

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.

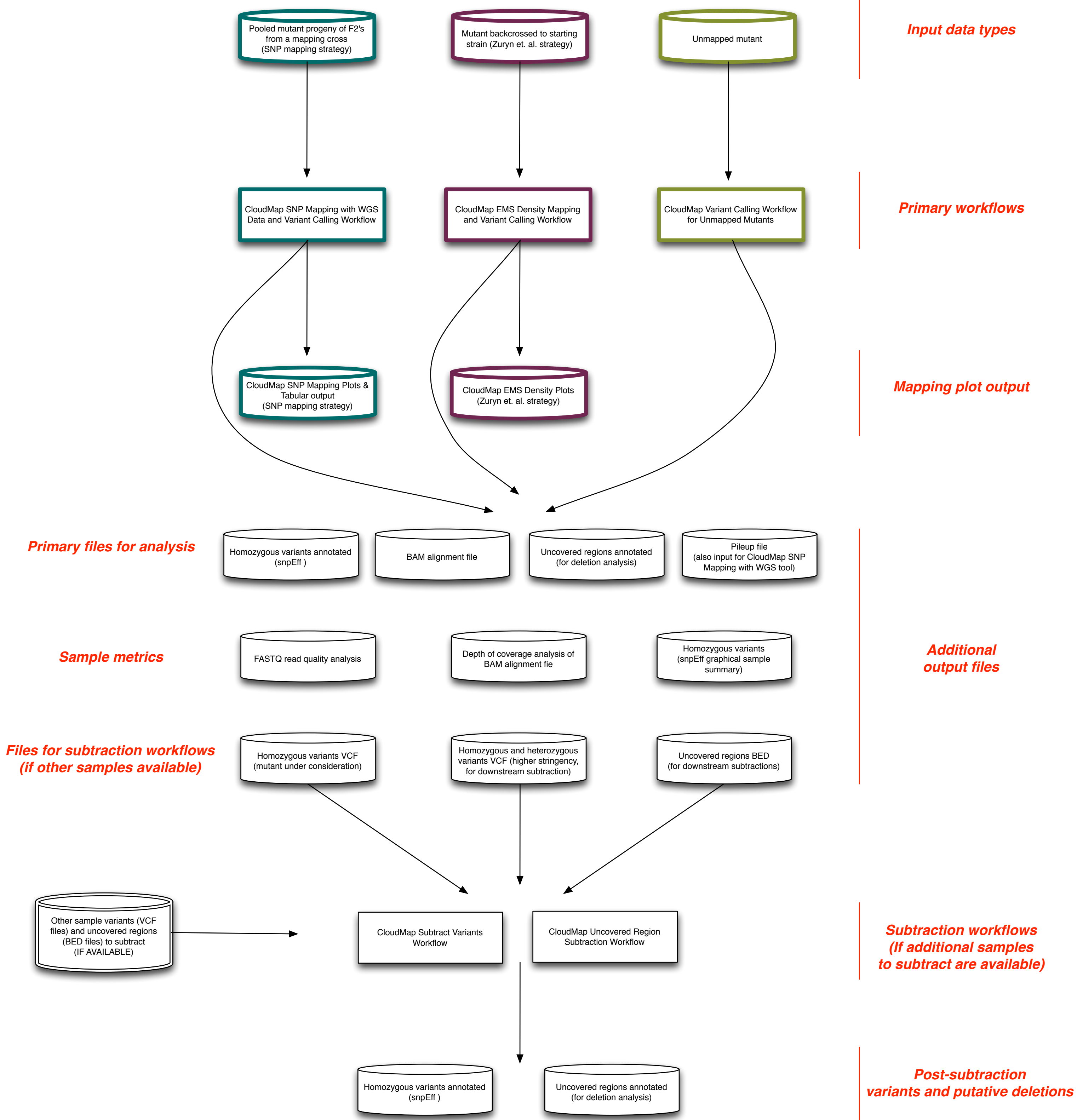
p33 -- **EMS Density Mapping Workflow.** This workflow is essentially the same as the **Unmapped Mutant** workflow followed by the **Subtract Variants** workflow with the addition of an **EMS density** plot of the final VCF variants file.

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



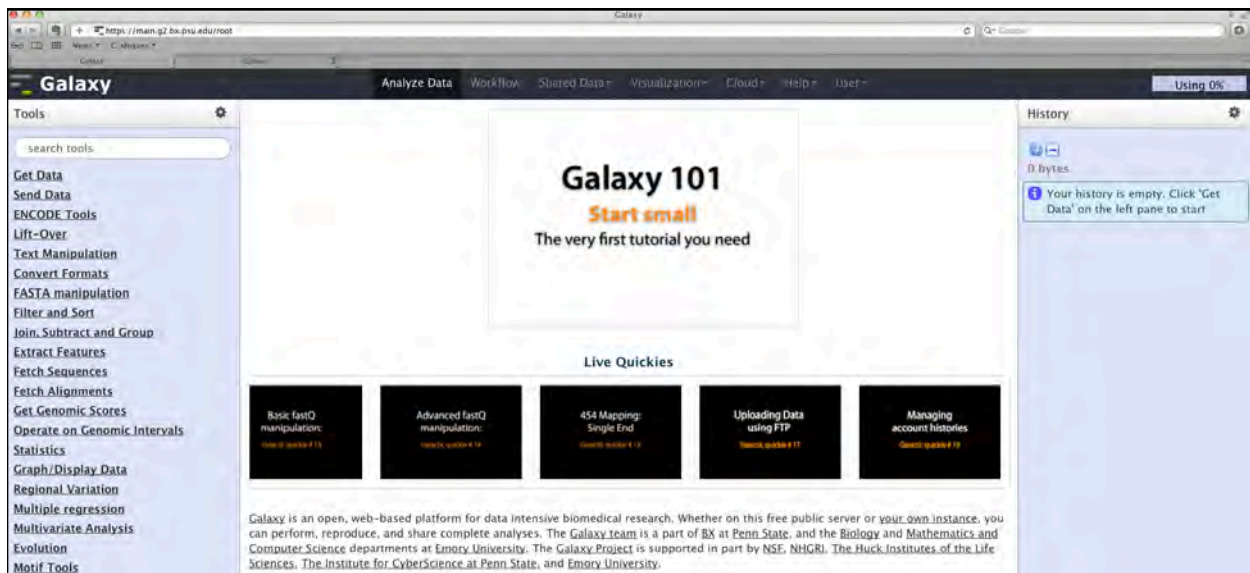
CloudMap Hawaiian Variant Mapping with WGS Data and Variant Calling Workflow (using *ot266* Proof of Principle example from the CloudMap paper). A video version of this user guide is available at: <http://usegalaxy.org/cloudmap>.

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).

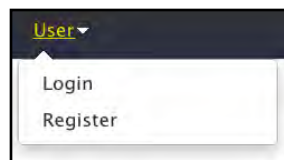
The aim in this user guide is to walk readers through Galaxy-based analysis of the *ot266* mutant using predefined CloudMap workflows which sequentially execute all of the steps required for common mutant analysis functions. This same workflow can be used for analysis of any mutant (from any species)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

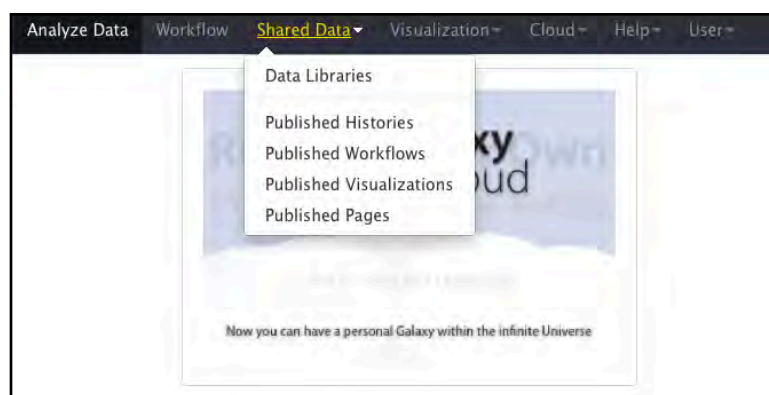
1) Navigate to <http://usegalaxy.org> (URL will resolve to something like <https://main.g2.bx.psu.edu>)



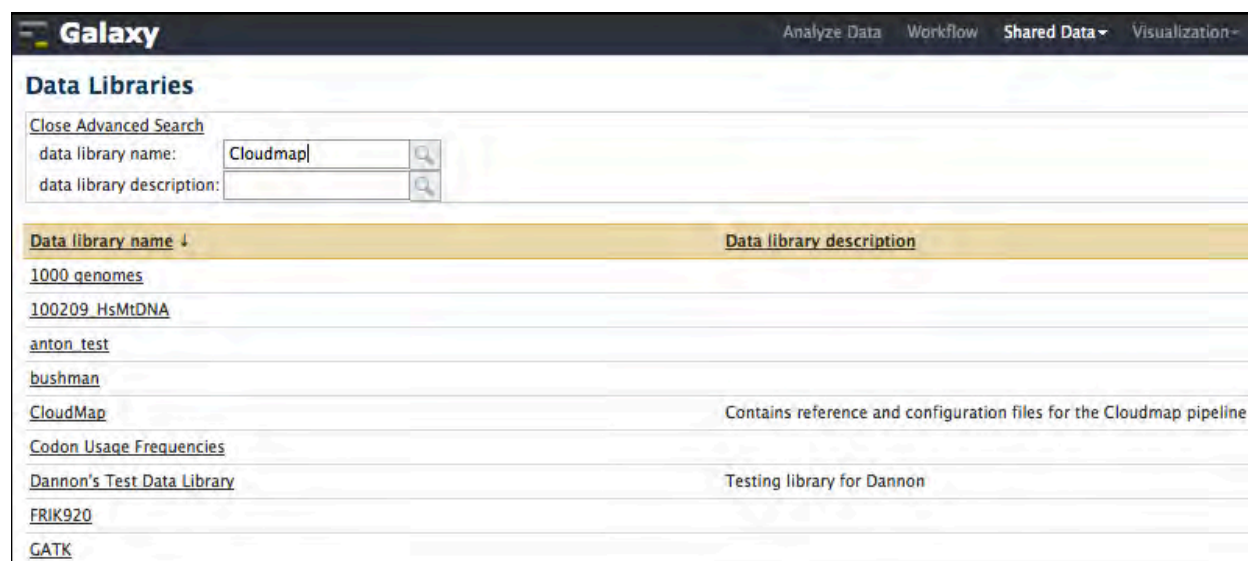
2) Register for an account or login if you already have an account:



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:



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.

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
<input checked="" type="checkbox"/> 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)			
<input checked="" type="checkbox"/> 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)			
<input checked="" type="checkbox"/> HA_SNPs_Unfiltered_112061Variants_WS220.64_chr.vcf		vcf	2012-09-23	4.6 MB
<input checked="" type="checkbox"/> ot266_ProofOfPrinciple_Small.fastqsanger	None	fastqsanger	2012-09-23	2.2 GB
<input checked="" type="checkbox"/> 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:

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:

Analyze Data | Workflow | Shared Data | Admin | Help | User

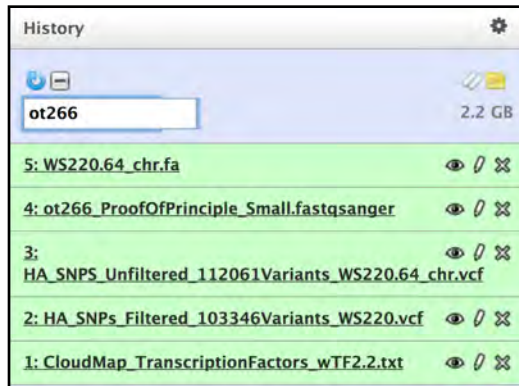
8) You will now see that the data files have been added to an unnamed history:

History

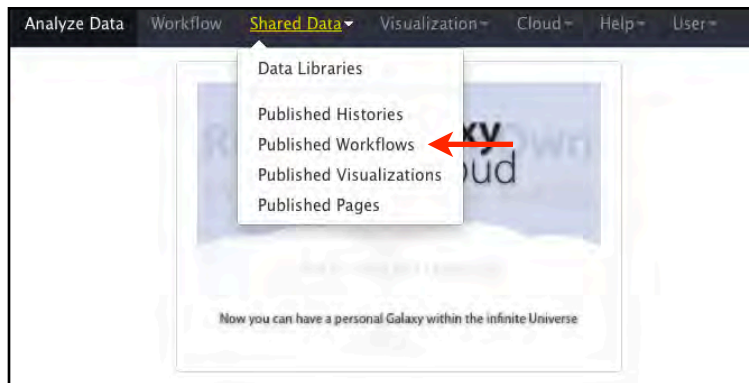
Unnamed history 2.2 GB

- 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

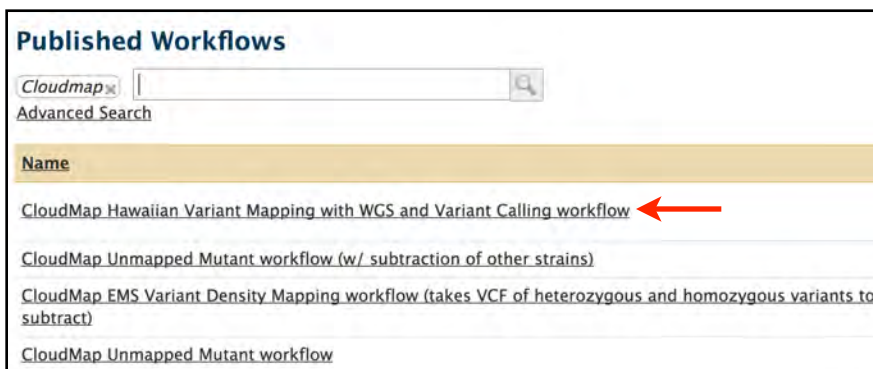
9) Name your history **ot266** after the sample that we will be analyzing:



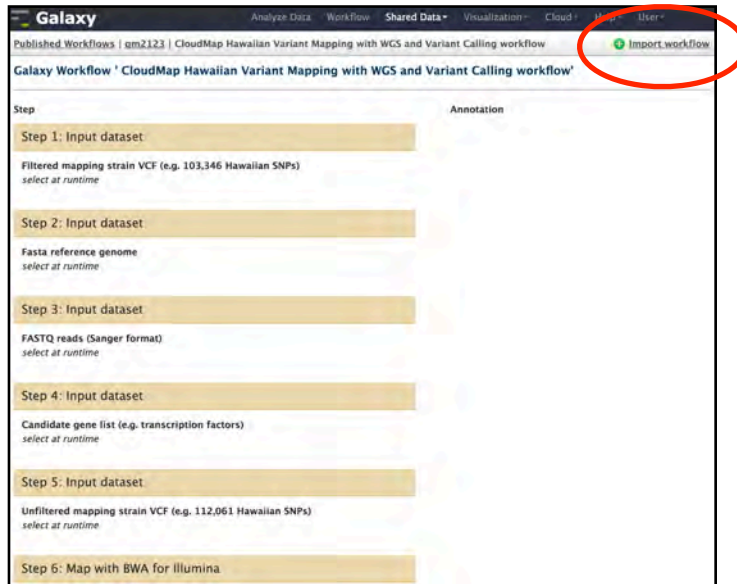
10) Again click on the **Shared Data** link at the top of the page and select **Published Workflows** :



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



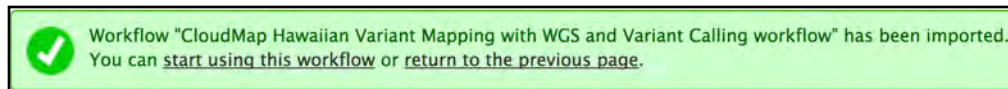
12) You will now have the option to **Import workflow**



The screenshot shows the Galaxy web interface. At the top, there is a navigation bar with tabs: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', and 'User'. The 'Workflow' tab is active. Below the navigation bar, there is a header for the workflow: 'Galaxy Workflow "CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow"'. A green button labeled 'Import workflow' is circled in red. Below the header, there is a list of steps for the workflow:

- 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**.



A green notification box with a checkmark icon on the left. The text inside reads: "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.



The screenshot shows the 'Your workflows' section. At the top right, there are two buttons: 'Create new workflow' and 'Upload or import workflow'. Below this is a table with two columns: 'Name' and '# of Steps'. The table contains one entry:

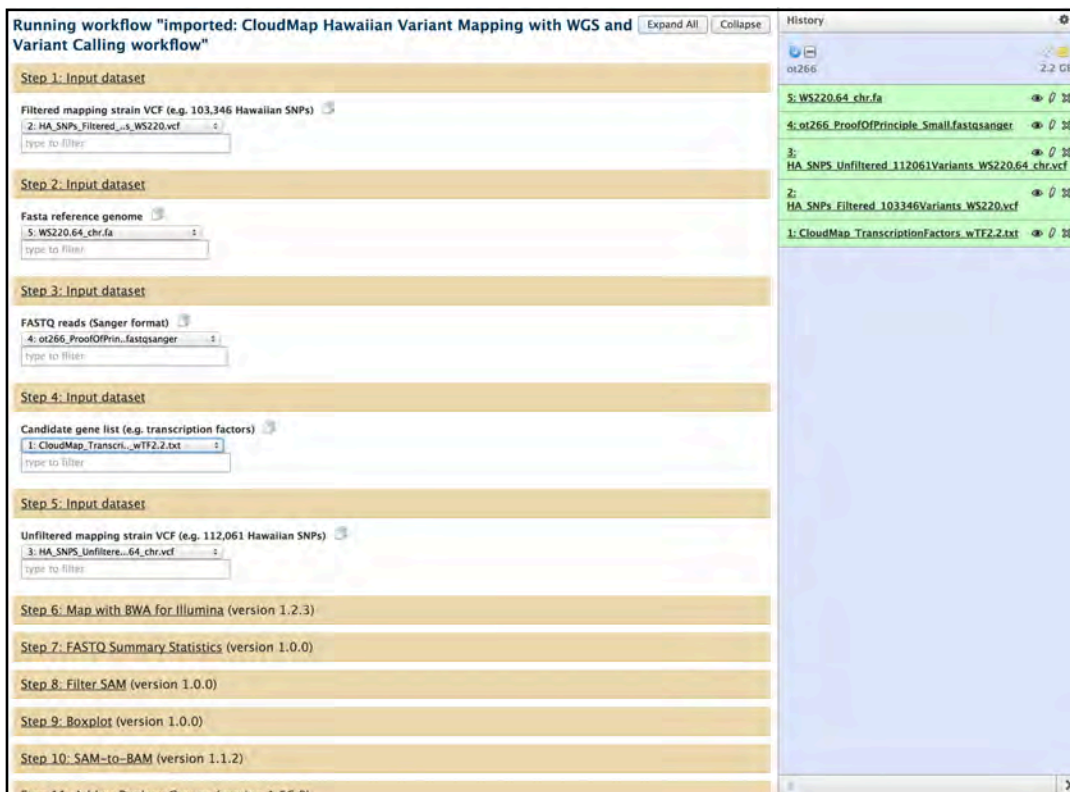
Name	# of Steps
imported: CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow	29

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

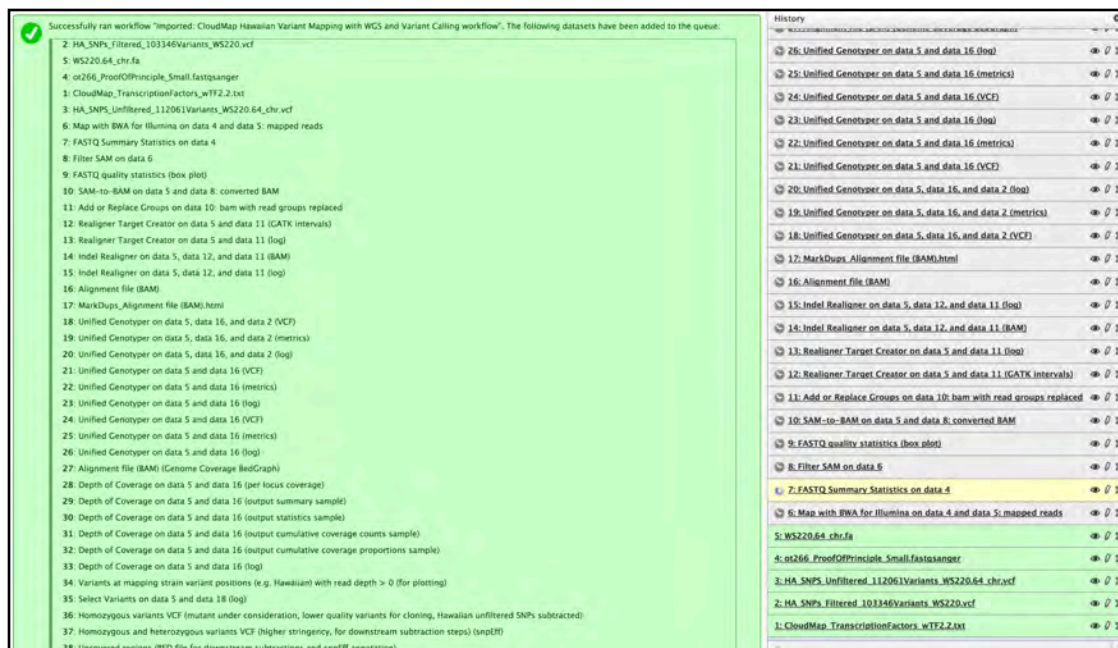


The screenshot shows the 'Your workflows' section. The workflow name 'imported: CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow' is selected. A context menu is open over the name, with the 'Run' option circled in red. The context menu options are: Edit, Run, Share or Publish, Download or Export, Clone, Rename, View, and Delete.

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.



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:



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

The screenshot shows the Galaxy web interface. On the left, there's a tutorial titled "Running Your Own Understanding how Galaxy works" with "Live Quickies" for various mapping and analysis tools. On the right, the "History" panel displays a list of workflow outputs, including files like "49: Homozygous variants annotated (snpEff)", "48: SnpEff on data 41", and "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:

This screenshot shows the "History" panel with a context menu open over the list of datasets. The menu includes options like "Create New", "Clone", "Copy Datasets", and "Show Hidden Datasets", which is highlighted with a red circle. The list of datasets in the background includes items like "49: Homozygous variants annotated (snpEff)", "48: SnpEff on data 41", and "1: CloudMap_TranscriptionFactors_wTF2.2.txt".

