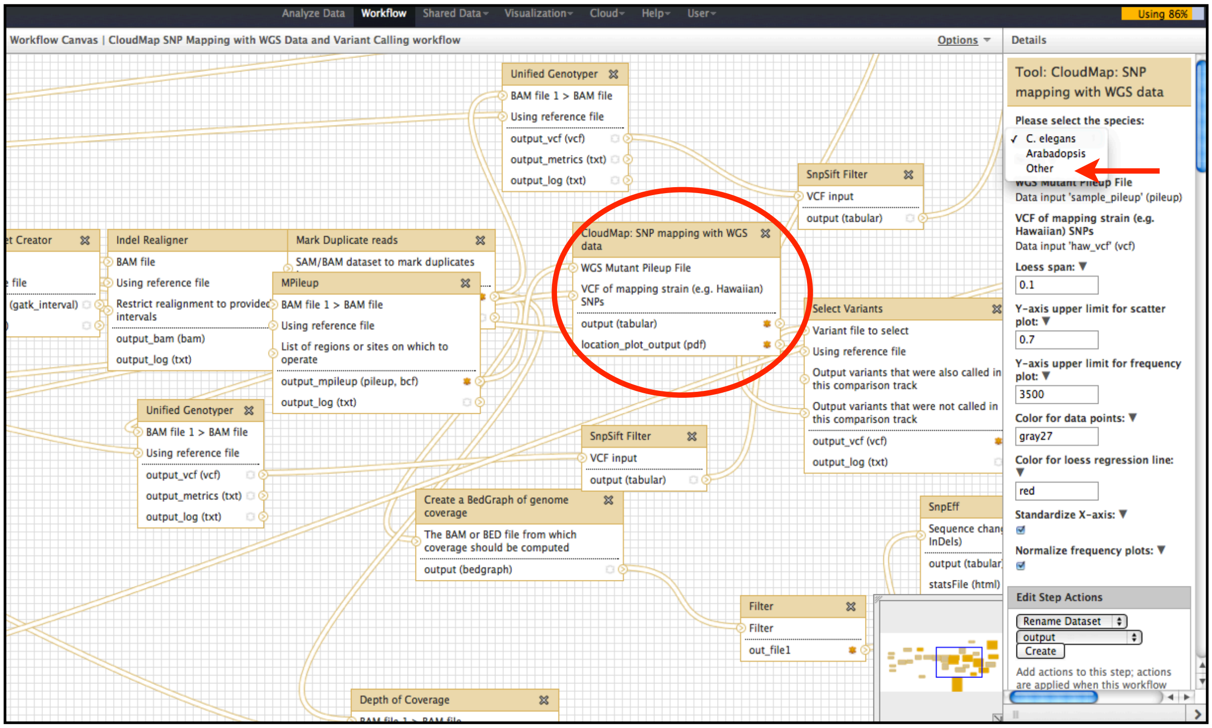
Input dataset

Name: Fasta reference genome

Edit Step Attributes

Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.



Workflow Canvas | CloudMap SNP Mapping with WGS Data and Variant Calling workflow

Details

Tool: CloudMap: SNP mapping with WGS data

Please select the species:
✓ C. elegans
Arabadopsis
Other

WGS Mutant Pileup File
Data input 'sample_pileup' (pileup)

VCF of mapping strain (e.g. Hawaiian) SNPs
Data input 'haw_vcf' (vcf)

