

This user guide serves as a simplified, graphic version of the CloudMap paper for applicationoriented 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.

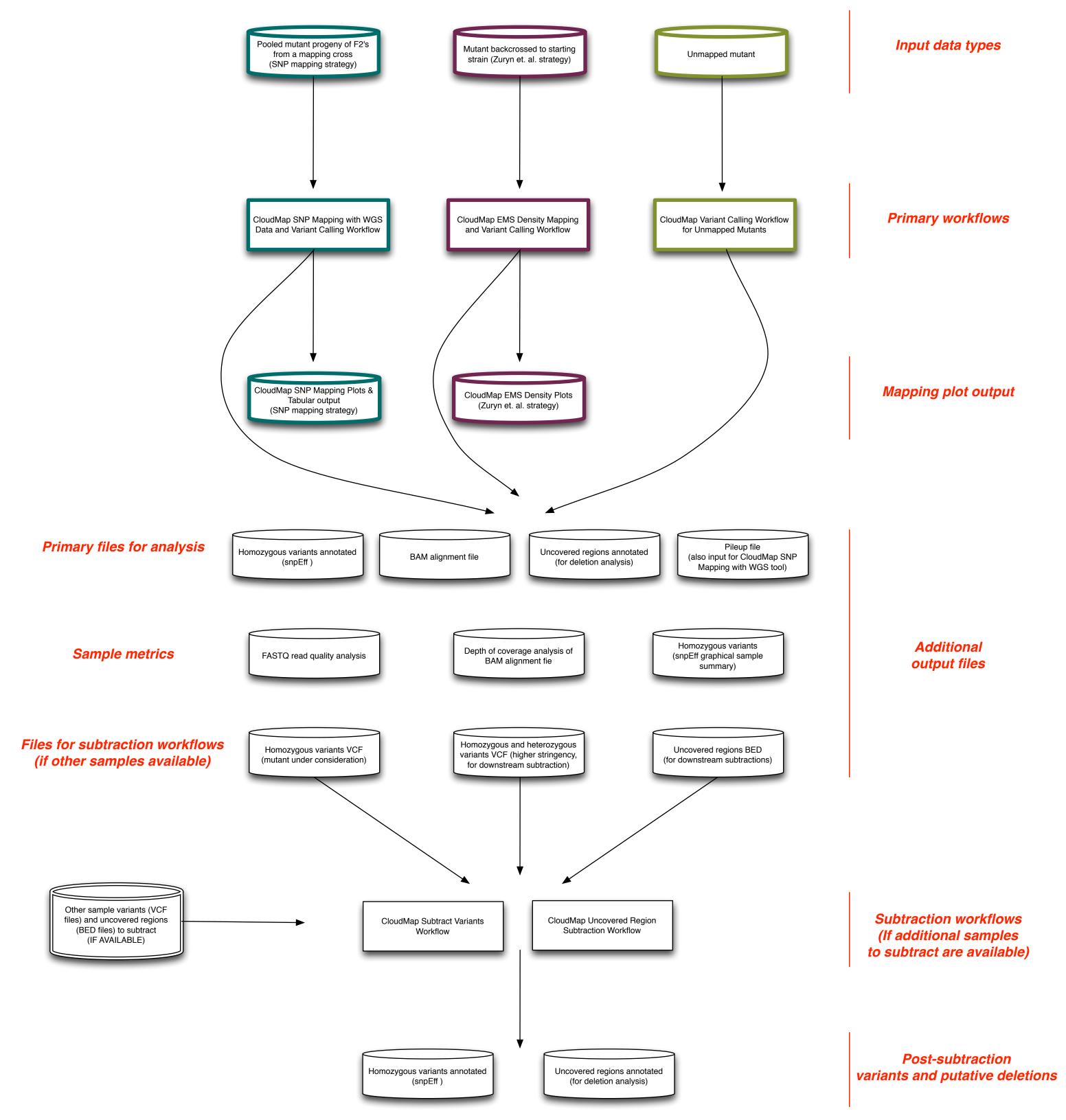
Helpful Galaxy screencasts are available at: http://wiki.g2.bx.psu.edu/Learn/Screencasts

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

*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: <u>http://usegalaxy.org/cloudmap</u>.