20) You may unhide any files that are hidden:

This screenshot shows a list of files in a CloudMap interface. At the top, a yellow warning banner states: "This dataset has been hidden. Click [here](#) to unhide." Below this, several analysis steps are listed, each with a hidden icon (an eye with a slash) and a delete icon (an X). The steps are:

- 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.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:

This screenshot shows a detailed view of a file in CloudMap. On the left, there are two plots:

- The top plot is a scatter plot titled "Ratio of mapping strain allelic reads" vs "Location (Mb)". It shows a red line fluctuating between 0.0 and 0.4 across a 20 Mb region.
- The bottom plot is a histogram titled "LG X" showing the "Normalized frequency of pure parental alleles" vs "Location (Mb)". The x-axis ranges from 0 to 20 Mb, and the y-axis ranges from 0 to 4000. The histogram shows a distribution of alleles with a prominent peak around 10 Mb.

On the right, a "History" panel lists various analysis steps. The file "40: CloudMap: Hawaiian Variant Mapping with WGS data on data 34" is selected, and its details are shown below:

- 6.8 MB
- Format: pdf, database: ce10
- Download icon (circled in red)
- Image in pdf format

Below these details is a list of other files in the history, including "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", and "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.

The screenshot displays the CloudMap Galaxy interface. On the left is the tool configuration panel for 'CloudMap: Hawaiian Variant Mapping with WGS data (version 1.0.0)'. It includes a species dropdown set to 'C. elegans', a WGS Mutant VCF File dropdown set to '34: (hidden) Variants at mapping plotting', and various parameters for Loess span, Y-axis upper limits, colors, and normalization. An 'Execute' button is at the bottom. On the right is the 'History' panel showing a list of jobs. Job 40, 'CloudMap: Hawaiian Variant Mapping with WGS data on data 34', is highlighted in green and has a red arrow pointing to its 'Run this job again' button.

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):

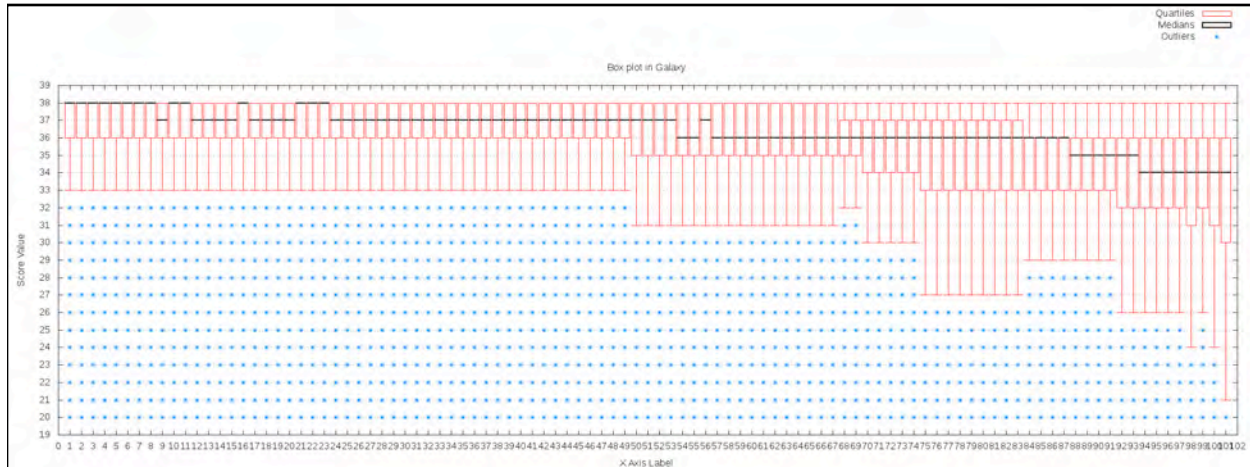
1. A **set of homozygous variants** (VCF file) in the entire sample that can be further filtered by subtracting variants present in other samples using the **CloudMap Subtract Variants** workflow (*GATK*). This VCF file is used as input into *snpeff* to generate the **annotated list of homozygous variants** 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:

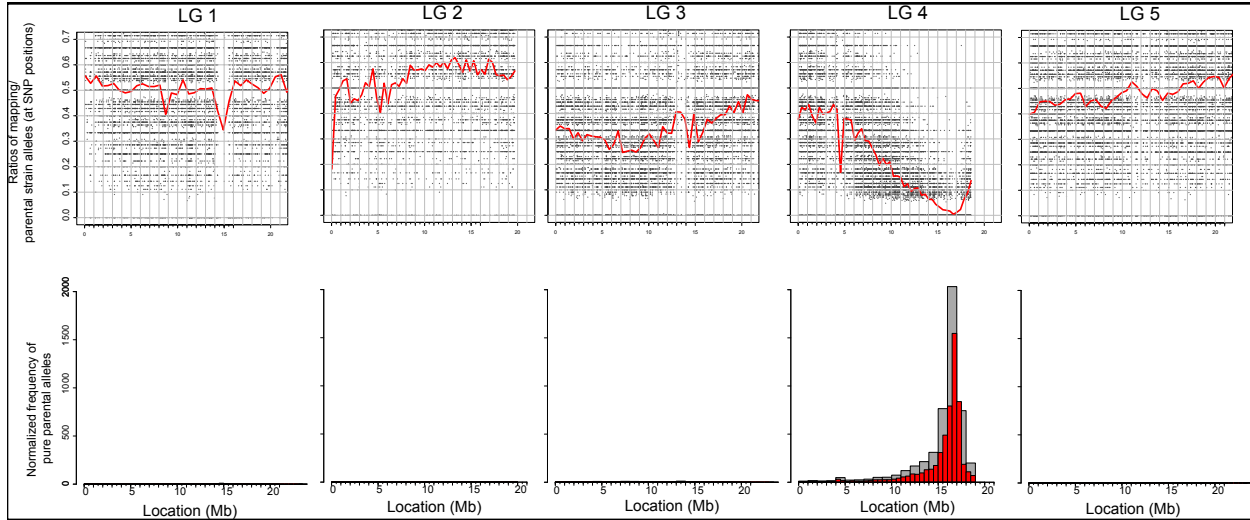
Contents
Summary
Change rate by chromosome
Variants by type
Number of variants by impact
Number of variants by functional class
Number of variants by effect
Quality histogram
Coverage histogram
Base change table
Transition vs transversions (ts/tv)
Frequency of alleles
Codon change table
Amino acid change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

Number of effects by type and region					
Type			Region		
Type (alphabetical order)	Count	Percent	Type (alphabetical order)	Count	Percent
CODON_INSERTION	1	0.001%	DOWNSTREAM	36,909	45.796%
DOWNSTREAM	36,909	45.796%	EXON	1,469	1.823%
FRAME_SHIFT	20	0.025%	INTERGENIC	22	0.027%
INTERGENIC	22	0.027%	INTRON	4,139	5.136%
INTRON	4,139	5.136%	NONE	199	0.247%
NON_SYNONYMOUS_CODING	724	0.898%	SPLICE_SITE_ACCEPTOR	3	0.004%
SPLICE_SITE_ACCEPTOR	3	0.004%	SPLICE_SITE_DONOR	1	0.001%
SPLICE_SITE_DONOR	1	0.001%	UPSTREAM	37,818	46.875%
START_GAINED	13	0.016%	UTR_3_PRIME	137	0.17%
START_LOST	1	0.001%	UTR_5_PRIME	98	0.122%
STOP_GAINED	12	0.015%			
SYNONYMOUS_CODING	711	0.882%			
TRANSCRIPT	199	0.247%			
UPSTREAM	37,818	46.875%			
UTR_3_PRIME	137	0.17%			
UTR_5_PRIME	98	0.105%			

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	4	3	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

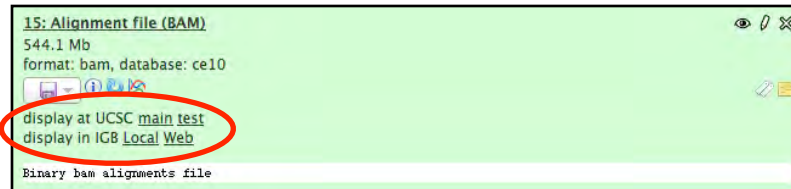
23.2) **Annotated set of homozygous variants (Fig.4) (snpeff)**

Fig. 4 - Sample screenshot of snpeff output

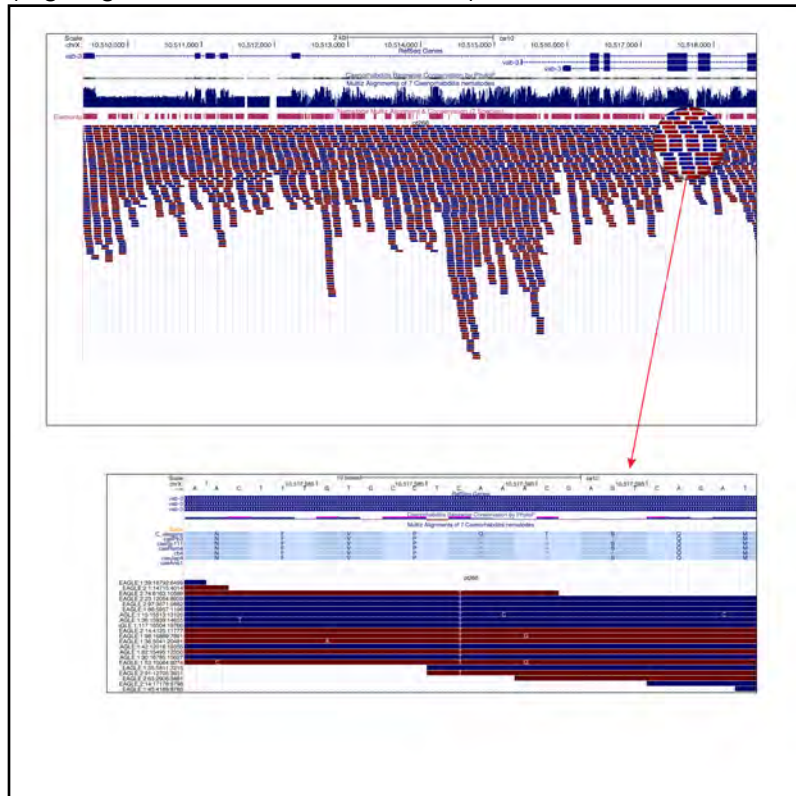
#	Chromo	Position	Reference	Change	Change_type	Quality	Coverage	Gene_ID	Gene_name	Bio_type	Transcript_ID	Exon_Rank	Effect	old_AA/new_AA	Old_codon/New_codon	Codon	Num(CDS)	CDS_size
1	V	19485472	*G	INS		299.66	10	V43188.17	V43188.17	pseudogene	V43188.17	17	TRANSCRIPT: V43188.17					621
2	X	2165878	*G	INS		2399.2	52	F4889.3	F4889.3	protein_coding	F4889.3	9	5 FRAME_SHIFT: F4889.3					585
3	X	3412021	*T	DEL		196.55	25	CD4F6.8	CD4F6.8	ncRNA	CD4F6.8	8	TRANSCRIPT: CD4F6.8					124
4	X	3503048	T	C	SNP	37.15	7	T2322.11	T2322.11	ncRNA	T2322.11	11	TRANSCRIPT: T2322.11					148
5	X	6338449	C	T	SNP	157.66	5	SS501.1	igom2	protein_coding	SS501.1	1	5 NON_SYNONYMOUS_CODING	G/R	Gag/Arg		188	1911
6	X	7037478	*G	INS		210.28	7	BD403.11	BD403.12	ncRNA	BD403.12	12	TRANSCRIPT: BD403.12					200
7	X	7037478	*G	INS		210.28	7	BD403.11	BD403.13	ncRNA	BD403.13	13	TRANSCRIPT: BD403.13					203
8	X	7191138	*C	INS		726.28	26	K03A1.1	K03A1.1	pseudogene	K03A1.1	1	TRANSCRIPT: K03A1.1					410
9	X	7719013	*C	INS		635.6	22	K09P5.11	K09P5.11	ncRNA	K09P5.11	11	TRANSCRIPT: K09P5.11					137
10	X	7719013	*C	INS		635.6	22	K09P5.10	K09P5.10	ncRNA	K09P5.10	10	TRANSCRIPT: K09P5.10					126
11	X	7825447	*T	INS		300.36	16	RD305.8	RD305.8	ncRNA	RD305.8	8	TRANSCRIPT: RD305.8					161
12	X	7866252	*A	DEL		1247.88	50	CS402.16	CS402.16	ncRNA	CS402.16	16	TRANSCRIPT: CS402.16					349
13	X	8026796	*T	INS		337.94	10	C34D10.2	C34D10.2	protein_coding	C34D10.2	2	UTR_3_PRIME: 1423 bases from CDS					27 - CCGH - 2 domains
14	X	8292734	C	T	SNP	1085.02	41	F1389.1	F1389.1	protein_coding	F1389.1b	1b	14 NON_SYNONYMOUS_CODING	S/F	TCG/TTC		1426	4845
15	X	8292734	C	T	SNP	1085.02	41	F1389.1	F1389.1	protein_coding	F1389.1a	1a	15 NON_SYNONYMOUS_CODING	A/F	TCG/TTC		1448	4899
16	X	8292734	C	T	SNP	1085.02	41	F1389.1	F1389.1	protein_coding	F1389.1c	1c	14 NON_SYNONYMOUS_CODING	V/F	TCG/TTC		1426	4820
17	X	8408774	*C	INS		476.87	12	F08F1.18	F08F1.18	ncRNA	F08F1.18	18	TRANSCRIPT: F08F1.18					283
18	X	8639239	*C	INS		775.11	16	F12D9.18	F12D9.18	ncRNA	F12D9.18	18	TRANSCRIPT: F12D9.18					88
19	X	8639239	*G	INS		775.11	16	F12D9.15	F12D9.15	lincRNA	F12D9.15	15	TRANSCRIPT: F12D9.15					91
20	X	8941351	*G	DEL		530.28	15	D1073.1	brn-1	protein_coding	D1073.1b	1b	15 FRAME_SHIFT: D1073.1b					2523
21	X	8941351	*G	DEL		530.28	15	D1073.1	brn-1	protein_coding	D1073.1a	1a	13 FRAME_SHIFT: D1073.1a					2115
22	X	934810	*A	INS		654.81	30	T2085.3	sgo-1	protein_coding	T2085.3a	3a	UTR_3_PRIME: 75 bases from CDS					
23	X	10882433	C	T	SNP	1776.49	42	C30D1.1	elt-2	protein_coding	C30D1.1	1	7 NON_SYNONYMOUS_CODING	S/F	TCG/TTC		811	1802 2f - DATA
24	X	10517587	C	T	SNP	376.64	16	F14F3.1	vab-3	protein_coding	F14F3.1b	1b	4 STOP_GAINED	Q/*	Caa/Taa		152	810 HD - PRD, Paired Domain - FULL
25	X	10517587	C	T	SNP	376.64	16	F14F3.1	vab-3	protein_coding	F14F3.1a	1a	9 STOP_GAINED	Q/*	Caa/Taa		338	1368 HD - PRD, Paired Domain - FULL
26	X	10517587	T	A	SNP	376.64	16	F14F3.1	vab-3	protein_coding	F14F3.1c	1c	4 STOP_GAINED	Q/*	Caa/Taa		179	891 HD - PRD, Paired Domain - FULL
27	X	1160051	C	T	SNP	572.86	22	T04R8.1	lfn-1.5	protein_coding	T04R8.1	1	5 NON_SYNONYMOUS_CODING	G/R	GCG/GAG		214	975
28	X	11695513	C	T	SNP	422.81	19	C44C10.4	C44C10.4	protein_coding	C44C10.4	4	7 NON_SYNONYMOUS_CODING	V/F	Ctc/Ttc		535	1614
29	X	12492601	*G	INS		681.86	18	F45E6.7	F45E6.7	ncRNA	F45E6.7	7	TRANSCRIPT: F45E6.7					145
30	X	14060338	C	SNP		85.86	8	C36G3.13	C36G3.13	ncRNA	C36G3.13	13	TRANSCRIPT: C36G3.13					71
31	X	14305870	C	T	SNP	1288.01	46	C11H1.2	C11H1.2	protein_coding	C11H1.2	2	7 SYNONYMOUS_CODING	K/K	aaG/aaa		292	1383
32	X	16509778	*A	DEL		809.66	24	F58C12.8	F58C12.8	ncRNA	F58C12.8	8	TRANSCRIPT: F58C12.8					275
33	X	17255200	T	C	SNP	45.01	14	V46C7B.1	V46C7B.1	protein_coding	V46C7B.1	1	1 SYNONYMOUS_CODING	V/V	zTA/mu		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:



(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 (snEff). 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/>)

J	A	B	C	D	E	F	G	H	I	J
# Chromo	Position	Reference	Homozygous	Coverage	Gene_name	Bio_type	Transcript_ID	Exon_ID	old_AA/new_AA	
2		2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.4	UPSTREAM: 8859 bases
3		2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.6	UPSTREAM: 8972 bases
4		2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.3	UPSTREAM: 7767 bases
5		2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.2	UPSTREAM: 8849 bases
6		2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.1	UPSTREAM: 8853 bases
7		2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.5	UPSTREAM: 8853 bases
8		2646	2664	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.1	DOWNSTREAM: 1473 bases
9		2646	2664	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.2	DOWNSTREAM: 1575 bases
10		2646	2664	Interval	0	Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6	DOWNSTREAM: 1101 bases
11		3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.4	UPSTREAM: 8037 bases
12		3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.6	UPSTREAM: 8150 bases
13		3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.3	UPSTREAM: 6945 bases
14		3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.2	UPSTREAM: 8027 bases
15		3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.1	UPSTREAM: 8031 bases
16		3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.5	UPSTREAM: 8031 bases
17		3468	3482	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.1	DOWNSTREAM: 651 bases
18		3468	3482	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.2	DOWNSTREAM: 753 bases
19		3468	3482	Interval	0	Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6	DOWNSTREAM: 279 bases
20		3926	4014	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.4	UPSTREAM: 7579 bases
21		3926	4014	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.6	UPSTREAM: 7692 bases
22		3926	4014	Interval	0	Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6	UPSTREAM: 17 bases
23		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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	2645	2664	0
2	chr1	3467	3482	0
3	chr1	3925	4014	0
4	chr1	8673	8703	0
5	chr1	8835	8995	0
6	chr1	9774	9787	0
7	chr1	11219	11317	0
8	chr1	11450	11469	0
9	chr1	15107	15117	0
10	chr1	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.

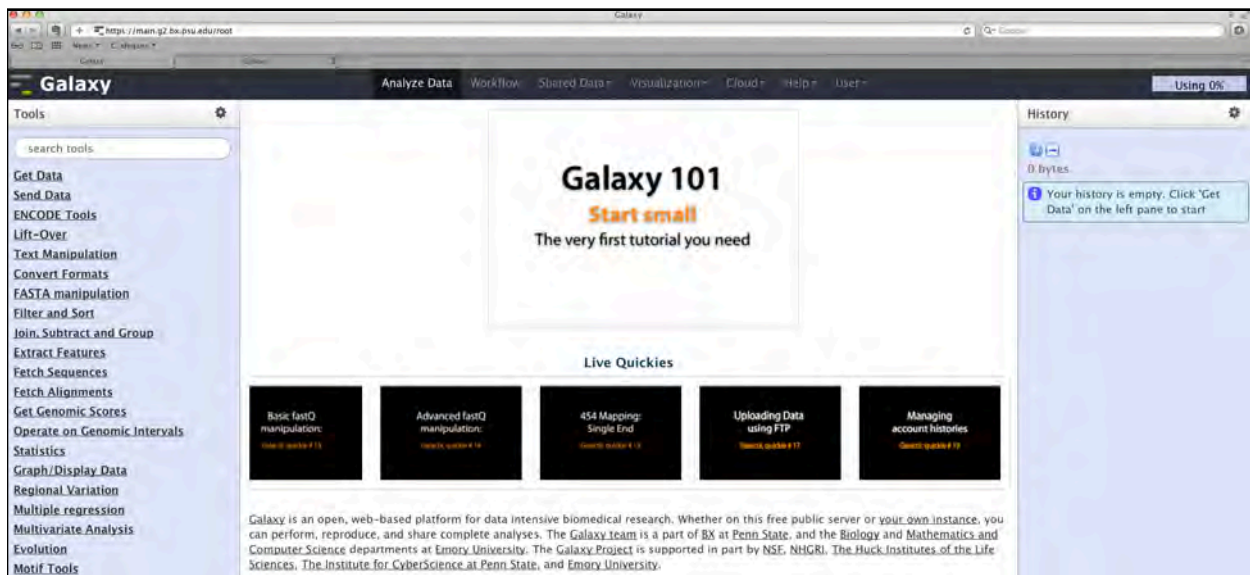
CloudMap UnMapped Mutant Workflow

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>.

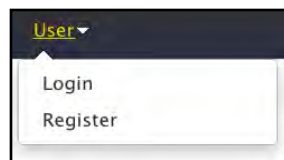
The *ot266* FASTQ file used in this example represents Hawaiian variant mapped data but for the purposes of this user guide, we perform an unmapped analysis. Users wishing to run their own unmapped data should also view the **Analyzing Your Own Data** section of this user guide before proceeding.