Loess span: ▼ 0.1

Y-axis upper limit for scatter plot: ▼ 0.7

Y-axis upper limit for frequency plot: ▼ 3500

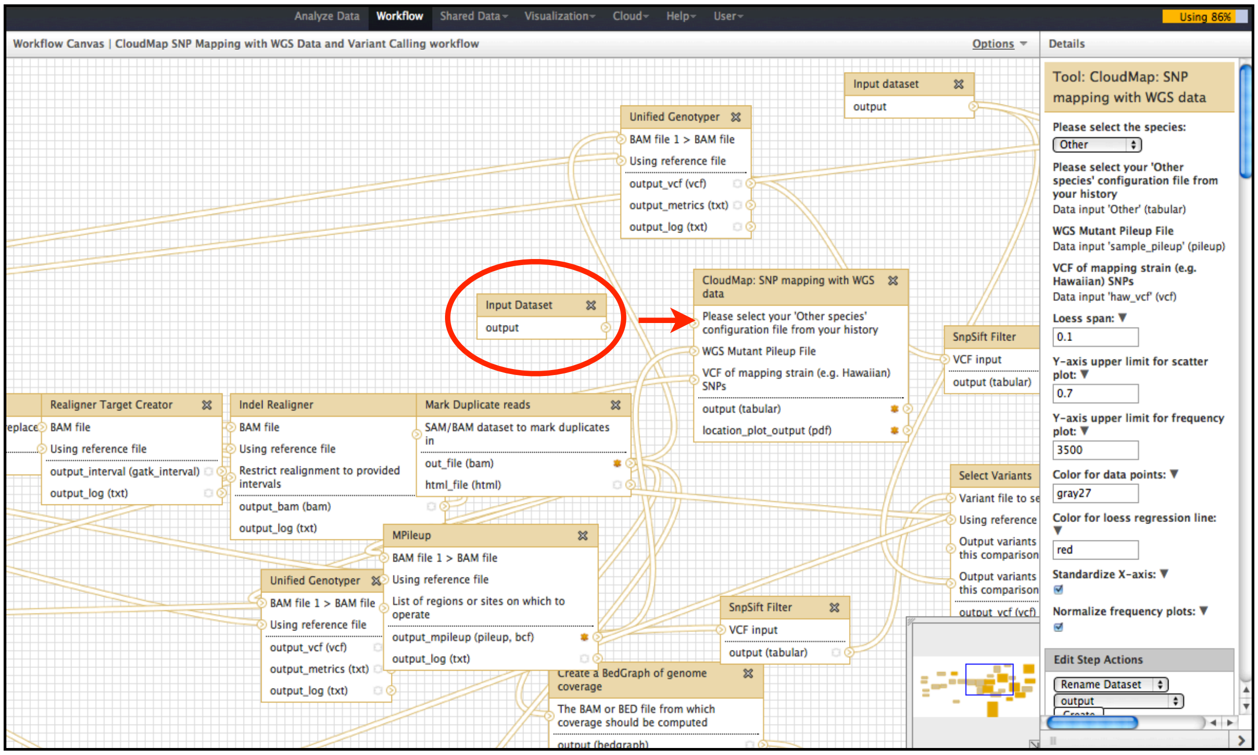
Color for data points: ▼ gray27

Color for loess regression line: ▼ red

Standardize X-axis: ▼ ☑

Normalize frequency plots: ▼ ☑

4) Select the ***CloudMap Hawaiian Variant Mapping with WGS Data*** tool, then select ***Other*** from species list.



Workflow Canvas | CloudMap SNP Mapping with WGS Data and Variant Calling workflow

Details

Tool: CloudMap: SNP mapping with WGS data

Please select the species: Other

Please select your 'Other species' configuration file from your history
Data input 'Other' (tabular)

WGS Mutant Pileup File
Data input 'sample_pileup' (pileup)

VCF of mapping strain (e.g. Hawaiian) SNPs
Data input 'haw_vcf' (vcf)