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 <a href="http://usegalaxy.org/cloudmap">http://usegalaxy.org/cloudmap</a>.



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

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:

Analyze Data	Workflow	Shared Data Visualization - Cloud - Help - User -
	R	Data Libraries Published Histories Published Workflows Published Visualizations Published Pages
	Ne	wyou can have a personal Galaxy within the infinite Universe

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

= Galaxy	Analyze Data Workflow Shared Data - Visualization-
Data Libraries	
Close Advanced Search       data library name:       data library description:	
Data library name 4	Data library description
1000 genomes	
100209_HsMtDNA	
anton_test	
bushman	
CloudMap	Contains reference and configuration files for the Cloudmap pipeline
Codon Usage Frequencies	
Dannon's Test Data Library	Testing library for Dannon
FRIK920	
GATK	

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

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 KE
CloudMap_ChromatinFactors.txt		tabular	2012-09-23	15.0 KB
CloudMap_TranscriptionFactors_wTF2.2.txt =		tabular	2012-09-23	19.2 KB
💿 🕨 📴 CloudMap EMS Variant Density Mapping 🕤	Use this dataset to try out the CloudMap EMS Variant Density Mapping tool			
🔄 🌱 🚤 CloudMap ot266 proof of principle dataset 🕤	Use these files to run the CloudMap ot266 proof of principle example			
📃 🌱 🍉 Hawaiian SNP reference files filtered (WS220.64) -	Filtered set of Hawaiian SNP variants (used by CloudMap SNP Mapping with WGS tool)			
HA_SNPs_Filtered_103346Variants_WS220.vcf -		vcf	2012-10-09	4.3 MB
🔲 🌱 📴 Hawaiian SNP reference files unfiltered (WS220.64) -	Unfiltered set of Hawailan SNP variants (used by CloudMap SNP Mapping with WGS tool)			
HA_SNPS_Unfiltered_112061Variants_WS220.64_chr.vcf -		vcf	2012-09-23	4.6 MB
ot266_ProofOfPrinciple_Small.fastqsanger =	None	fastqsanger	2012-09-23	2.2 GB
WS220.64_chr.fa -		fasta	2012-09-23	97.6 MB
🗇 🕨 📴 CloudMap user guides 🕤	Detailed guides for using the CloudMap pipeline			
b 20 ot260 and ot263 BEDs for uncovered subtraction =	Use these BED files for the CloudMap ot266 proof of principle for uncovered region subtraction			
Description - 1 Provide August Aug	Use these VCF files for the CloudMap ot266 proof of principle variant subtraction			

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:



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

Shared Data Analyze Data Workflow Admin Help User

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

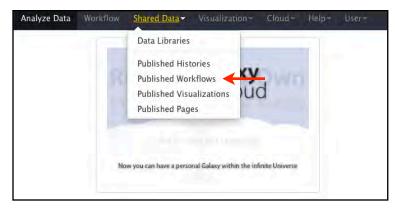
History	<	¥
Unnamed history	2.2 0	B
5: WS220.64_chr.fa	• 0 8	×
4: ot266_ProofOfPrinciple_Small.fastqsanger	• 0 5	×
3: HA_SNPS_Unfiltered_112061Variants_WS220.64_d	O S	*
2: HA_SNPs_Filtered_103346Variants_WS220.vcf		3
1: CloudMap_TranscriptionFactors_wTF2.2.txt		×



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

History	\$
<u>0</u>	42
ot266	2.2 GB
5: WS220.64_chr.fa	• 0 %
4: ot266_ProofOfPrinciple_Small.fastqsanger	•0%
3: HA_SNPS_Unfiltered_112061Variants_WS220.64_0	● Ø 器 chr.vcf
2: HA_SNPs_Filtered_103346Variants_WS220.vcf	
1: CloudMap_TranscriptionFactors_wTF2.2.txt	• 0 %

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



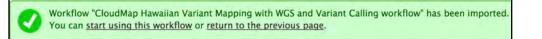
11) Use the search term "CloudMap" to view the automated workflows. Select the CloudMap Hawaiian Variant Mapping with WGS Data and Variant Calling workflow.