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

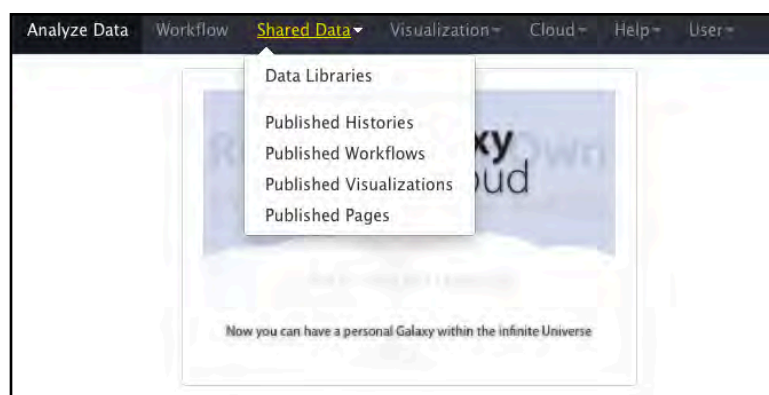


The screenshot displays the Galaxy web interface. The main content area features a large banner for "Galaxy 101 Start small" with the subtitle "The very first tutorial you need". Below this banner is a section titled "Live Quickies" containing five tool cards: "Basic fastQ manipulation", "Advanced fastQ manipulation", "454 Mapping: Single End", "Uploading Data using FTP", and "Managing account histories". The left sidebar lists various tool categories such as "Get Data", "Send Data", "ENCODE Tools", "Lift-Over", "Text Manipulation", "Convert Formats", "FASTA manipulation", "Filter and Sort", "Join, Subtract and Group", "Extract Features", "Fetch Sequences", "Fetch Alignments", "Get Genomic Scores", "Operate on Genomic Intervals", "Statistics", "Graph/Display Data", "Regional Variation", "Multiple regression", "Multivariate Analysis", "Evolution", and "Motif Tools". The right sidebar shows a "History" section with "0 bytes" and a message: "Your history is empty. Click 'Get Data' on the left pane to start". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Cloud", "Help", and "User".

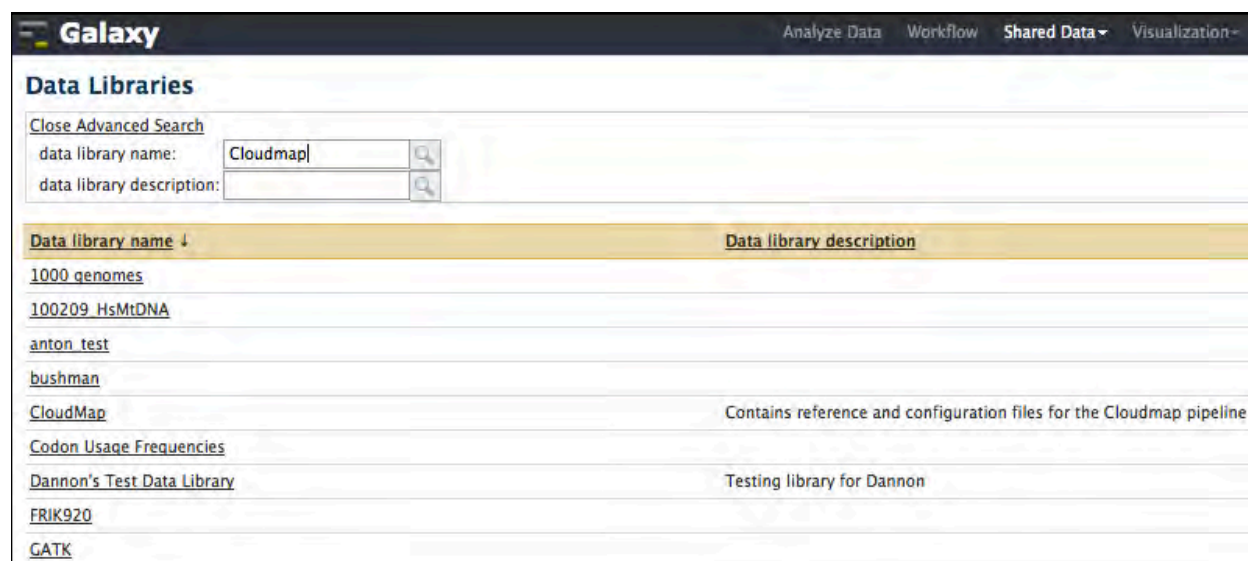
2) Register for an account or login if you already have an account:



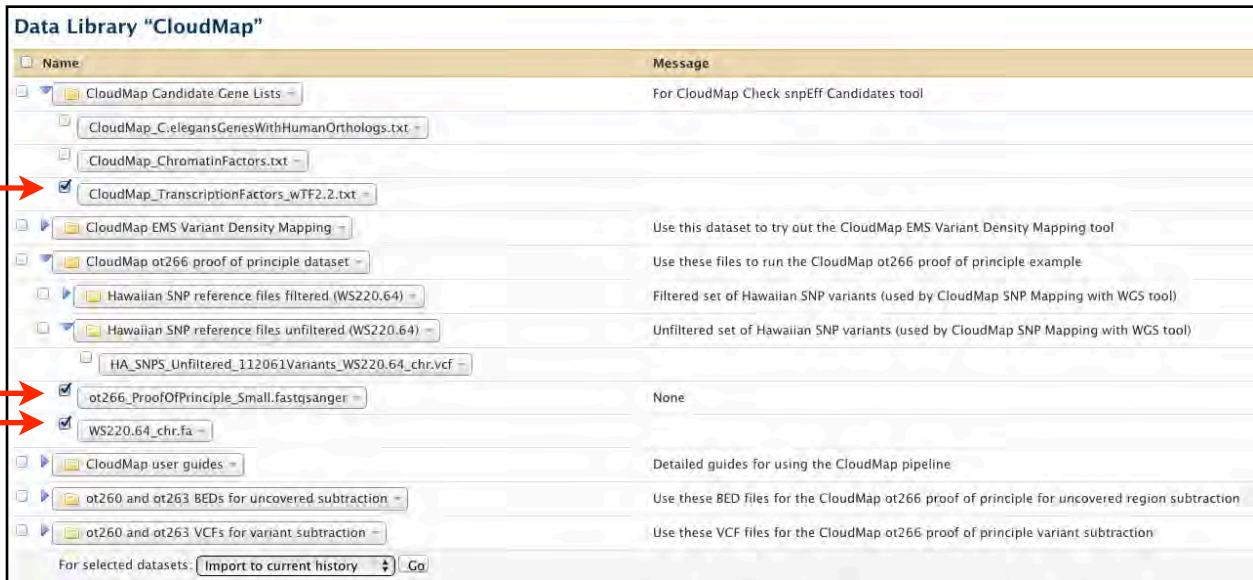
3) Once you are logged in using your email address, click on the **Shared Data** link at the top of the page:



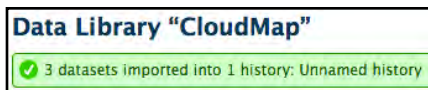
4) Click on **Data Libraries** and search for the CloudMap data library:



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.



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



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



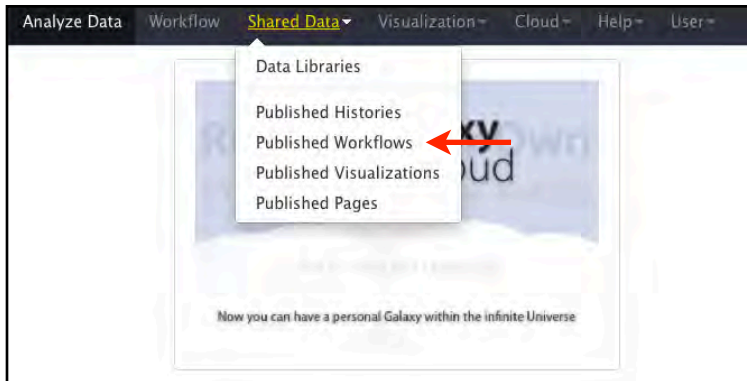
8) You will now see that the data files have been added to an unnamed history:



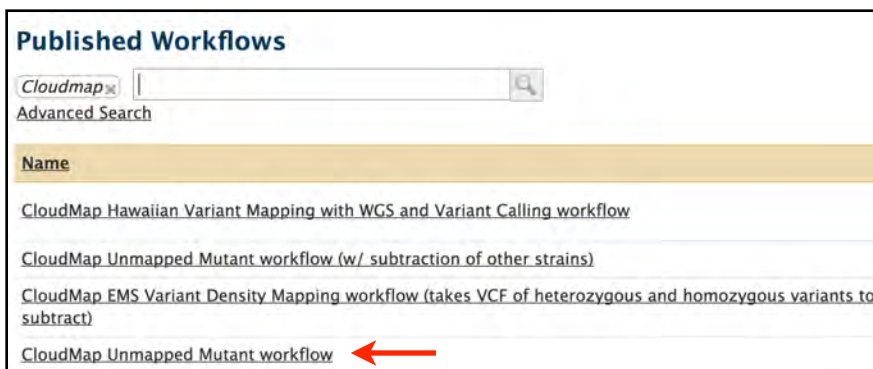
9) Name your history **ot266** after the sample that we will be analyzing:



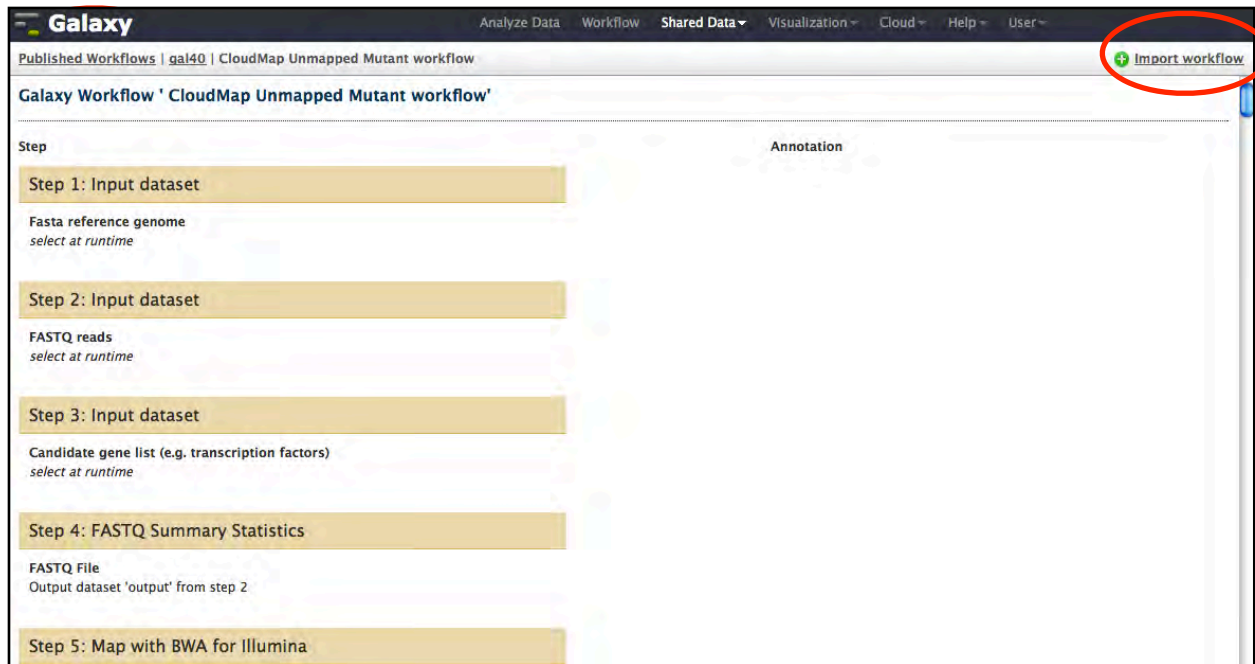
10) Again click on the **Shared Data** link at the top of the page and select **Published Workflows** :



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

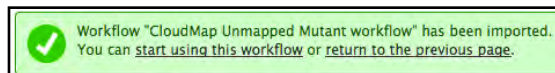


12) You will now have the option to **Import workflow**



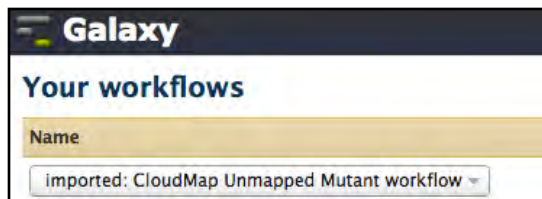
The screenshot shows the Galaxy web interface. At the top, there is a navigation bar with the 'Galaxy' logo and several menu items: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. Below the navigation bar, the breadcrumb trail reads 'Published Workflows | gal40 | CloudMap Unmapped Mutant workflow'. In the top right corner, a green button with a plus sign and the text 'import workflow' is circled in red. The main content area displays a workflow titled 'Galaxy Workflow ' CloudMap Unmapped Mutant workflow''. The workflow is organized into five steps, each with a yellow header bar and a description of the step's input and output. The steps are: 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**.



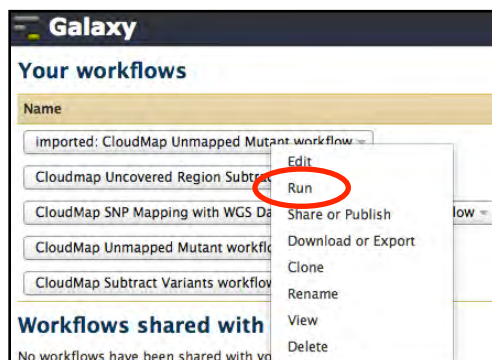
A green notification box with a white checkmark icon on the left. The text inside reads: 'Workflow "CloudMap Unmapped Mutant workflow" has been imported. You can start using this workflow or return to the previous page.'

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



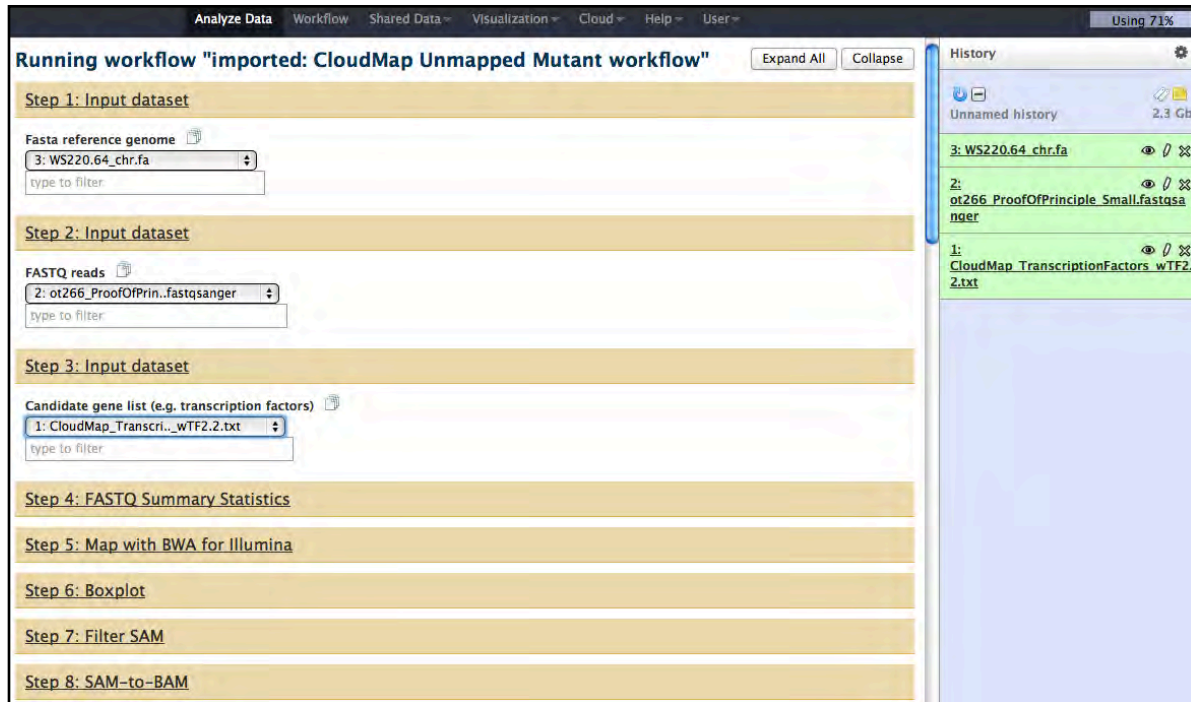
The screenshot shows the 'Your workflows' section of the Galaxy interface. The title 'Your workflows' is in blue. Below it, there is a table with a 'Name' column. The first entry in the table is 'imported: CloudMap Unmapped Mutant workflow'.

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

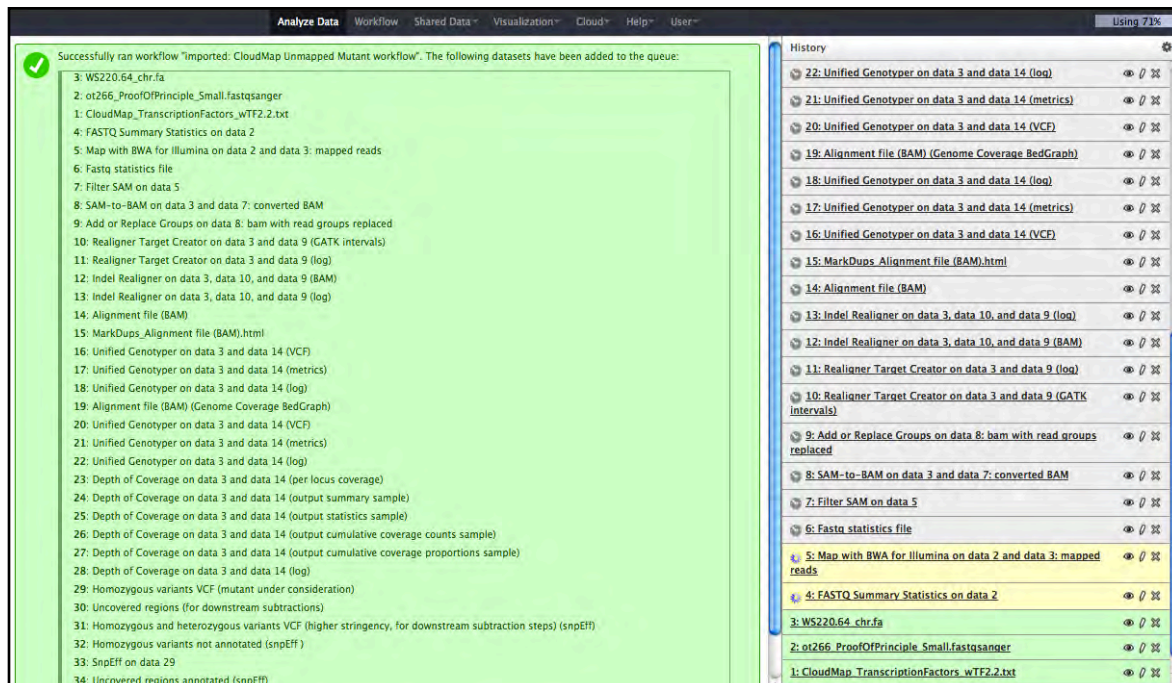


The screenshot shows the 'Your workflows' section of the Galaxy interface. A context menu is open over the workflow 'imported: CloudMap Unmapped Mutant workflow'. The menu items are: 'Edit', 'Run', 'Share or Publish', 'Download or Export', 'Clone', 'Rename', 'View', and 'Delete'. The 'Run' option is circled in red.

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.



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:



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

The screenshot shows the Galaxy web interface. The main content area has a yellow warning box that says "Hello world! This is galaxy test." and a blue information box that says "Galaxy Test has usage quotas." Below these is a graphic of a funnel with the text "the test is for breaking". The right-hand side features a "History" panel with a list of 36 output files, numbered 1 to 36. The files include "CloudMap_TranscriptionFactors_wTF2.2.txt", "ot266_ProofOfPrinciple_Small.fastqsanger", "WS220.64_chr.fa", "Fastq statistics file", "Alignment file (BAM)", "Depth of Coverage on data 3 and data 14 (output summary sample)", "Homozygous variants VCF (mutant under consideration)", "Uncovered regions (for downstream subtractions)", "Homozygous and heterozygous variants VCF (higher stringency for downstream subtraction steps) (snpEff)", "Uncovered regions annotated (snpEff)", and "Homozygous variants annotated (snpEff)".

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:

This is a close-up of the History menu in the Galaxy interface. The menu is open, showing a list of actions. A red arrow points to the "Show Hidden Datasets" option. Other visible options include "Create New", "Clone", "Copy Datasets", "Share or Publish", "Extract Workflow", "Dataset Security", "Show Deleted Datasets", "Purge Deleted Datasets", "Show Structure", "Export to File", "Delete", "Delete Permanently", and "Import from File". The background shows the same list of 36 output files as in the previous screenshot.

20) You may unhide any files that are hidden:

History

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:

Galaxy

Analyze Data | Workfiles | Shared Data | Visualization | Cloud | Help | Logout

Tools: search tools

Chroma	Position	Reference	Change	Change_type	Homozygous	Quality	Coverage	Warnings	Gene_ID	Gene
I	42899	C	A	SNP	Hom	75.03	3		V48G1C.12	V48G
I	42899	C	A	SNP	Hom	75.03	3		V48G1C.4	
I	42899	C	A	SNP	Hom	75.03	3		Y74C9A.1	Y74C
I	62642	T	C	SNP	Hom	48.77	2		V48G1C.4	Y48G
I	62642	T	C	SNP	Hom	48.77	2		V48G1C.5	Y48G
I	62642	T	C	SNP	Hom	48.77	2		V48G1C.2	Y48G
I	62642	T	C	SNP	Hom	48.77	2		V48G1C.2	Y48G
I	72355	A	G	SNP	Hom	48.77	2		V48G1C.2	Y48G
I	72355	A	G	SNP	Hom	48.77	2		V48G1C.10	Y48G
I	72355	A	G	SNP	Hom	48.77	2		V48G1C.2	Y48G
I	72355	A	G	SNP	Hom	48.77	2		V48G1C.2	Y48G
I	72355	A	G	SNP	Hom	48.77	2		V48G1C.11	Y48G
I	72355	A	G	SNP	Hom	48.77	2		V48G1C.5	Y48G
I	200948	A	T	SNP	Hom	317.33	9		V48G1B1.7	Y48G
I	200948	A	T	SNP	Hom	317.33	9		K10E9.1	K10E
I	200948	A	T	SNP	Hom	317.33	9		V48G1B1.2	Y48G
I	200949	C	T	SNP	Hom	309.05	9		V48G1B1.7	Y48G
I	200949	C	T	SNP	Hom	309.05	9		K10E9.1	K10E
I	200949	C	T	SNP	Hom	309.05	9		V48G1B1.2	Y48G
I	341300	-	-G	DEL	Hom	181.31	6		V48G1A.6	Y48G
I	341300	-	-G	DEL	Hom	181.31	6		V48G1A.6	Y48G
I	341300	-	-G	DEL	Hom	181.31	6		V48G1A.3	Y48G
I	341300	-	-G	DEL	Hom	181.31	6		V48G1A.1	Y48G
I	341300	-	-G	DEL	Hom	181.31	6		V48G1A.2	Y48G
I	341300	-	-G	DEL	Hom	181.31	6		V48G1A.2	Y48G
I	346149	T	A	SNP	Hom	85.77	3		V48G1A.3	Y48G
I	346149	T	A	SNP	Hom	85.77	3		V48G1A.1	Y48G
I	346149	T	A	SNP	Hom	85.77	3		V48G1A.6	Y48G
I	346149	T	A	SNP	Hom	85.77	3		V48G1A.6	Y48G
I	346149	T	A	SNP	Hom	85.77	3		V48G1A.2	Y48G
I	346149	T	A	SNP	Hom	85.77	3		V48G1A.2	Y48G
I	361325	C	A	SNP	Hom	232.91	7		R119.7	R119
I	361325	C	A	SNP	Hom	232.91	7		V48G1A.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

History

This dataset has been hidden. Click [here](#) to unhide.