Loess span: ▼ 0.1

Y-axis upper limit for scatter plot: ▼ 0.7

Y-axis upper limit for frequency plot: ▼ 3500

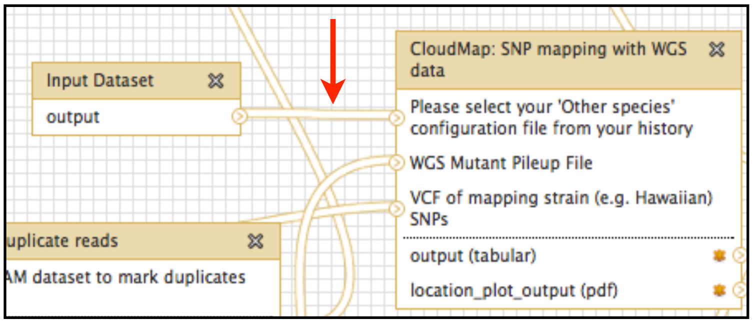
Color for data points: ▼ gray27

Color for loess regression line: ▼ red

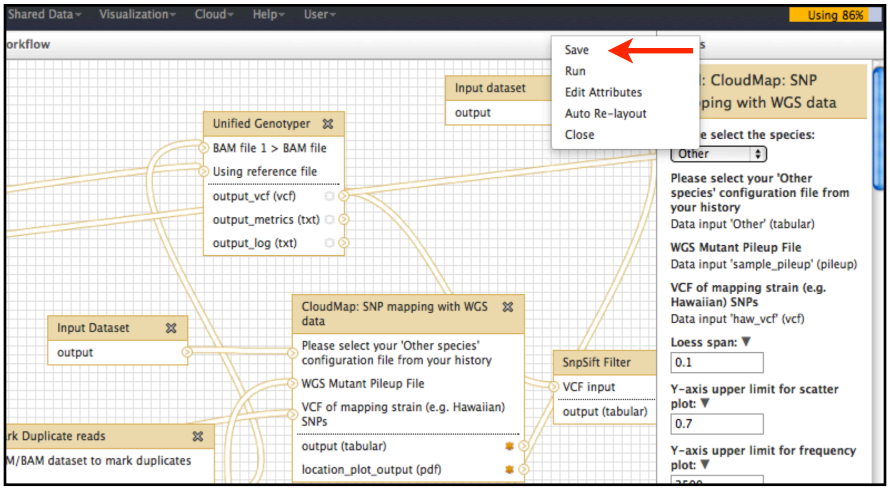
Standardize X-axis: ▼ ☑

Normalize frequency plots: ▼ ☑

6) Connect the ***Other species*** input dataset to the ***CloudMap Hawaiian Variant Mapping with WGS Data*** tool by clicking and dragging the arrow on the side of the Input dataset tool.


Input Dataset
output
uplicate reads
AM dataset to mark duplicates
CloudMap: SNP mapping with WGS data
Please select your 'Other species' configuration file from your history
WGS Mutant Pileup File
VCF of mapping strain (e.g. Hawaiian) SNPs
output (tabular)
location_plot_output (pdf)

7) Now save and close the workflow and you're ready to run it.



e select the species: Other

Please select your 'Other species' configuration file from your history
Data input 'Other' (tabular)

WGS Mutant Pileup File
Data input 'sample_pileup' (pileup)

VCF of mapping strain (e.g. Hawaiian) SNPs
Data input 'haw_vcf' (vcf)

Loess span: 0.1

Y-axis upper limit for scatter plot: 0.7

Y-axis upper limit for frequency plot:

CloudMap | Cloud-based Pipeline for Analysis of Mutant Genome Sequences

This document contains **Frequently Asked Questions** (FAQs) regarding CloudMap and Galaxy. The document will be continually updated. For more details, please see the CloudMap paper or visit the CloudMap website at: http://usegalaxy.org/cloudmap. Video versions of these user guides are available at the CloudMap website.

Your first stop for Galaxy-related FAQs:
http://wiki.g2.bx.psu.edu/Support
http://wiki.g2.bx.psu.edu/Learn/FAQ

http://seqanswers.com/ is a very useful next generation sequencing forum.

FAQs:

Cloudmap questions:

1) My workflow is missing steps mentioned in the user guide, how do I get the latest version?

2) I would like to change some aspect of the plots, how can I do this?

Galaxy questions:

1) My tool turned red after execution and no output file was created. What should I do?

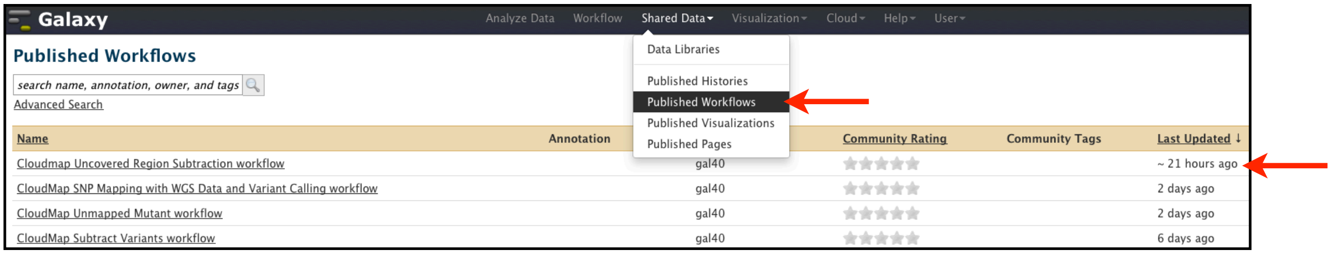
2) I see my data in my history but the tool won't recognize it. What's wrong?

3) I want to use a specific genome build that isn't available in Galaxy. How can I do this?