Published Workflows	
Cloudmap	E4,
Advanced Search	
Name	
CloudMap Hawaiian Variant Mapping with	WGS and Variant Calling workflow
CloudMap Unmapped Mutant workflow (w/	/ subtraction of other strains)
CloudMap EMS Variant Density Mapping we subtract)	orkflow (takes VCF of heterozygous and homozygous variants to
CloudMap Unmapped Mutant workflow	

12) You will now have the option to Import workflow



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



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

Your workflows	Create new workflow
Name	# of Steps
imported: CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow -	29

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

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

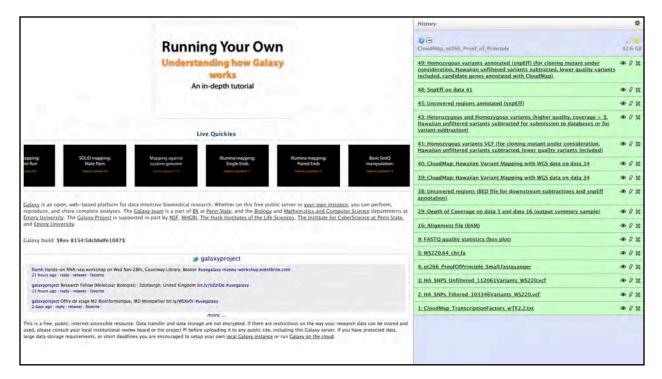
16) You will see all the steps in the workflow prior to running it. Make sure that each of the input fields corresponds to the appropriate file in your history.

Running workflow "imported: CloudMap Hawaiian Variant Mapping with WGS and Expand All Collapse	History	0
Variant Calling workflow"	DE	
Step 1: Input dataset	01266	2.2 68
Filtered mapping strain VCF (e.g. 103,346 Hawaiian SNPs) 5 2: HA SNPs Filtered, .s. WS220.vt =	5: W5220.64 chr.fa 4: ot266 ProofOfPrinciple Small.fastqsanger	000 X
Note to title:	3: HA SNPS Unfiltered 112061Variants WS220.64	a 0 1
Step 2: Input dataset	Z: HA SNPs Filtered 103346Variants WS220.vcf	@ Ø 3
Fasta reference genome S 5: WS220.64, cbr.fa 1 type to Tilin/	1: CloudMap TranscriptionFactors wTF2.2.txt	@ 0 %
Step 3: Input dataset		
FASTQ reads (Sanger format) 4: 01266.ProofO/Prin,fastqsanger 1 Sype to There		
Step 4: Input dataset		
Candidate gene list (e.g. transcription factors)		
Step 5: Input dataset		
Unfiltered mapping strain VCF (e.g. 112,061 Hawaiian SNPs) 3: HA_SNPS_Unfiltere64_chr.vcf z type to filter		
Step 5: Map with BWA for Illumina (version 1.2.3)		
Step 7: FASTQ Summary Statistics (version 1.0.0)		
Step 8: Filter SAM (version 1.0.0)		
Step 9: Boxplot (version 1.0.0)		
Step 10: SAM-to-BAM (version 1.1.2)		
From 11. Add on Benlann Common Austrian 1 FC 01		>