36: Homozygous variants annotated (snpeff)

30: 595 lines, 3 comments

Tabular, database: ce10

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)

28: Death of Coverage on data 3 and data 14 (output_summary sample)

14: Alignment file (BAM)

6: Fastq statistics file

3: WS220.64_chr1a

2: ot266_ProofOfPrinciple_Smallfastxanalyzer

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)

Candidate List:
1: CloudMap_Transcri..._wTF2.2.txt

Execute

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/ui/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 Nekutenko, 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.)

History

ot266 12.7 Gb

36: Homozygous variants annotated (snpEff)

Run this job again

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/pool12/files/000/384/dataset_384498.dat								
# Chromo	Position	Reference	Change	Change_type	Homozyg			
Interval_ID								
I	42899	G	A	SNP	Non	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.fastq.sanger

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:

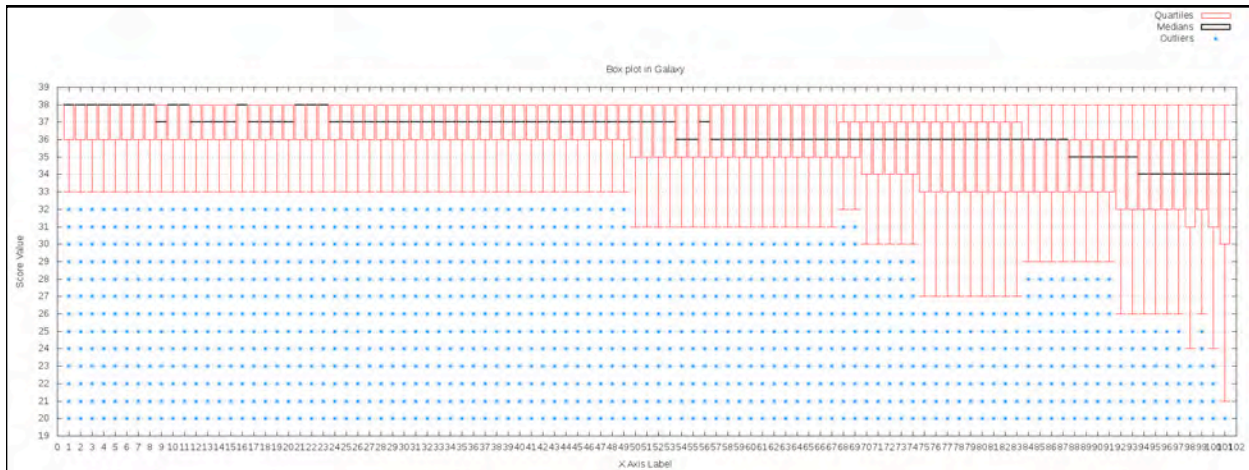
1. An **annotated set of homozygous variants** in the entire sample (*snpeff*). (For more information on file format, see: <http://snpeff.sourceforge.net/>)
2. A **BAM alignment file** that can be viewed in your choice of alignment viewers (*SAMtools*). (For more information on file format, see: <http://genome.ucsc.edu/FAQ/FAQformat>)
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	Region	
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	98	0.122%

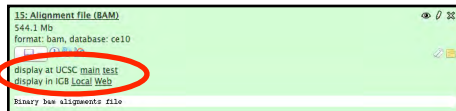
Examples of **primary set of files for analysis** (mentioned in step 23 above):

24.1) Annotated set of homozygous variants (Fig.4) (snpeff)

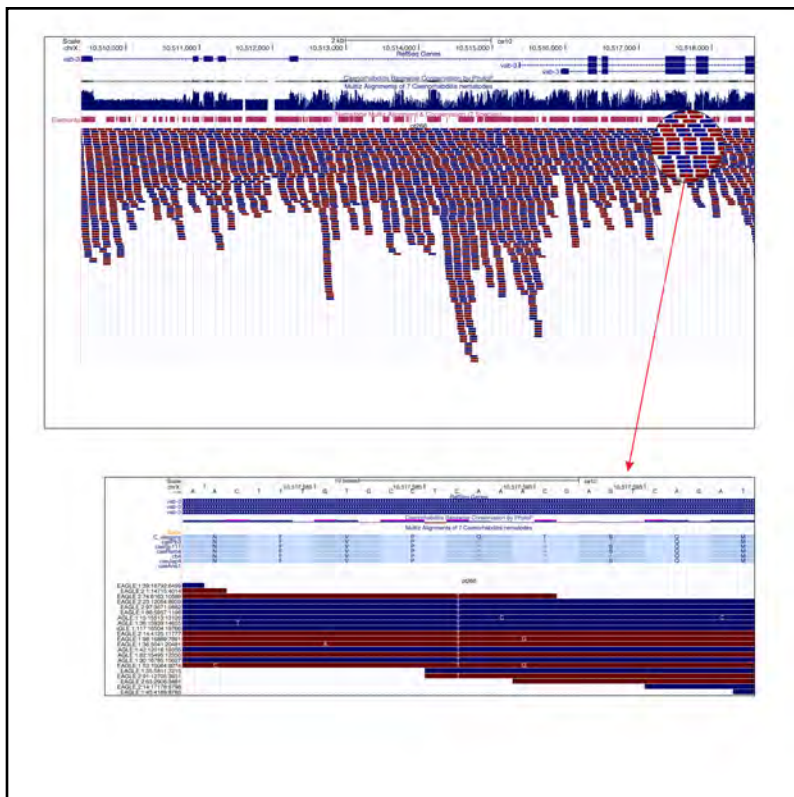
Fig. 4: Sample screenshot of snpEff output

Chrom	Position	Reference	Change	Type	Quality	Coverage	Gene_ID	Gene_name	Bio_Type	Transcript_ID	Exon_Rank	Effect	old_AA/new_AA	Old_Codon/New_Codon	Codon_Num(CDS)	CDS_size		
1	3942472	A	G	INS	299	66	10	Y43F88.17	psiudogene	Y43F88.17		TRANSCRIPT: Y43F88.17				621		
1	2163879	A	G	INS	2399	2	12	F4889.3	protein_coding	F4889.3	5	5	FRAMESHIFT	F4889.3		585		
1	3412021	A	T	DEL	196	55	25	C04F6.8	ncRNA	C04F6.8		TRANSCRIPT: C04F6.8				124		
1	3903048	T	C	SNP	37	15	2	T2282.11	ncRNA	T2282.11		TRANSCRIPT: T2282.11				148		
1	6384849	C	T	SNP	157	66	5	S5501.1	lincRNA	S5501.1		5	NON_SYNONYMOUS_CODING	G/R	Gag/Arg	188	1911	
1	7037478	A	G	INS	210	28	7	80403.12	ncRNA	80403.12		TRANSCRIPT: 80403.12				200		
1	7037478	A	G	INS	210	28	7	80403.11	ncRNA	80403.11		TRANSCRIPT: 80403.11				203		
1	7310138	A	C	INS	726	28	26	K03A1.1	psiudogene	K03A1.1		TRANSCRIPT: K03A1.1				410		
1	7739013	A	C	INS	635	6	22	K09F5.11	ncRNA	K09F5.11		TRANSCRIPT: K09F5.11				137		
1	7739013	A	C	INS	635	6	22	K09F5.10	ncRNA	K09F5.10		TRANSCRIPT: K09F5.10				126		
1	7823447	A	T	INS	350	18	18	R0365.8	ncRNA	R0365.8		TRANSCRIPT: R0365.8				353		
1	7864252	A	A	DEL	1247	88	50	C5402.16	ncRNA	C5402.16		TRANSCRIPT: C5402.16				349		
1	8026796	A	T	INS	317	34	10	C34010.2	protein_coding	C34010.2	1	UTR_3_PRIME	1423 bases from CDS			27	CCCH_2 domain	
1	8292734	C	T	SNP	1085	62	41	F1389.1	protein_coding	F1389.1	14	14	NON_SYNONYMOUS_CODING	S/F	ICG/IT	1426	4845	
1	8292734	C	T	SNP	1085	62	41	F1389.1	protein_coding	F1389.1	15	15	NON_SYNONYMOUS_CODING	S/F	ICG/IT	1448	4899	
1	8292734	C	T	SNP	1085	62	41	F1389.1	protein_coding	F1389.1	14	14	NON_SYNONYMOUS_CODING	S/F	ICG/IT	1426	4830	
1	8468774	A	C	INS	476	87	12	F08F1.18	ncRNA	F08F1.18		TRANSCRIPT: F08F1.18				283		
1	8639239	A	CG	INS	775	11	16	F1209.18	ncRNA	F1209.18		TRANSCRIPT: F1209.18				88		
1	8639239	A	CG	INS	775	11	16	F1209.15	ncRNA	F1209.15		TRANSCRIPT: F1209.15				71		
1	8941351	A	GATC	DEL	530	28	15	D1073.1b	protein_coding	D1073.1b	15	15	FRAME_SHIFT	D1073.1b		2523		
1	8941351	A	GATC	DEL	530	28	15	D1073.1a	protein_coding	D1073.1a	12	12	FRAME_SHIFT	D1073.1a		2112		
1	9345610	A	A	INS	654	81	39	T2085.3a	protein_coding	T2085.3a		UTR_3_PRIME	75 bases from CDS					
1	10482433	C	T	SNP	1776	49	42	C3303.1	nc-2	C3303.1		7	NON_SYNONYMOUS_CODING	S/F	ICG/IT	811	1802	7T-GATA
1	10517587	C	T	SNP	376	64	16	F14F3.1	nc-2	F14F3.1		4	STOP_GAINED	Q/*	CaA/Tea	152	810	HD - PRD, Paired Domain - FULL
1	10517587	C	T	SNP	376	64	16	F14F3.1	nc-3	F14F3.1		9	STOP_GAINED	Q/*	Eaa/Tea	338	1368	HD - PRD, Paired Domain - FULL
1	10517587	C	T	SNP	376	64	16	F14F3.1	nc-3	F14F3.1		4	STOP_GAINED	Q/*	Eaa/Tea	179	891	HD - PRD, Paired Domain - FULL
1	11640051	C	T	SNP	573	86	22	T0418.1	nc-1.5	T0418.1		5	NON_SYNONYMOUS_CODING	G/R	Gga/Arg	214	975	
1	11695333	C	T	SNP	427	81	19	C44C10.4	protein_coding	C44C10.4		7	NON_SYNONYMOUS_CODING	S/F	Ccg/Trt	535	1634	
1	12492668	A	G	INS	831	86	18	F4566.7	ncRNA	F4566.7		TRANSCRIPT: F4566.7				145		
1	14060338	T	C	SNP	85	86	3	C3303.13	ncRNA	C3303.13		TRANSCRIPT: C3303.13				71		
1	18305870	C	T	SNP	1285	61	46	C11H1.2	protein_coding	C11H1.2		7	SYNONYMOUS_CODING	A/K	aaG/aaa	272	1383	
1	15568729	A	AG	DEL	899	56	24	F5K12.8	ncRNA	F5K12.8		TRANSCRIPT: F5K12.8				275		
1	22232200	T	C	SNP	45	61	14	V40C78.1	protein_coding	V40C78.1		1	SYNONYMOUS_CODING	V/V	gTA/mu	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:



(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	Transcript_ID	Exon_ID	old_AA/new_AA	
2		2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.4	UPSTREAM: 8859 bases
3		2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.6	UPSTREAM: 8972 bases
4		2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.3	UPSTREAM: 7767 bases
5		2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.2	UPSTREAM: 8849 bases
6		2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.1	UPSTREAM: 8853 bases
7		2646	2664	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.5	UPSTREAM: 8853 bases
8		2646	2664	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.1	DOWNSTREAM: 1473 bases
9		2646	2664	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.2	DOWNSTREAM: 1575 bases
10		2646	2664	Interval	0	Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6	DOWNSTREAM: 1101 bases
11		3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.4	UPSTREAM: 8037 bases
12		3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.6	UPSTREAM: 8150 bases
13		3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.3	UPSTREAM: 6945 bases
14		3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.2	UPSTREAM: 8027 bases
15		3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.1	UPSTREAM: 8031 bases
16		3468	3482	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.5	UPSTREAM: 8031 bases
17		3468	3482	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.1	DOWNSTREAM: 651 bases
18		3468	3482	Interval	0	Y74C9A.3	Y74C9A.3	protein_coding	Y74C9A.3.2	DOWNSTREAM: 753 bases
19		3468	3482	Interval	0	Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6	DOWNSTREAM: 279 bases
20		3926	4014	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.4	UPSTREAM: 7579 bases
21		3926	4014	Interval	0	Y74C9A.2	nlp-40	protein_coding	Y74C9A.2.6	UPSTREAM: 7692 bases
22		3926	4014	Interval	0	Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6	UPSTREAM: 17 bases
23		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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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>)

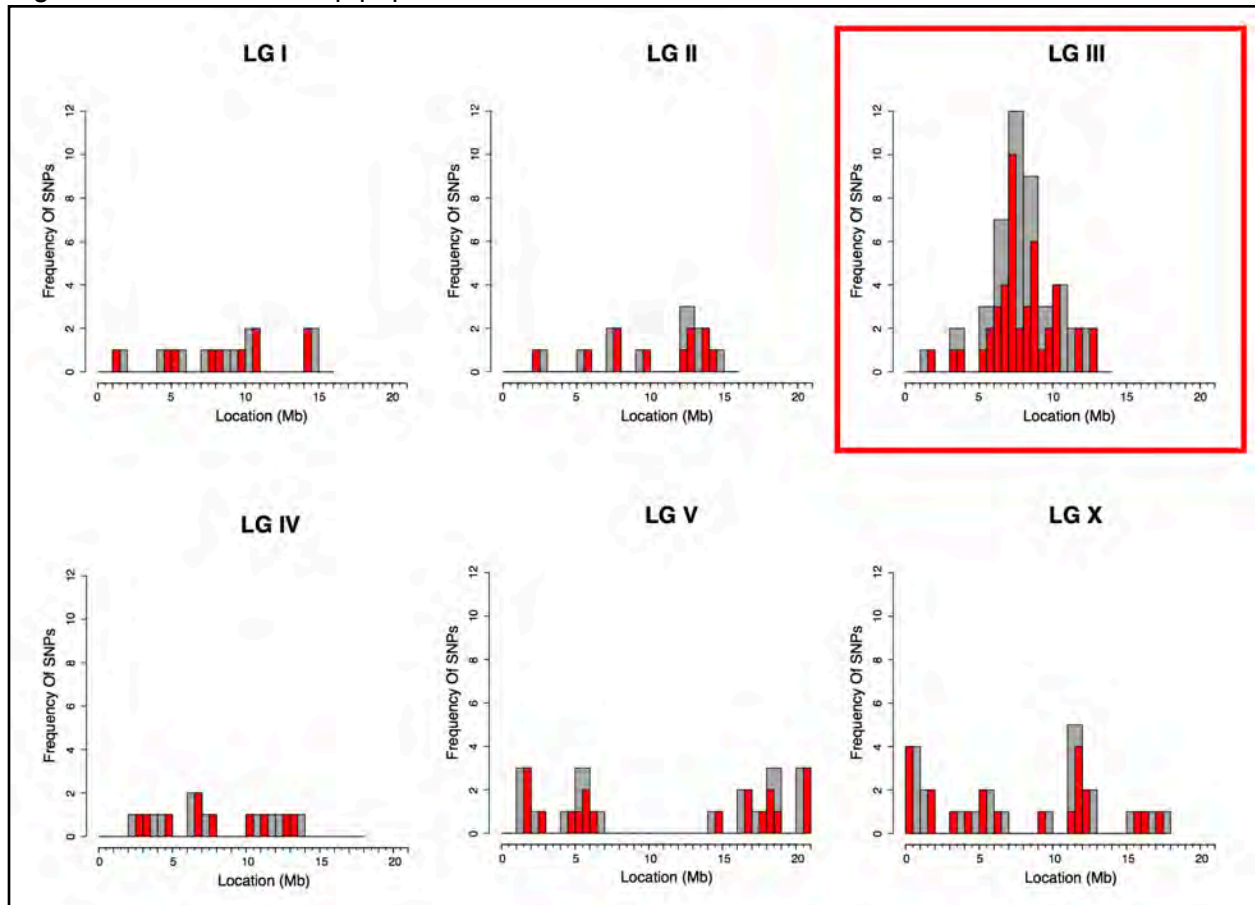
	A	B	C	D
1	chr1	2645	2664	0
2	chr1	3467	3482	0
3	chr1	3925	4014	0
4	chr1	8673	8703	0
5	chr1	8835	8995	0
6	chr1	9774	9787	0
7	chr1	11219	11317	0
8	chr1	11450	11469	0
9	chr1	15107	15117	0
10	chr1	15635	15767	0

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:

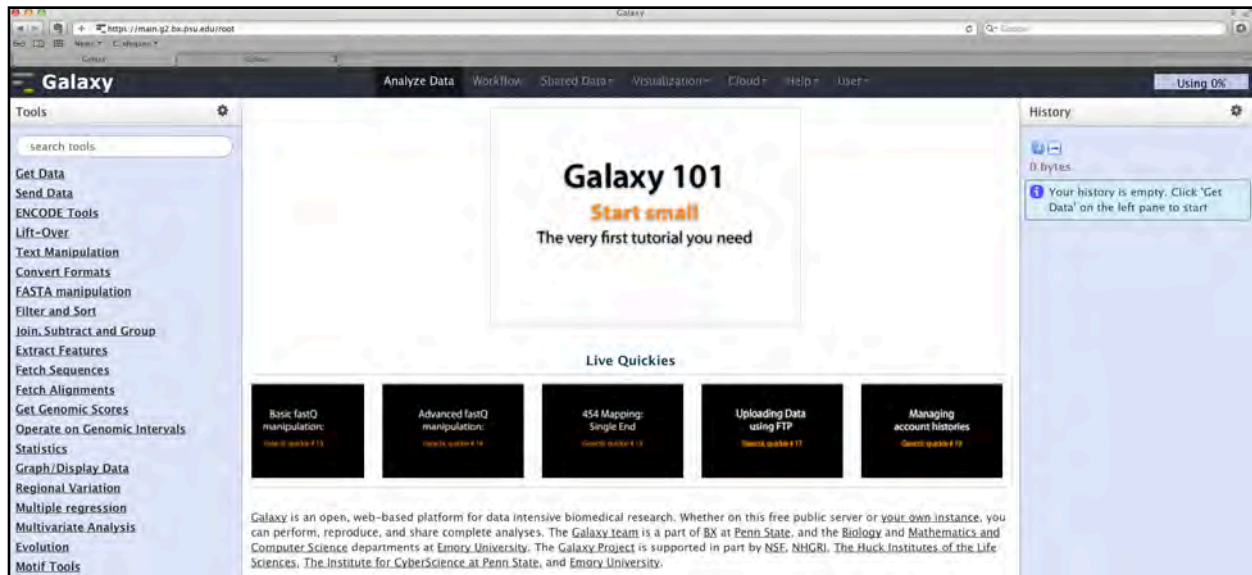


CloudMap Subtract Variants workflow (using *ot266* Proof of Principle example from the CloudMap paper). A video version of this user guide is available at: <http://usegalaxy.org/cloudmap>.