Cloudmap questions:

My workflow is missing steps mentioned in the user guide, how do I get the latest version?

Make sure you re-import your workflows to get the latest versions. Check under Shared Data —> Published Workflows to see when workflow were last updated.



Published Workflows

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated ↓
Cloudmap Uncovered Region Subtraction workflow		gal40	☆☆☆☆☆		~ 21 hours ago
CloudMap SNP Mapping with WGS Data and Variant Calling workflow		gal40	☆☆☆☆☆		2 days ago
CloudMap Unmapped Mutant workflow		gal40	☆☆☆☆☆		2 days ago
CloudMap Subtract Variants workflow		gal40	☆☆☆☆☆		6 days ago

I would like to change some aspect of the plots, how can I do this?

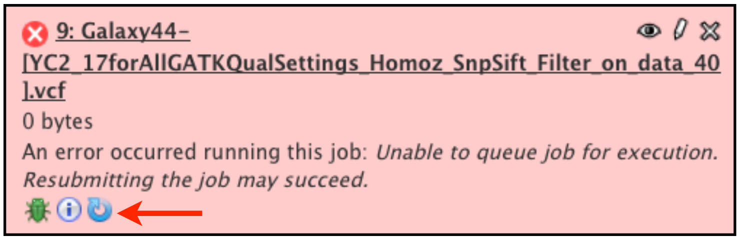
You can email us with your request at gm2123@columbia.edu or or38@columbia.edu. If you want to make the change yourself and run the tool locally, you can download the source code from the Galaxy Tool Shed at: http://toolshed.g2.bx.psu.edu/

Read more about the Galaxy Tool Shed here: http://wiki.g2.bx.psu.edu/Tool%20Shed

Galaxy questions:

My tool turned red after execution and no output file was created. What should I do?

First check that you provided the correct type of input file and settings for the tool. Next try rerunning the tool by clicking the ***run this job again*** arrow.

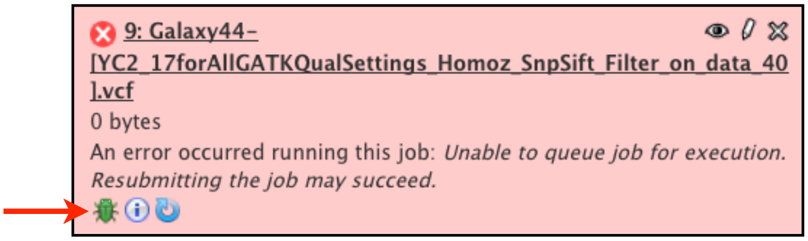


9: Galaxy44-[YC2_17forAllGATKQualSettings_Homoz_SnpSift_Filter_on_data_40].vcf

0 bytes

An error occurred running this job: *Unable to queue job for execution. Resubmitting the job may succeed.*

Failing that, submit a bug report to Galaxy by clicking on the bug icon.



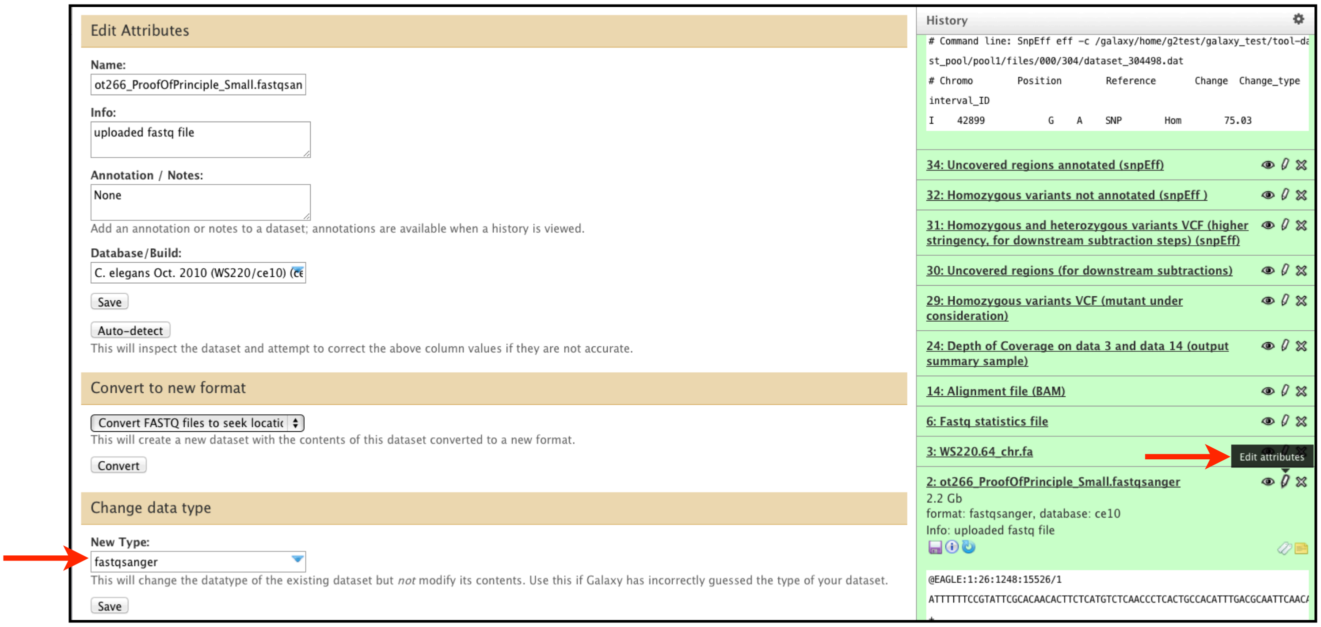
9: Galaxy44-[YC2_17forAllGATKQualSettings_Homoz_SnpSift_Filter_on_data_40].vcf

0 bytes

An error occurred running this job: *Unable to queue job for execution. Resubmitting the job may succeed.*

I see my data in my history but the tool won't recognize it. What's wrong?

This is one of the most common problems users encounter within Galaxy. Use the pencil icon to change the data type to the correct type. http://wiki.g2.bx.psu.edu/Learn/Managing%20Datasets



Edit Attributes

Name: ot266_ProofOfPrinciple_Small.fastqsan

Info: uploaded fastq file

Annotation / Notes: None

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

Database/Build: C. elegans Oct. 2010 (WS220/ce10) (ce

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

Convert to new format

Convert FASTQ files to seek locatic

This will create a new dataset with the contents of this dataset converted to a new format.

Change data type

New Type: fastqsanger

This will change the datatype of the existing dataset but *not* modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.

History

```
# Command line: SnpEff eff -c /galaxy/home/g2test/galaxy_test/tool-da
st_pool/pool1/files/000/304/dataset_304498.dat
# Chromo        Position        Reference       Change  Change_type
interval_ID
I     42899           G     A     SNP       Hom       75.03
```

34: Uncovered regions annotated (snpEff)

32: Homozygous variants not annotated (snpEff)

31: Homozygous and heterozygous variants VCF (higher stringency, for downstream subtraction steps) (snpEff)

30: Uncovered regions (for downstream subtractions)

29: Homozygous variants VCF (mutant under consideration)

24: Depth of Coverage on data 3 and data 14 (output summary sample)

14: Alignment file (BAM)

6: Fastq statistics file

3: WS220.64_chr.fa

2: ot266_ProofOfPrinciple_Small.fastqsanger

2.2 Gb

format: fastqsanger, database: ce10

Info: uploaded fastq file

```
@EAGLE:1:26:1248:15526/1
ATTTTTTCCGTATTCGCACAACACTTCTCATGTCTCAACCCTCACTGCCACATTTGACGCAATTCAAC
```

I want to use a specific genome build that isn't available in Galaxy. How can I do this?

For the vast majority of the tools (BWA, Bowtie aligners especially), you can upload genome reference files (FASTA) and use these for the duration of the history. If you're using a tool that only takes genome builds that are "hard-coded" within Galaxy and you want to support a specific genome, please check the Galaxy support page: http://wiki.g2.bx.psu.edu/Support.

If you plan to use an uploaded FASTA file with the ***Hawaiian Variant Mapping with WGS Data*** tool, make sure that the FASTA headers (lines starting with >) contain only the chromosome name in one of the following formats:

>CHROMOSOME_<number>

>CHROM_<number>

><number>

If you plan to use an uploaded FASTA file with the ***Hawaiian Variant Mapping with WGS Data*** tool and your FASTA file is for a species other than *C.elegans* or *Arabidopsis*, make sure the chromosome naming convention in the ***Other species*** configuration file matches that of the FASTA file. Please see sample ***Other species*** configuration files in the CloudMap data library in the ***Hawaiian Variant Mapping with WGS Data Other Species Config Files*** folder.