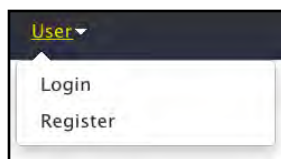
This workflow should be used downstream of either of the following workflows: **Hawaiian Variant Mapping with WGS data and Variant Calling**, **EMS Density Mapping**, or **Unmapped Mutant workflows**. Here we demonstrate the workflow using the *ot266* example from the Cloudmap paper (**Fig.8**). Users 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>



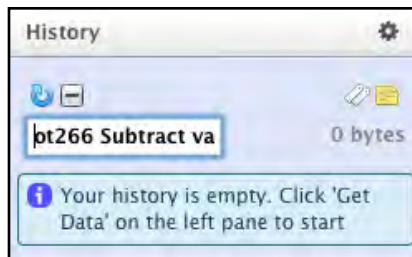
2) You should already have a Galaxy account at this point because you have run earlier workflows:



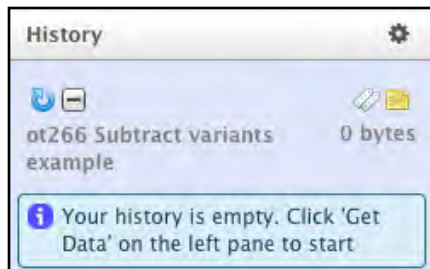
3) Once you are logged in using your email address, create a new history:



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



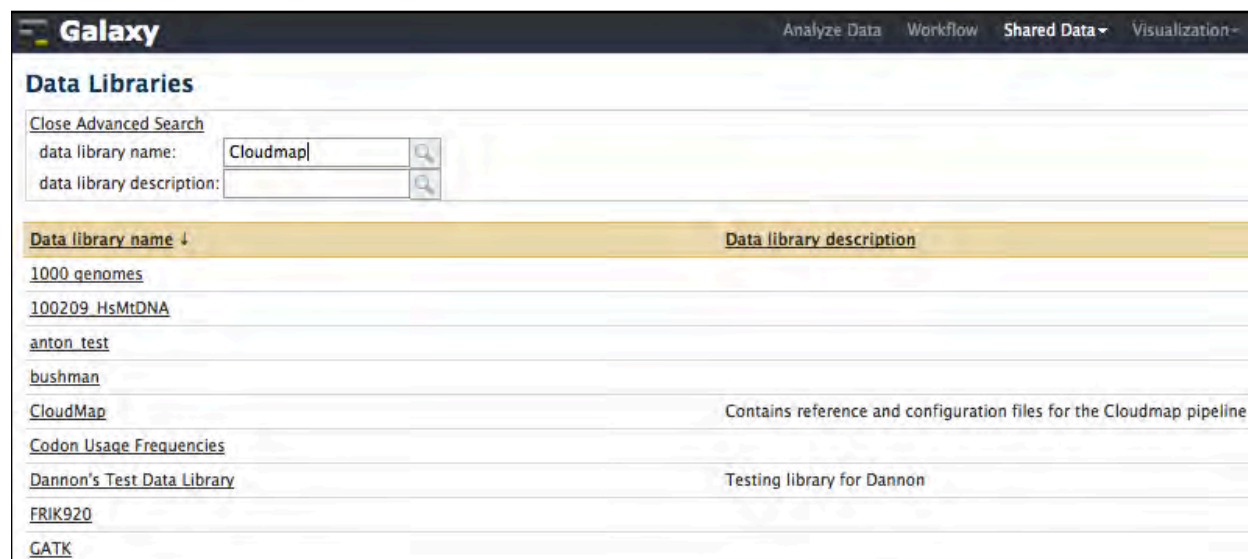
5) You now need to import the **ot266 Proof of principle** files or your own files to run the workflow:



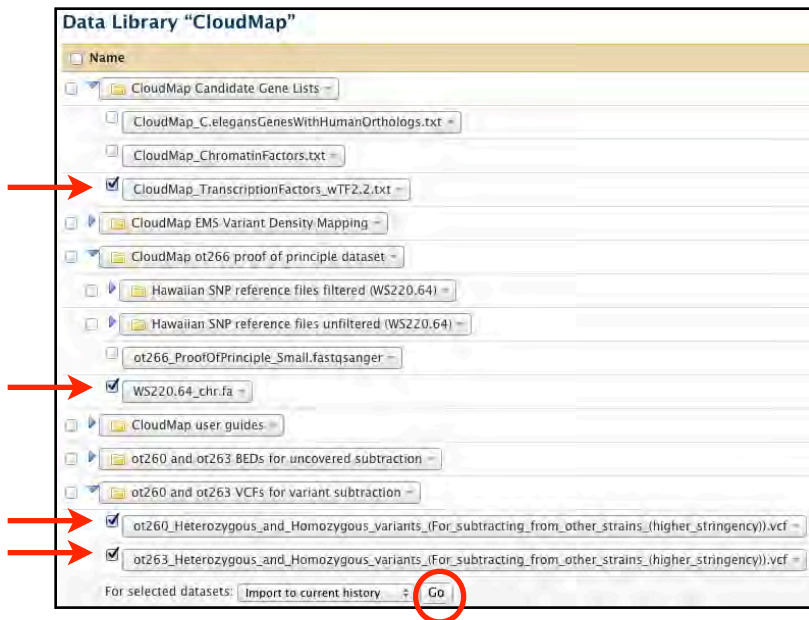
6) Click on the **Shared Data** link at the top of the page:



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

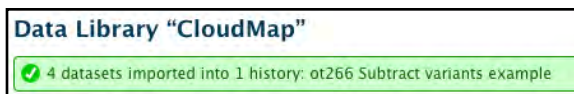


8) Click on the **CloudMap** library and select the 4 data files below for the *ot266* example. Then click “Go” to import these files into your history.



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:



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

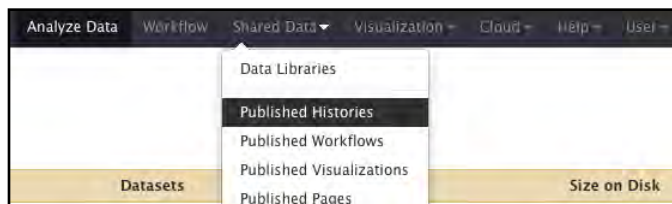


11) You will now see these files in your history:

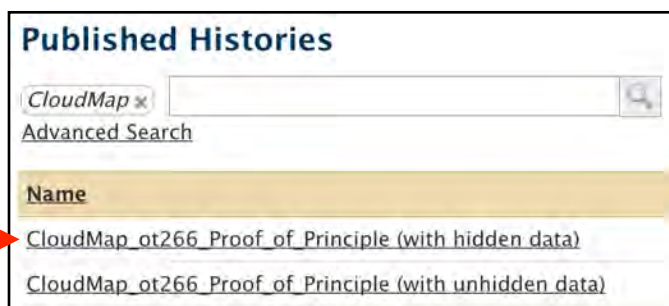


12) You will also need to import homozygous variants (VCF file) from the workflow you performed earlier. In this example, we will use the *ot266* homozygous variants from running the **Hawaiian Variant Mapping with WGS Data and Variant Calling** workflow. The *ot266* example history is shared so we will 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)**:



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) Import history


Galaxy History ' CloudMap_ot266_Proof_of_Principle (with hidden data)'

Dataset

- 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
[View](#) [Download](#)
display at UCSC [main](#)
- 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)

1. Chrom	2. Pos	3. ID	4. Ref	5. Alt	6. Qual	7. Filter
##FILTER=ID=vcfFilter,Description=VcfFilter v0.2, Expression used: isHom1 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">						

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 (snEff) (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 (snEff)		👁	🗑
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 hidden data)			
12.6 GB			
49: Homozygous variants annotated (snEff) (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 (snEff)		👁	🗑
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)		👁	🗑

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

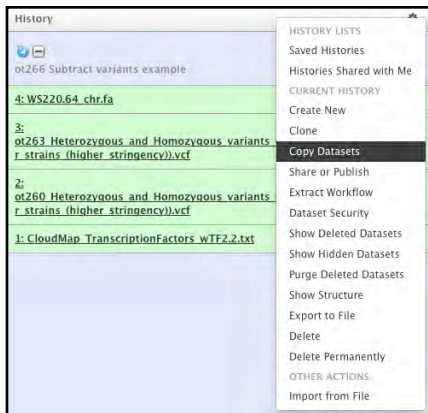
OTHER ACTIONS

- Import from File

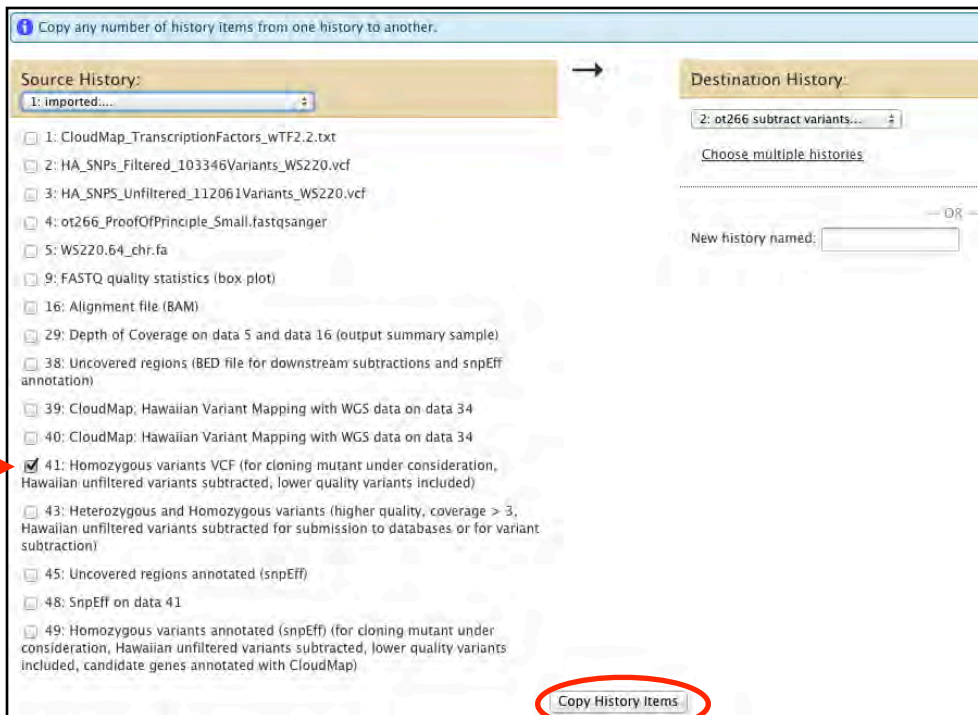
18) Click on the **ot266 Subtract Variants example** history and click **Switch** to return to that history:



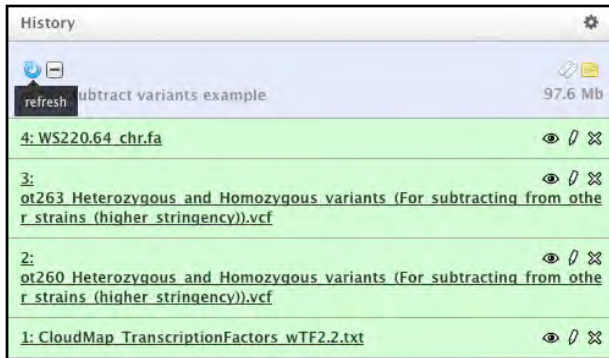
19) To copy the **ot266** homozygous variants into this history, click **Copy Datasets** in your history options:



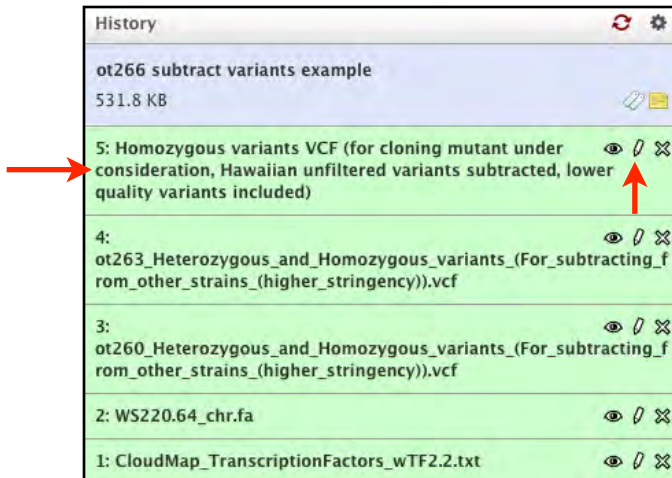
20) Copy the **ot266 Homozygous variants VCF** from the newly imported **ot266** history:



21) Hit refresh in your history:



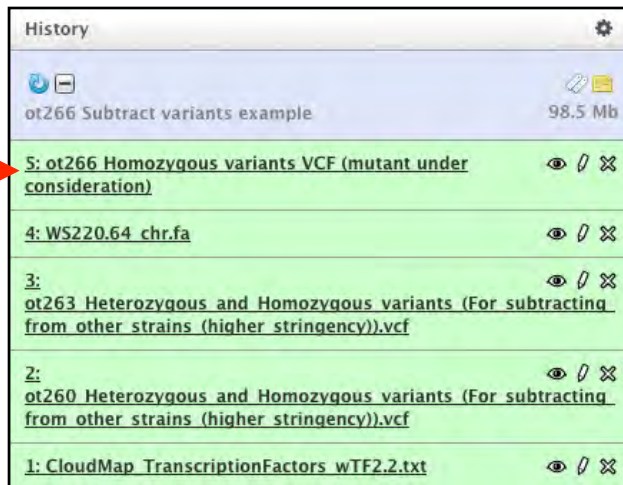
22) You will now see the **ot266 Homozygous Variants** (VCF) in your history. Click on the pencil icon to change the name of the file to add the *ot266* prefix.



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

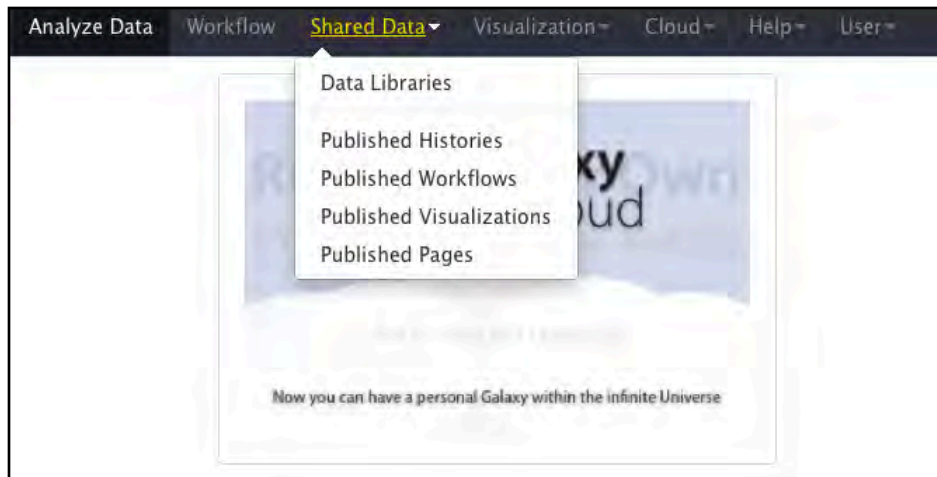


24) You will see that the file name has been updated:

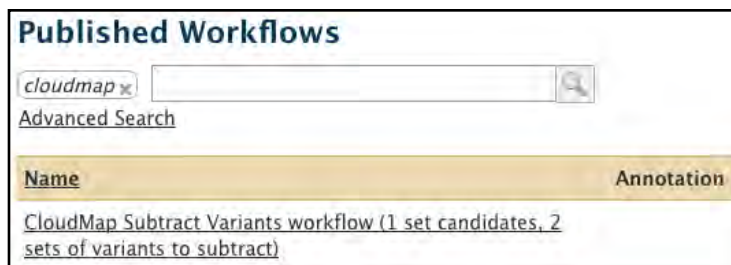


History	
ot266 Subtract variants example	98.5 Mb
5: <u>ot266 Homozygous variants VCF (mutant under consideration)</u>	👁️ ✂️ 🗑️
4: WS220.64 chr.fa	👁️ ✂️ 🗑️
3: <u>ot263 Heterozygous and Homozygous variants (For subtracting from other strains (higher stringency)).vcf</u>	👁️ ✂️ 🗑️
2: <u>ot260 Heterozygous and Homozygous variants (For subtracting from other strains (higher stringency)).vcf</u>	👁️ ✂️ 🗑️
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 x	🔍
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) **Import workflow**

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

Step	Annotation
Step 1: Input dataset Fasta reference <i>select at runtime</i>	
Step 2: Input dataset Candidate gene list <i>select at runtime</i>	
Step 3: Input dataset Variants for mutant under analysis (VCF file) (e.g. ot266 in CloudMap paper) <i>select at runtime</i>	
Step 4: Input dataset Variants to subtract 1 (VCF file) (e.g. ot260 or ot263 in CloudMap paper) <i>select at runtime</i>	
Step 5: Input dataset Variants to subtract 2 (VCF file) (e.g. ot260 or ot263 in CloudMap paper) <i>select at runtime</i>	
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:

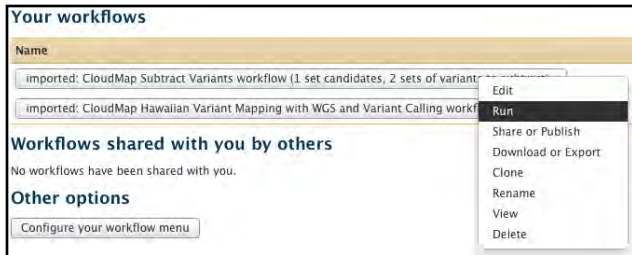
Workflow "CloudMap Subtract Variants workflow (1 set candidates, 2 sets of variants to subtract)" has been imported. You can [start using this workflow](#) or [return to the previous page](#).

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.

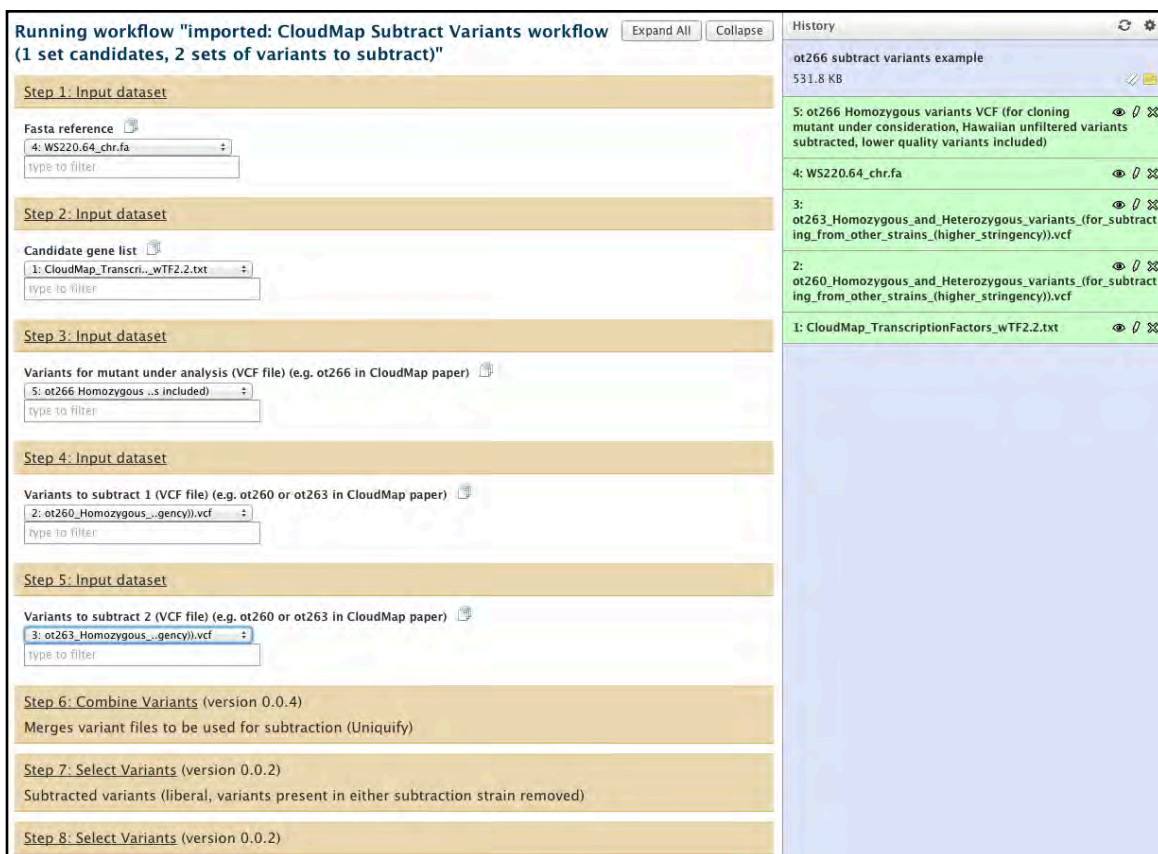
Your workflows

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

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



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.



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:

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

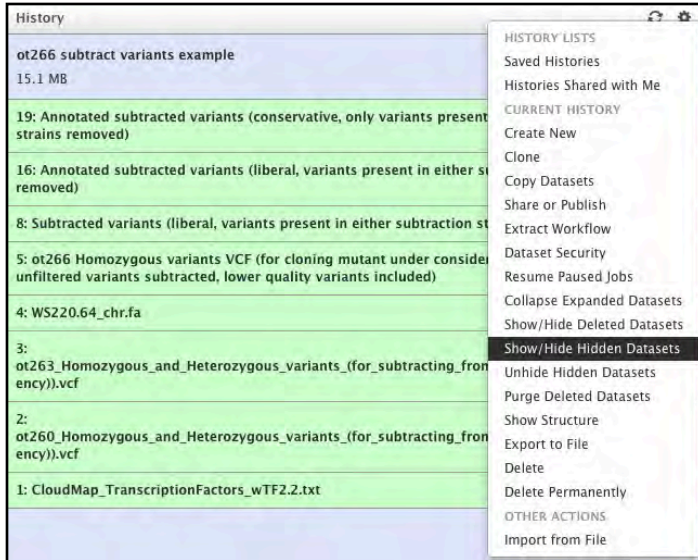
Successfully ran workflow "imported: CloudMap Subtract Variants workflow (1 set candidates, 2 sets of variants to subtract)". The following datasets have been added to the queue:

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

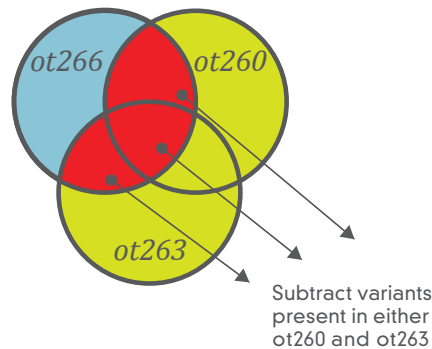
History

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:



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.



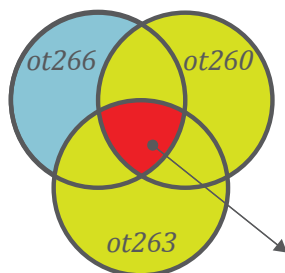
This file contains *ot266* homozygous variants after both homozygous and heterozygous variants present in **either** *ot260* **or** *ot263* have been subtracted. This file should be downloaded to be easily viewed in its entirety. The first several lines in any VCF file are header lines starting with “#” 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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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	chr1	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.

#	Chrom	Position	Reference	Change	Change_type	Homozygous	Quality	Coverage	Warnings	Gene_ID	Gene_name	Bin_type	Transcript_ID	Exon_ID	Exon_Rank	Effect	alt_AA	new_GH	codon/P	Codon_Num	Codon_Degr	CO5_score
1		62642	T	C	SNP	Hom	48.77	2		Y48G1C.4	ppp-1	protein_codi	Y48G1C.4			DOWNSTREAM: 8216 bases						
2		62642	T	C	SNP	Hom	48.77	2		Y48G1C.5	Y48G1C.5	protein_codi	Y48G1C.5			INTRON						3486
3		62642	T	C	SNP	Hom	48.77	2		Y48G1C.2	csk-1	protein_codi	Y48G1C.2.1			UPSTREAM: 9216 bases						
4		62642	T	C	SNP	Hom	48.77	2		Y48G1C.2	csk-1	protein_codi	Y48G1C.2.2			UPSTREAM: 9236 bases						
5		62642	T	C	SNP	Hom	48.77	2		Y48G1C.2	csk-1	protein_codi	Y48G1C.2.3			UPSTREAM: 9869 bases						
6		346149	T	A	SNP	Hom	85.77	3		Y48G1A.3	Y48G1A.3	protein_codi	Y48G1A.3			DOWNSTREAM: 8104 bases						
7		346149	T	A	SNP	Hom	85.77	3		Y48G1A.1	Y48G1A.1	protein_codi	Y48G1A.1			UPSTREAM: 2389 bases						
8		346149	T	A	SNP	Hom	85.77	3		Y48G1A.6	mbt-1	protein_codi	Y48G1A.6a			INTRON						1658
9		346149	T	A	SNP	Hom	85.77	3		Y48G1A.6	mbt-1	protein_codi	Y48G1A.6a			INTRON						1659
10		346149	T	A	SNP	Hom	85.77	3		Y48G1A.2	Y48G1A.2	protein_codi	Y48G1A.2.2			UPSTREAM: 1316 bases						
11		346149	T	A	SNP	Hom	85.77	3		Y48G1A.2	Y48G1A.2	protein_codi	Y48G1A.2.1			UPSTREAM: 1323 bases						
12		346149	T	A	SNP	Hom	85.77	3		Y48G1A.2	Y48G1A.2	protein_codi	Y48G1A.2.1			UPSTREAM: 1323 bases						
13		369870	C	T	SNP	Hom	48.08	2		R119.3	R119.3	protein_codi	R119.3.1			DOWNSTREAM: 3480 bases						
14		369870	C	T	SNP	Hom	48.08	2		R119.3	R119.3	protein_codi	R119.3.2			DOWNSTREAM: 3704 bases						
15		369870	C	T	SNP	Hom	48.08	2		R119.1	R119.1	protein_codi	R119.1			UPSTREAM: 5966 bases						
16		369870	C	T	SNP	Hom	48.08	2		R119.4	pan-50	protein_codi	R119.4.1			DOWNSTREAM: 7608 bases						
17		369870	C	T	SNP	Hom	48.08	2		R119.2	R119.2	protein_codi	R119.2			INTRON						1089
18		369870	C	T	SNP	Hom	48.08	2		R119.7	mpb-8	protein_codi	R119.7			DOWNSTREAM: 1358 bases						
19		369870	C	T	SNP	Hom	48.08	2		R119.4	pan-59	protein_codi	R119.4.2			DOWNSTREAM: 9377 bases						
20		369871	C	T	SNP	Hom	48.77	2		R119.3	R119.3	protein_codi	R119.3.1			DOWNSTREAM: 3479 bases						
21		369871	C	T	SNP	Hom	48.77	2		R119.3	R119.3	protein_codi	R119.3.2			DOWNSTREAM: 3703 bases						
22		369871	C	T	SNP	Hom	48.77	2		R119.1	R119.1	protein_codi	R119.1			UPSTREAM: 5967 bases						
23		369871	C	T	SNP	Hom	48.77	2		R119.4	pan-50	protein_codi	R119.4.1			DOWNSTREAM: 7607 bases						
24		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 **both** *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 **both** *ot260* and *ot263*.



Subtract variants present in both *ot260* and *ot263*

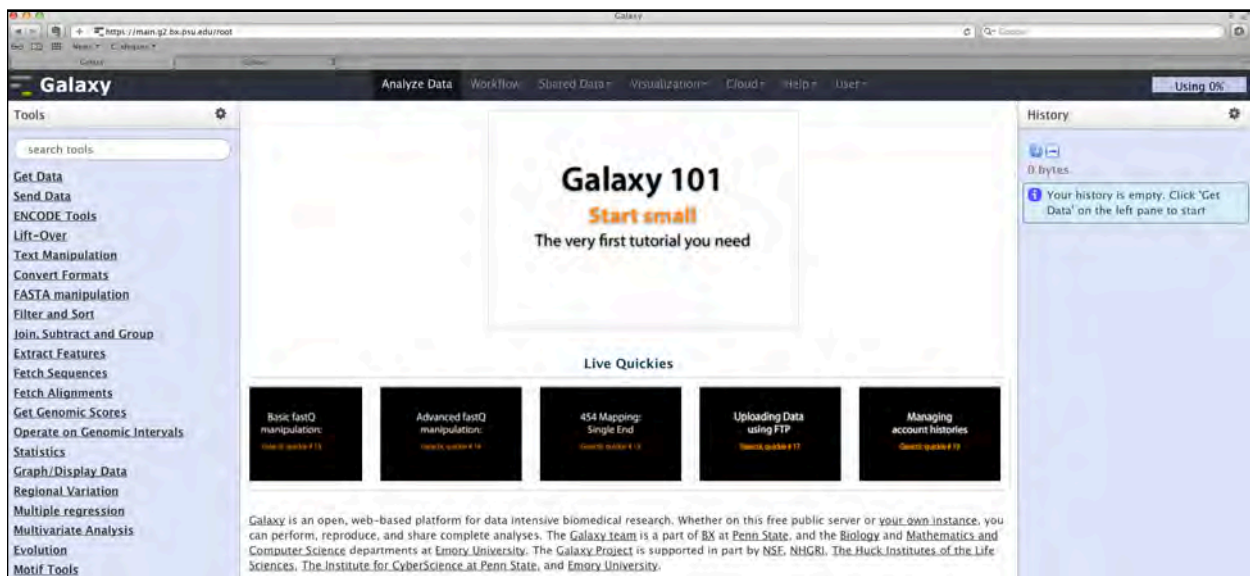
Note: We strongly suggest that users employ the ***Uncovered Region Subtraction*** workflow using the same strains (from their own screens) used in this workflow for variant subtraction. The general concept is shown in **Fig.5** of the CloudMap paper and is the same as used in this ***Subtract Variants***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

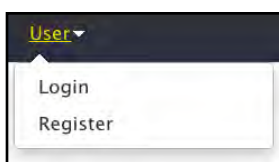
CloudMap Uncovered Region Subtraction workflow (using *ot266* Proof of Principle example from the CloudMap paper). A video version of this user guide is available at: <http://usegalaxy.org/cloudmap>. This workflow should be used downstream of either of the following workflows: **Hawaiian Variant Mapping 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>



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



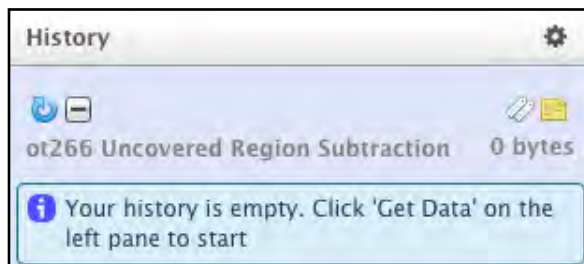
3) Once you are logged in using your email address, create a new history:



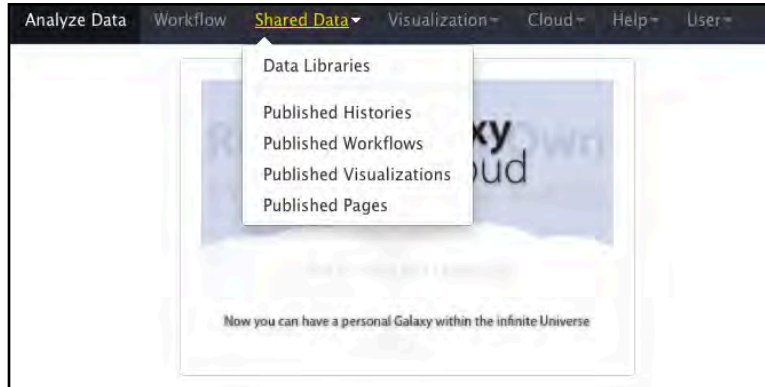
4) Now name that history:



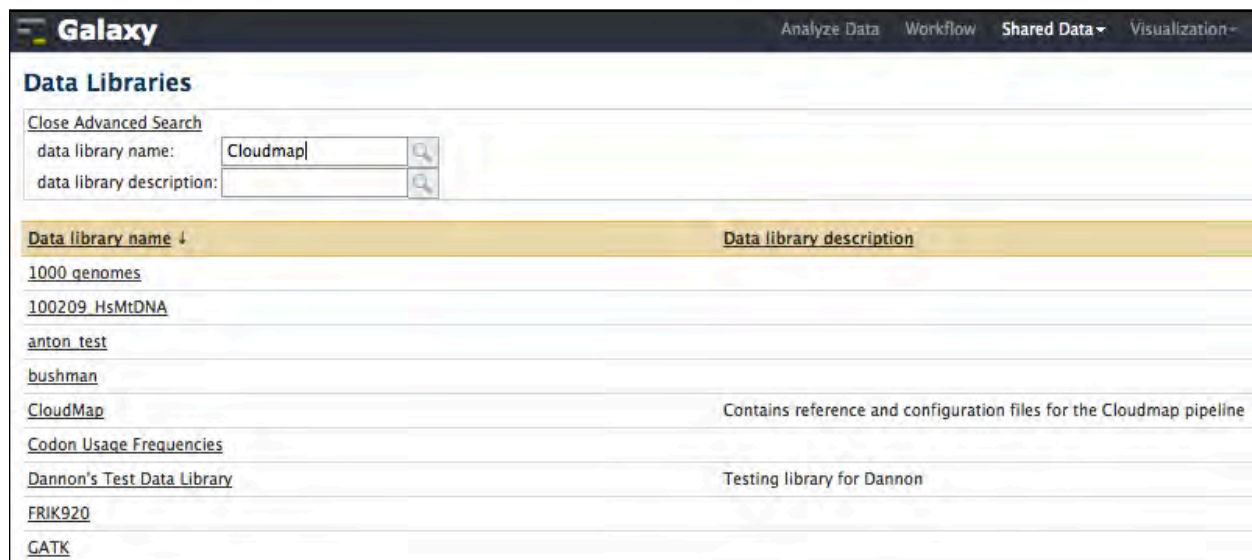
5) The history has been renamed.



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:



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.

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			
<input checked="" type="checkbox"/> ot260_Uncovered_regions.bed		bed	2012-07-17	1.0 Mb
<input checked="" type="checkbox"/> ot263_Uncovered_regions.bed		bed	2012-07-17	1.7 Mb
<input checked="" type="checkbox"/> 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.fastqanger	Sample FASTQ file for ot266 Proof of principle	fastqanger	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 WCS tool)	vcf	2012-06-11	4.3 Mb
WS220.64_chr.fa	WS220.64 genomic reference file	fasta	2012-06-11	87.6 Mb
SNP Mapping with WCS Data Other Species Config Files	Use these config files if you want to use the SNP Mapping with WCS 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:

Analyze Data Workflow Shared Data Visualization Cloud Help User

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

History

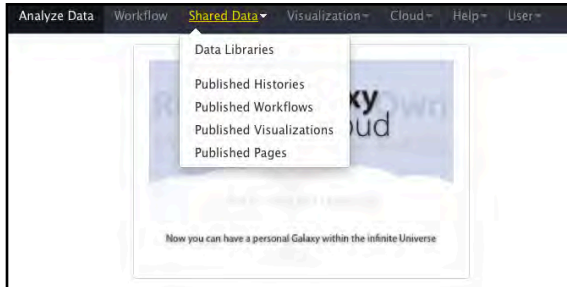
ot266 Uncovered Region Subtraction 0 bytes

3: ot266_Uncovered_regions.bed

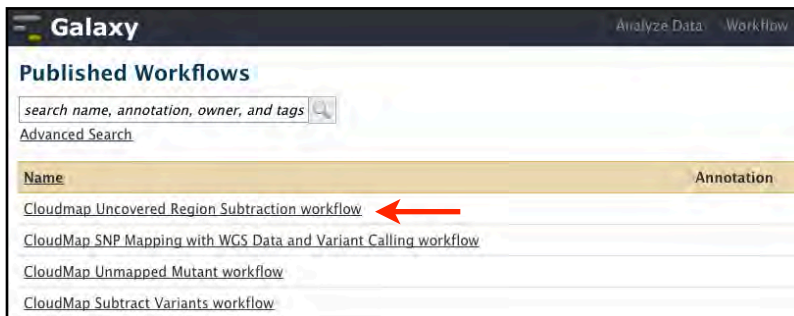
2: ot263_Uncovered_regions.bed

1: ot260_Uncovered_regions.bed

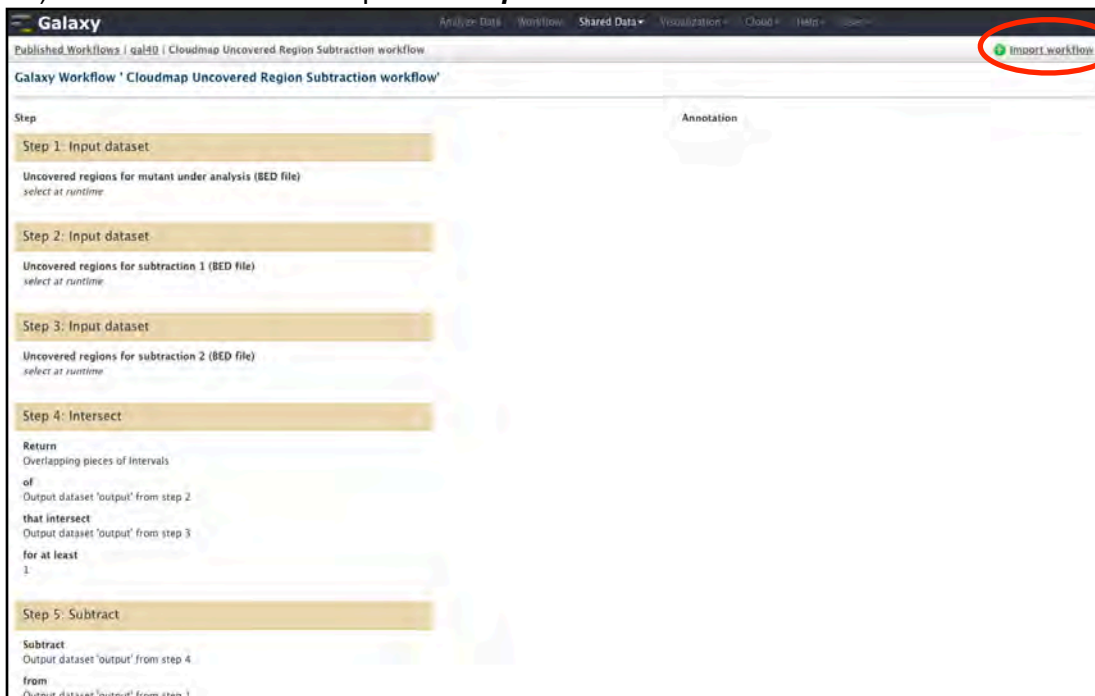
12) Now you have all the files ready to run the **Uncovered Region Subtraction** workflow. Click on the **Shared Data** → **Published Workflows** link at the top of the page:



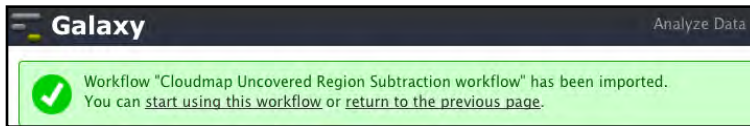
13) Select the **Uncovered Region Subtraction** workflow:



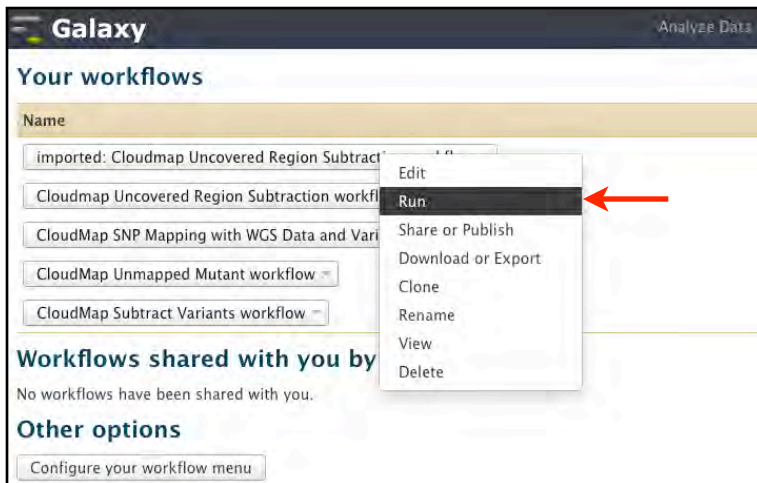
14) You will now have the option to **Import workflow**.



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



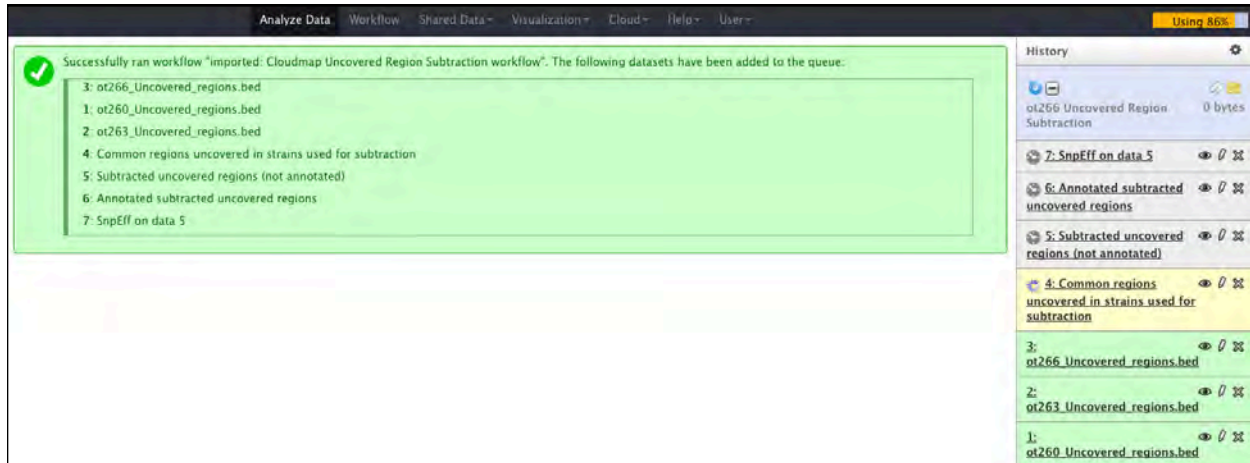
16) Click **Start using this workflow** and you will see that the workflow has been imported. From now on, you can easily access this workflow under the **Workflow** tab. Click on the workflow and select **Run**:



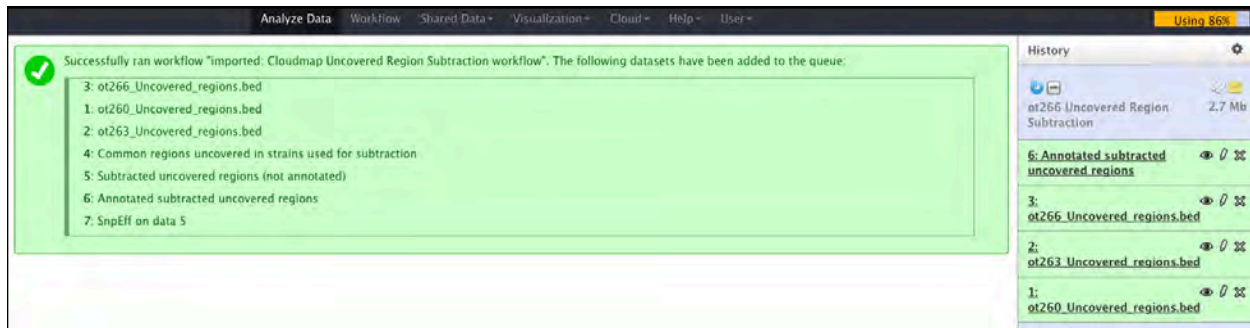
17) You will see all the steps in the workflow prior to running it. Make sure that each of the input fields corresponds to the appropriate file in your history. In our example, we want to subtract uncovered regions present in both *ot260* and *ot263* from the uncovered regions in *ot266*. Click **Run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



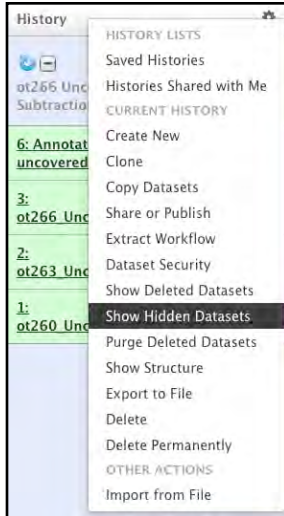
18) All of the automated functions have the appropriate default parameters configured, although experienced users may want to modify these prior to running. Once you are ready to run the workflow, press **Run Workflow** and the workflow will start (this step takes a minute or two to begin, be patient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



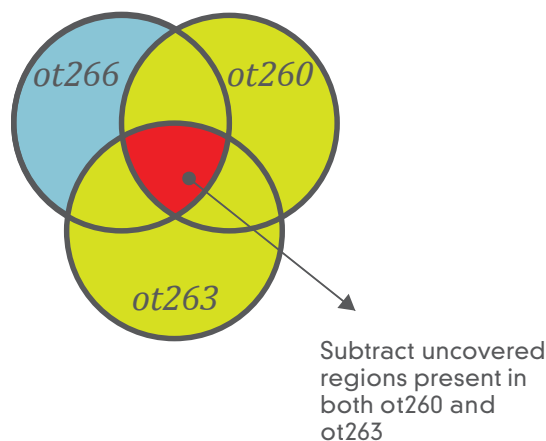
19) The workflow has finished running and you can view the resulting output:



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.



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	Homozygosity	Quality	Coverage	Warnings	Gene_ID	Gene_name	Biotype	Transcript_ID	Exon_ID	Exon_Rank	Effect
2	1		2646	2664		Interval			0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.4		UPSTREAM: 2859 bases
3	1		2646	2664		Interval			0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.6		UPSTREAM: 8972 bases
4	1		2646	2664		Interval			0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.3		UPSTREAM: 7767 bases
5	1		2646	2664		Interval			0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.2		UPSTREAM: 8840 bases
6	1		2646	2664		Interval			0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.1		UPSTREAM: 8853 bases
7	1		2646	2664		Interval			0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.5		UPSTREAM: 8853 bases
8	1		2646	2664		Interval			0	0		Y74C9A.3	Y74C9A.3	protein_codi	Y74C9A.3.1		DOWNSTREAM: 1473 bases
9	1		2646	2664		Interval			0	0		Y74C9A.3	Y74C9A.3	protein_codi	Y74C9A.3.2		DOWNSTREAM: 1575 bases
10	1		2646	2664		Interval			0	0		Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6		DOWNSTREAM: 1101 bases
11	1		3468	3482		Interval			0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.4		UPSTREAM: 8037 bases
12	1		3468	3482		Interval			0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.6		UPSTREAM: 8150 bases
13	1		3468	3482		Interval			0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.3		UPSTREAM: 6945 bases
14	1		3468	3482		Interval			0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.2		UPSTREAM: 8027 bases
15	1		3468	3482		Interval			0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.1		UPSTREAM: 8031 bases
16	1		3468	3482		Interval			0	0		Y74C9A.2	nlp-40	protein_codi	Y74C9A.2.5		UPSTREAM: 8031 bases
17	1		3468	3482		Interval			0	0		Y74C9A.3	Y74C9A.3	protein_codi	Y74C9A.3.1		DOWNSTREAM: 651 bases
18	1		3468	3482		Interval			0	0		Y74C9A.3	Y74C9A.3	protein_codi	Y74C9A.3.2		DOWNSTREAM: 753 bases
19	1		3468	3482		Interval			0	0		Y74C9A.6	Y74C9A.6	snoRNA	Y74C9A.6		DOWNSTREAM: 279 bases
20	1		3926	4018		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)

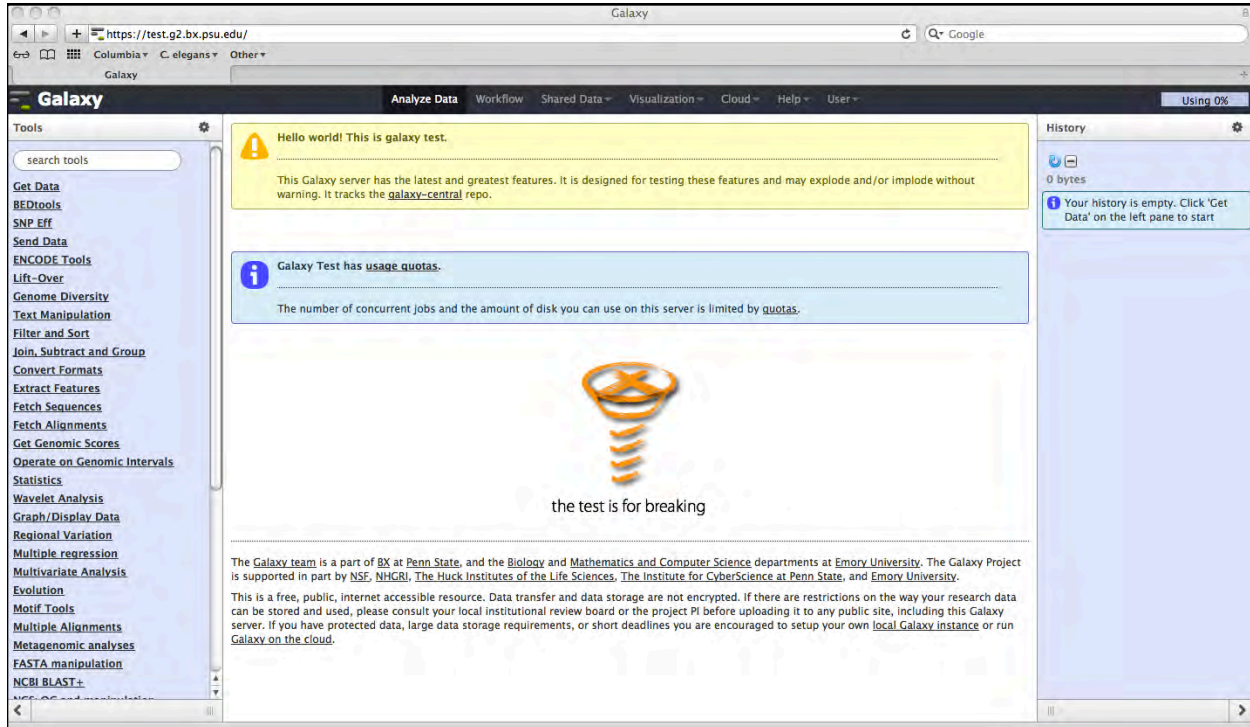
2) CONCATENATING FILES

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

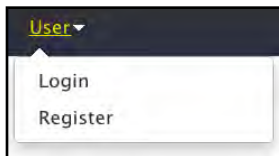
4) CONFIGURING THE *SNP MAPPING WITH WGS DATA* WORKFLOW TO SUPPORT SPECIES OTHER THAN *C.ELEGANS* AND *ARABIDOPSIS*:

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

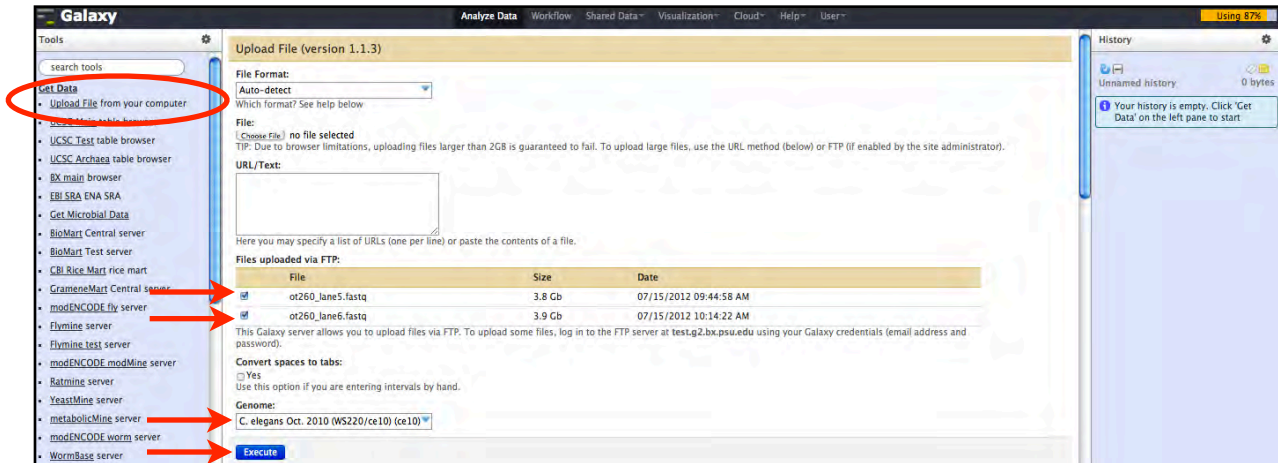
1) Navigate to the Galaxy site (<http://usegalaxy.org>)



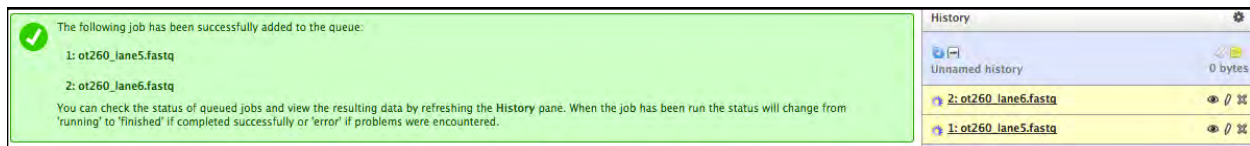
2) Register for an account or login if you already have an account:



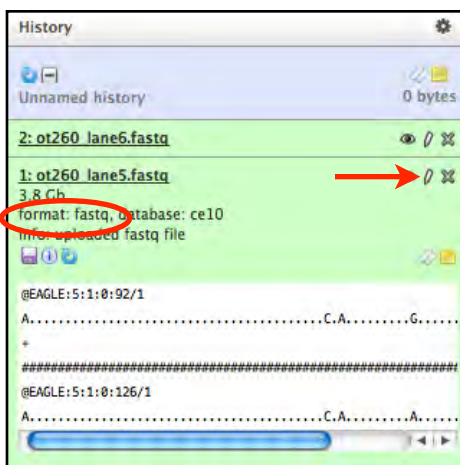
3) Once you are logged in using your email address, click on the **Get Data** link in the tools section on the left side of the screen. If the file you want to upload is < 2Gb, you can select the file through the **Choose file** link in the browser. Otherwise, you will need to upload your files via FTP (<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.



4) The files will be added to your history:



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:



- 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

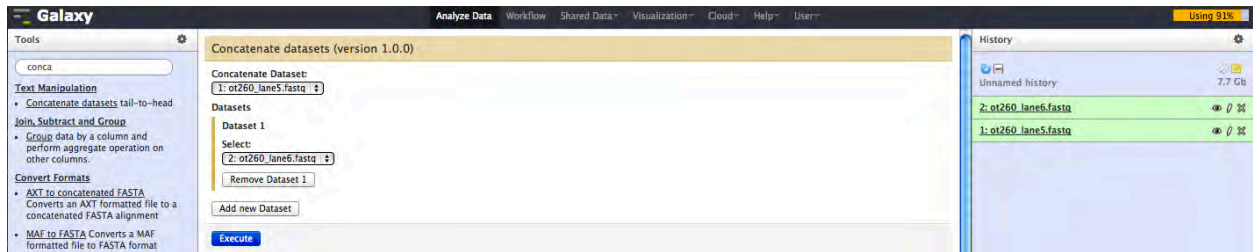
The screenshot shows the Galaxy interface for the FASTQ Groomer tool. The main panel is divided into three sections: 'Edit Attributes', 'Convert to new format', and 'Change data type'. In 'Edit Attributes', the name is 'ot260_lane5.fastq', info is 'uploaded fastq file', and the database is 'C. elegans Oct. 2010 (WS220/ce10)'. The 'Convert to new format' section has a dropdown menu set to 'Convert FASTQ files to seek locatic'. The 'Change data type' section has a dropdown menu set to 'fastqsanger'. The right sidebar shows the history with the dataset selected, and a preview of the FASTQ file content is visible.

- 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.

The screenshot shows a close-up of the Galaxy History panel. The dataset '1: ot260_lane5.fastq' is selected, and its format is shown as 'format: fastqsanger, database: ce10'. The text 'format: fastqsanger' is circled in red.

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:



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](#)

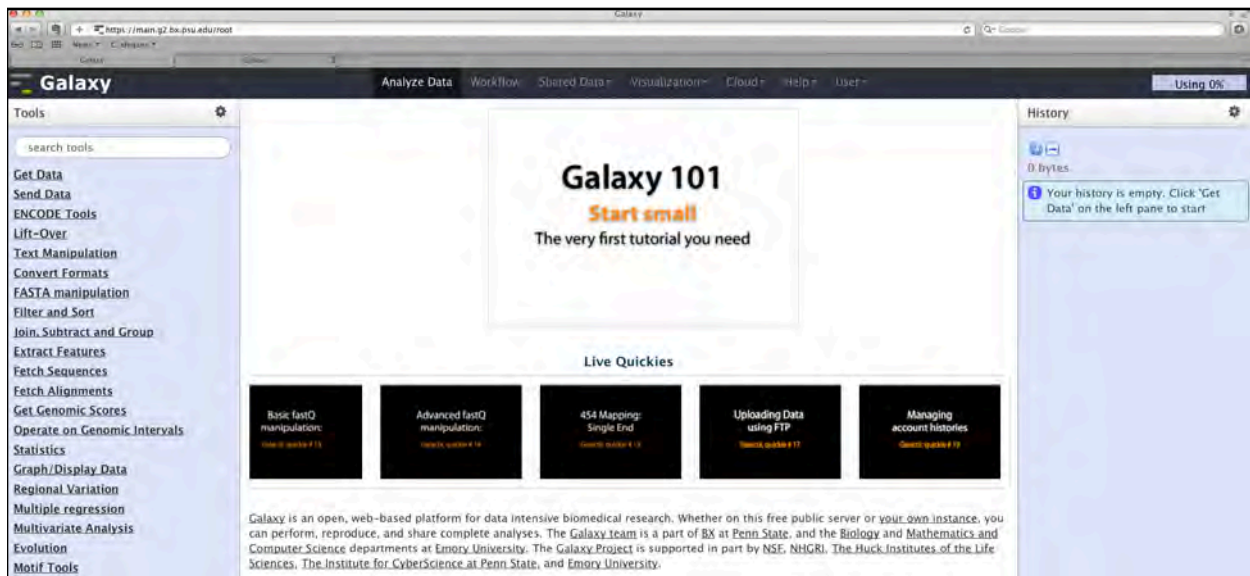
[Create workflow from scratch](#)

[Import workflow](#)

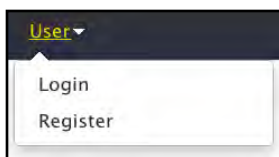
[Edit workflow](#)

[Convert workflow in a tool](#)

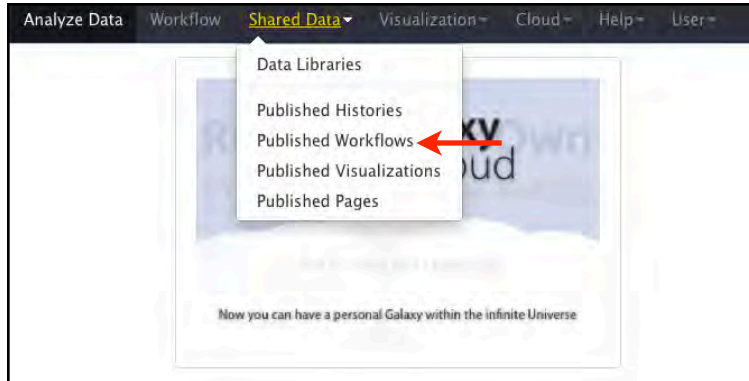
1) Let's assume that you haven't yet imported any CloudMap workflows. Navigate to <http://usegalaxy.org/>



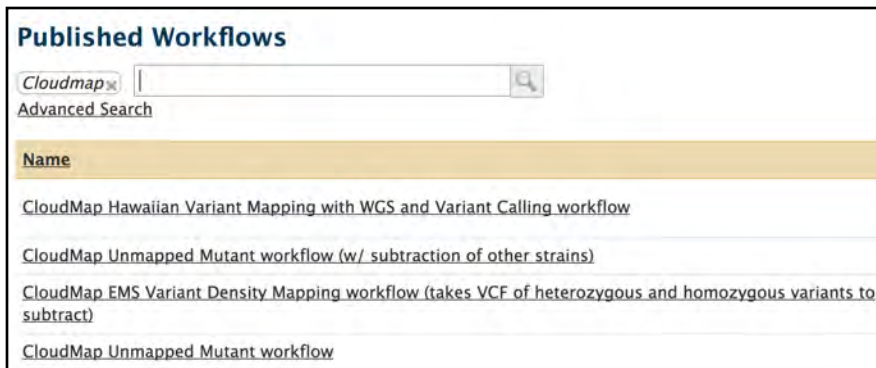
2) Register for an account or login if you already have an account:



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**.



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

The screenshot shows the Galaxy workflow editor interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. Below the navigation bar, there is a breadcrumb trail: 'Published Workflows | qm2123 | CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow'. A green '+ Import workflow' button is visible in the top right corner. The main content area is titled 'Galaxy Workflow ' CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow''. It displays a vertical sequence of six steps, each with a yellow bar and a description:

- 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:

A green message box with a white checkmark icon on the left. The text reads: '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** .

The screenshot shows the 'Your workflows' section in Galaxy. At the top right, there are two buttons: 'Create new workflow' and 'Upload or import workflow'. Below these buttons is a table with two columns: 'Name' and '# of Steps'. The table contains one entry:

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

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



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:

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

Options Details

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

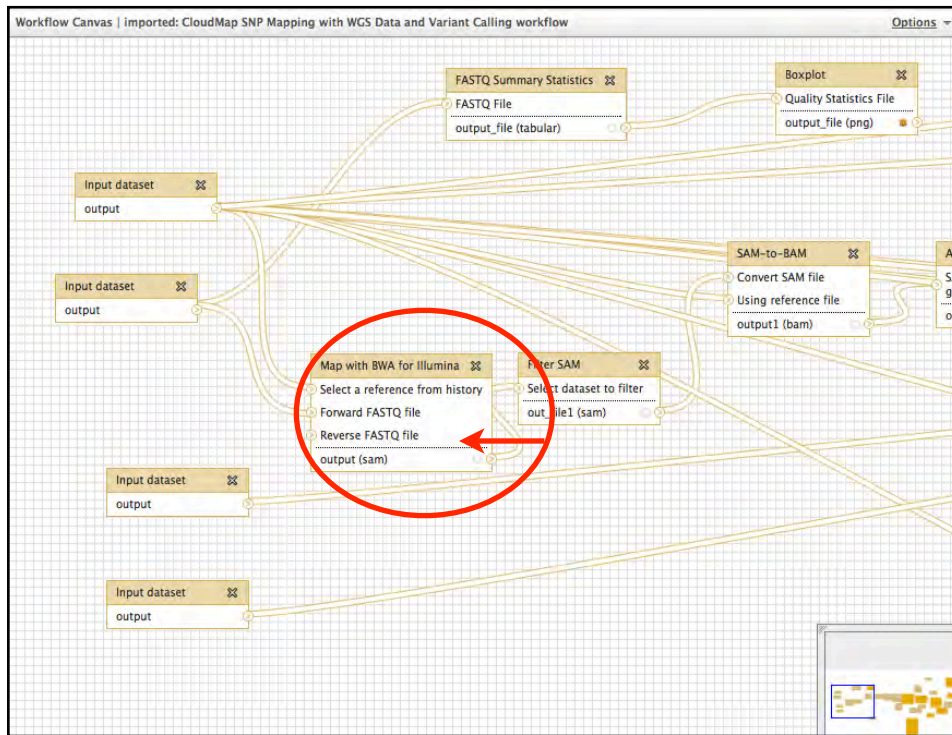
Disallow insertion/deletion within [value] bp towards the end (aln -i):
5

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

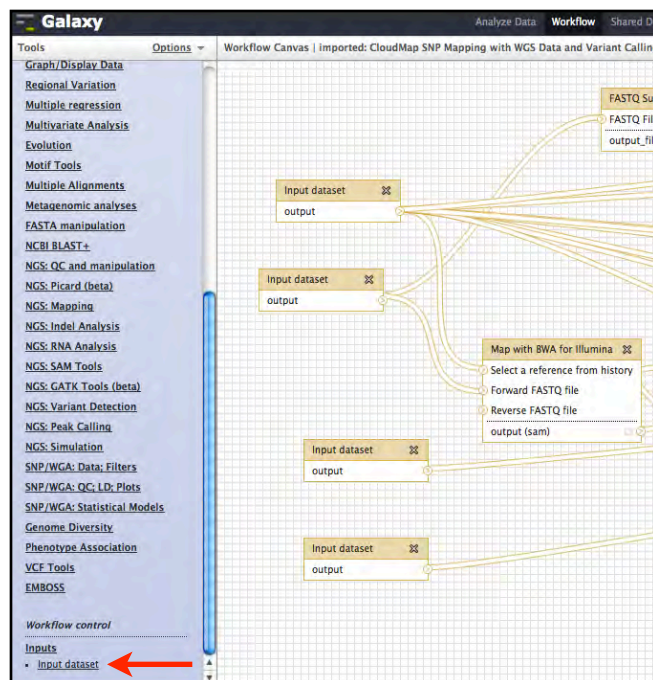
Maximum edit distance in the seed (aln -k):
2

Mismatch penalty (aln -M):

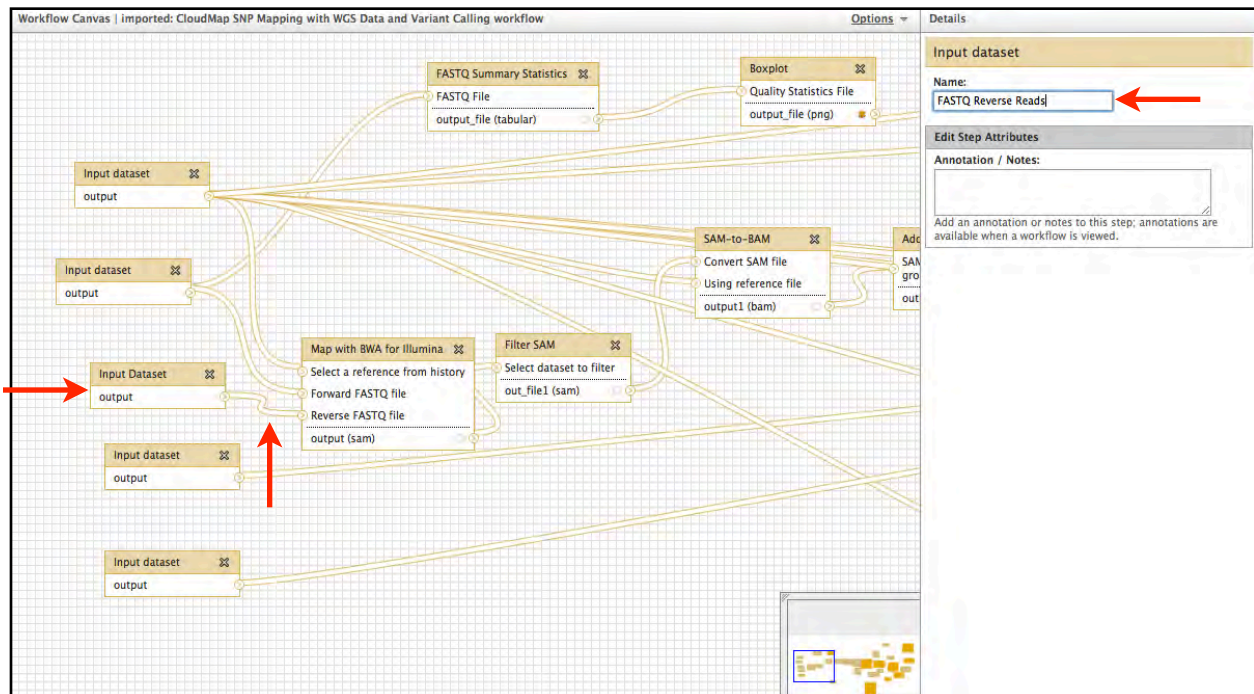
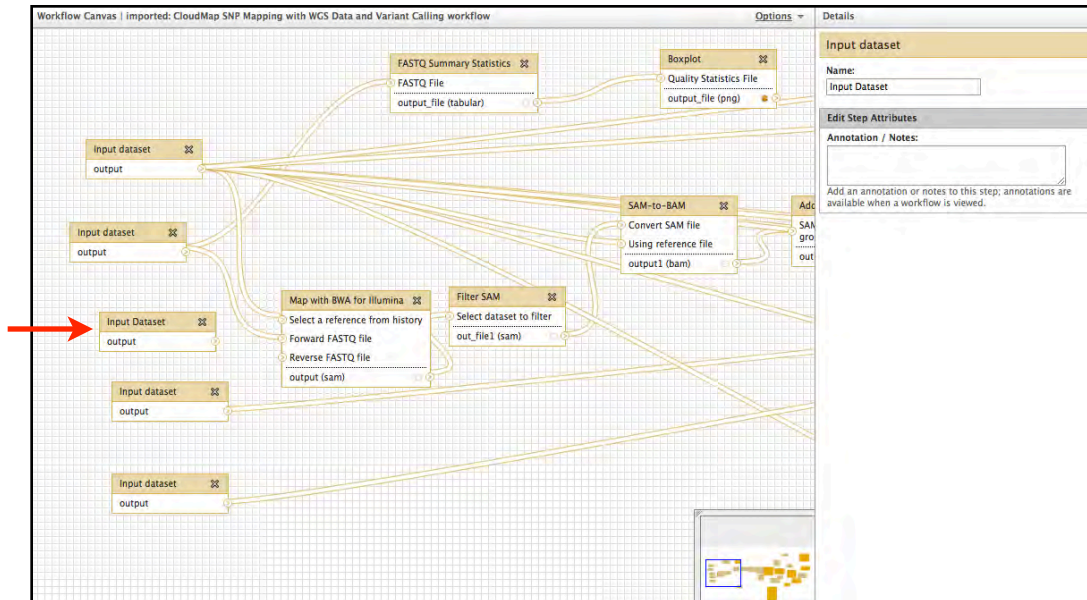
10) Once you select **paired-end** as the data type, the BWA mapping tool will now expect another input dataset.



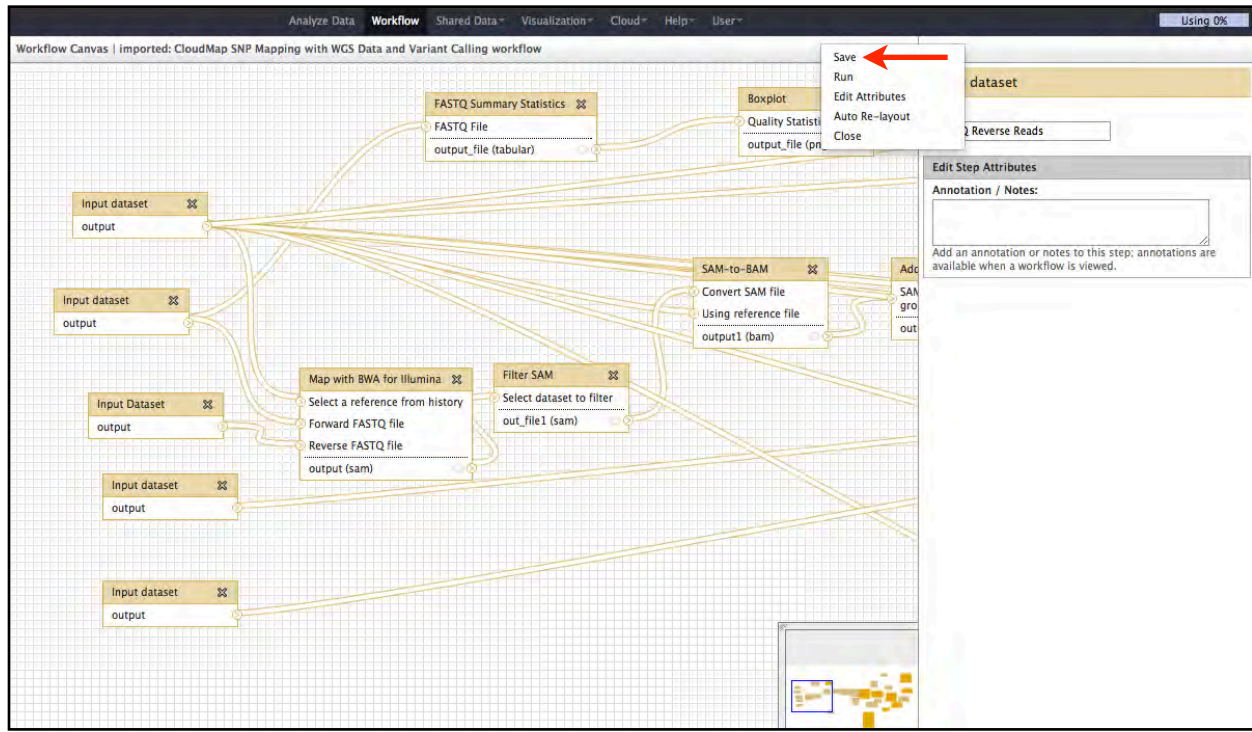
11) To add another input dataset, click **input dataset** under Galaxy tools:



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:



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



14) You can now run the modified workflow:



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.:

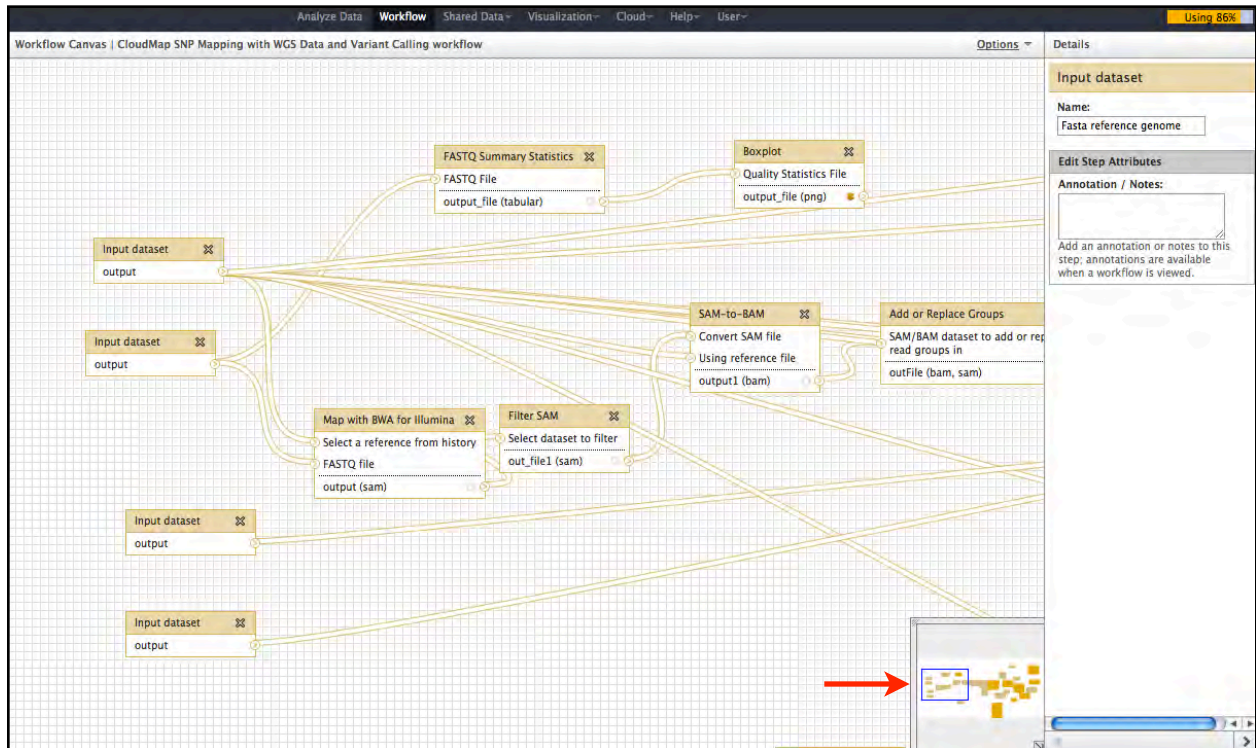
```
>CHROMOSOME_1  
>CHROM_1  
>1
```

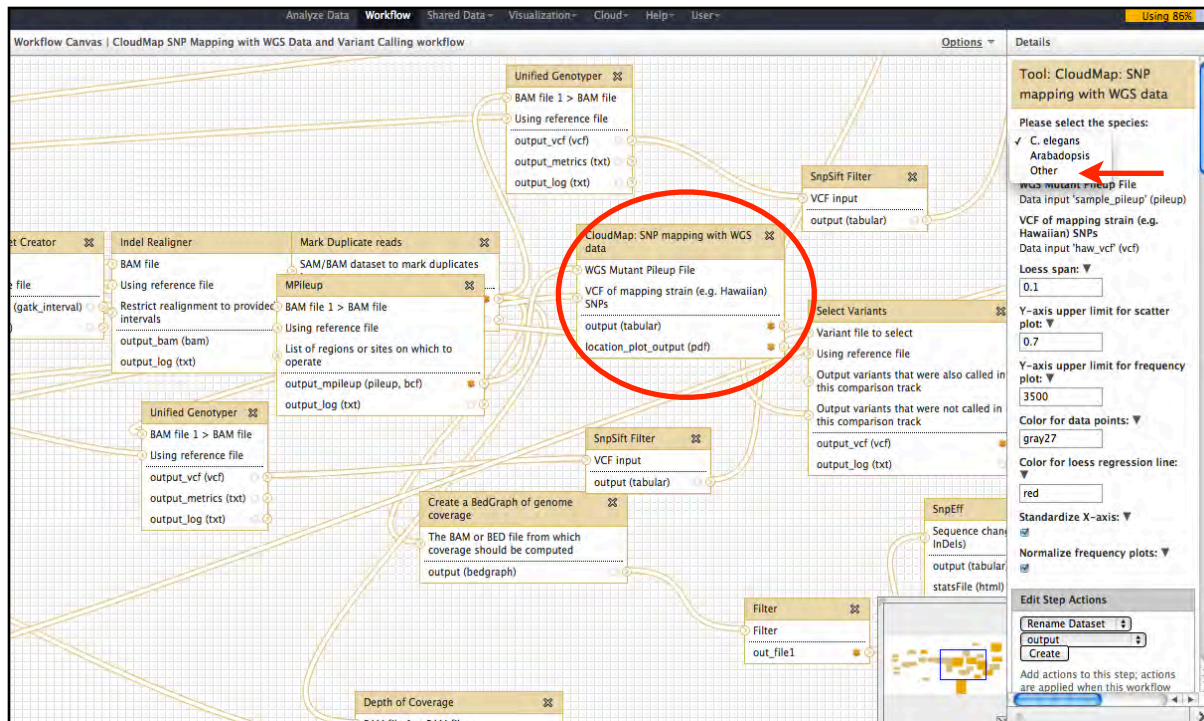
Sample *D.rerio* configuration file:

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

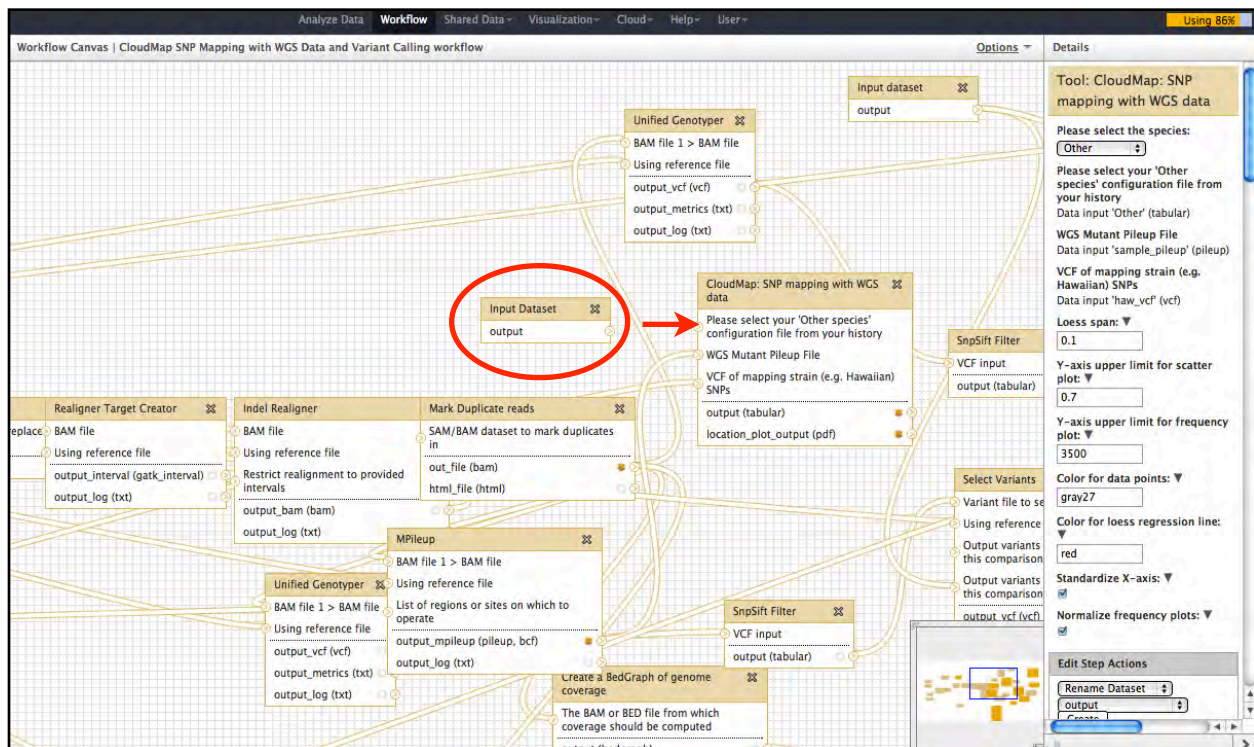
Please see more sample **Other species** configuration files in the CloudMap data library in the ***Hawaiian Variant Mapping with WGS Data Other Species Config Files*** folder.

- 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.

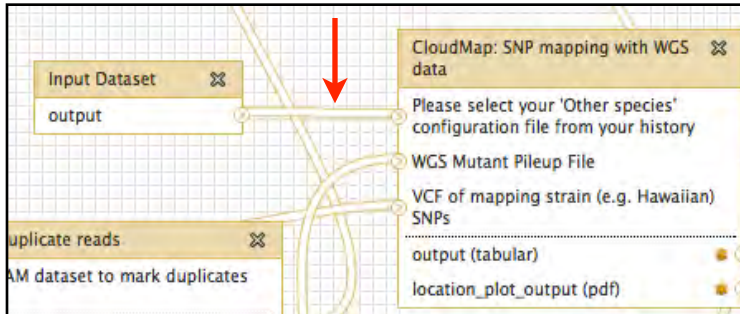




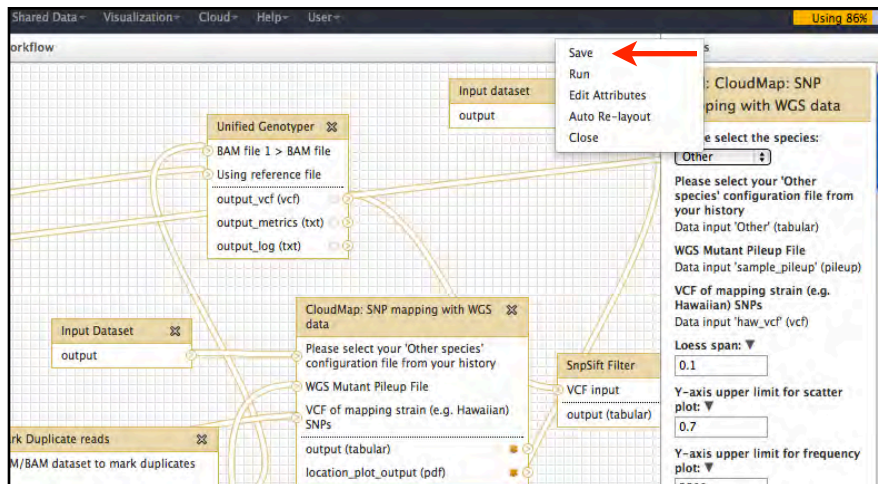
4) Select the **CloudMap Hawaiian Variant Mapping with WGS Data** tool, then select **Other** from species list.



- 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.



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



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.

Name	Annotation	Community Rating	Community Tags	Last Updated ↓
Cloudmap Uncovered Region Subtraction workflow		☆☆☆☆☆		- 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?

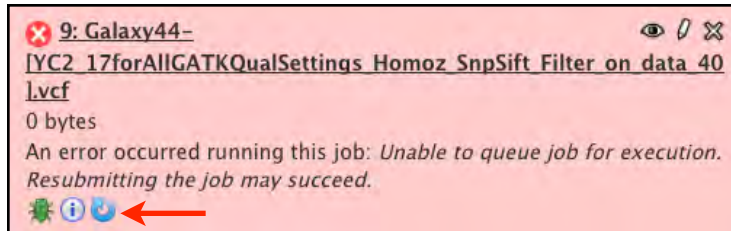
You can email us with your request at gm2123@columbia.edu or or38@columbia.edu. If you want to make the change yourself and run the tool locally, you can download the source code from the Galaxy Tool Shed at: <http://toolshed.g2.bx.psu.edu/>

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

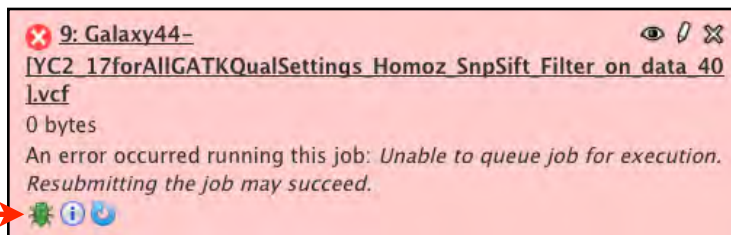
Galaxy questions:

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.

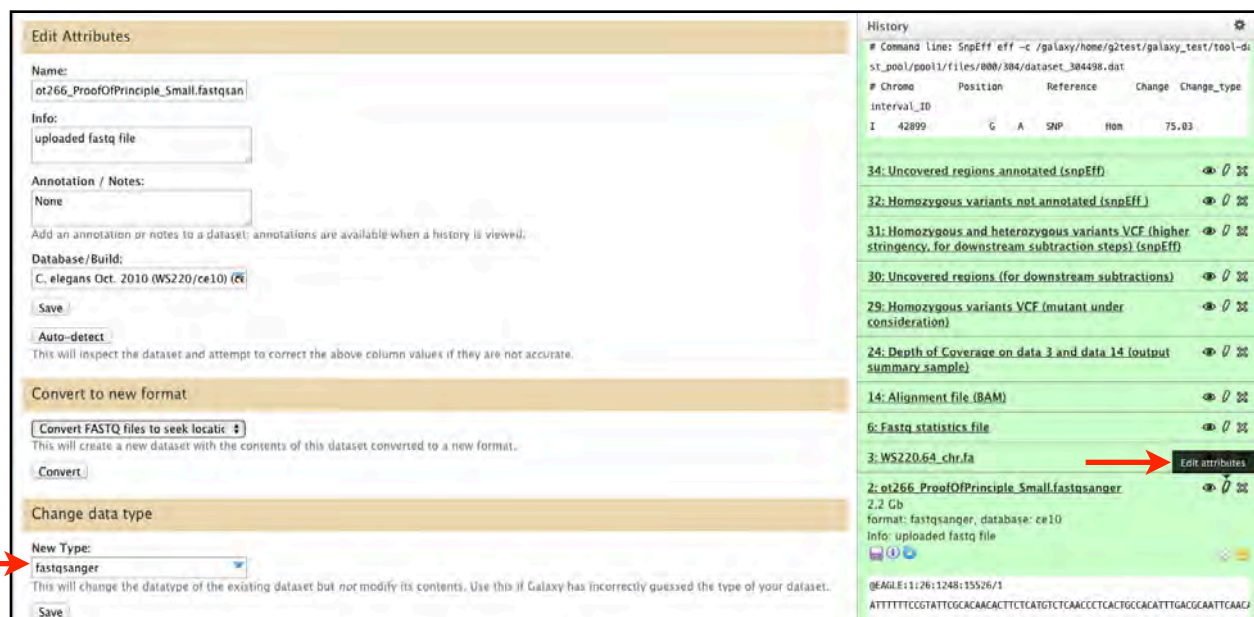


Failing that, submit a bug report to Galaxy by clicking on the bug icon.



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>



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.