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:

ccessfully ran workflow "imported: CloudMap Hawaian Variant Mapping with WGS and Variant Calling workflow". The following datasets have been added to the queue:	History	
2: HA_SNPs_Filtered_103346Variants_WS220.vcf	26: Unified Genotyper on data 5 and data 16 (log)	
5: W5220.64, chr.fa		
4: or266_ProofO@Principle_Small.fastqsanger	25: Unified Genotyper on data 5 and data 16 (metrics)	
1: CloudMap_TranscriptionFactors_wTF2.2.bxt	24: Unified Genotyper on data 5 and data 16 (VCF)	
3: HA_SNPS_Linflikered_112061Variants_WS220.64_chr.vcf	3 23: Unified Genotyper on data 5 and data 16 (log)	
6. Map with BWA for Illumina on data 4 and data 5: mapped reads		
7: FASTQ Summary Statistics on data 4	22: Unified Genotyper on data 5 and data 16 (metrics)	
8: Filter SAM on data 6 9: FASTD quality statistics (box plot)	C 21: Unified Genotyper on data 5 and data 16 (VCF)	
9: PASING quarky statistics (dox prot) 10: SAM-to-RAM on data 5 and data 8: converted BAM	20: Unified Genotyper on data 5, data 16, and data 2 (log)	
11: Add or Replace Groups on data 10: bam with read groups replaced		
12: Realigner Target Creator on data 5 and data 11 (CATK intervals)	19: Unified Genotyper on data 5, data 16, and data 2 (metrics)	æ Ø 1
13: Realigner Target Creator on data 5 and data 11 (log)	18: Unified Genotyper on data 5, data 16, and data 2 (VCF)	.001
14 Indel Realigner on data 5, data 12, and data 11 (BAM)	3 17: MarkDups Alignment file (8AM).html	001
15; Indel Realigner on data 5, data 12, and data 11 (log)	3 16: Alignment file (BAM)	
16: Alignment file (BAM)	S 10: Ansimment mill (BAM)	00.0
17: MarkDups_Alignment file (BAM).html	15: Indel Realigner on data 5, data 12, and data 11 (log)	
18: Unified Genotyper on data 5, data 16, and data 2 (NCF)	3 14: Indel Realigner on data 5, data 12, and data 11 (BAM)	
19: Unified Genotyper on data 5, data 16, and data 2 (metrics)	3 13: Realigner Target Creator on data 5 and data 11 (log)	a 0 1
20. Unified Genotyper on data 5, data 16, and data 2 (log)	S 13: Realigner Larget Creator on data 5 and data 11 Diogr	ap (/ )
11: Unified Genotyper on data 5 and data 16 (VCF)	C 12: Realigner Target Creator on data 5 and data 11 (GATK intervals)	
22: Unified Genotyper on data 5 and data 16 (metrics) 23: Unified Genotyper on data 5 and data 16 (log)	C 11: Add or Replace Groups on data 10: bam with read groups replaced	4
24. Unified Cenotyper on data 5 and data 16 (VCF)	C 10: SAM-to-BAM on data 5 and data 8: converted BAM	
25: Unified Genotyper on data 5 and data 16 (metrics)		
26: Unified Cenotyper on data 5 and data 16 (log)	9: FASTQ quality statistics (box plot)	
27: Alignment file (BAN) (Cenome Coverage BedGraph)	S: Filter SAM on data 6	
28: Depth of Coverage on data 5 and data 16 (per locus coverage)	E 7: FASTQ Summary Statistics on data 4	
29: Depth of Coverage on data 5 and data 16 (output summary sample)		
30: Depth of Coverage on data 5 and data 16 (output statistics sample)	5: Map with BWA for Illumina on data 4 and data 5: mapped reads	
31: Depth of Coverage on data 5 and data 16 (output cumulative coverage counts sample)	5: W5220,64 chr.fa	
32: Depth of Coverage on data 5 and data 16 (output cumulative coverage proportions sample)	4: ot266 ProofOfPrinciple Small.fastqsanger	0.0
33: Depth of Coverage on data 5 and data 16 (log)	3: HA SNPS Unfiltered 112061Variants WS220.64 chr.vcf	- 0
34 Variants at mapping strain variant positions (e.g. Hawaiian) with read depth > 0 (for plotting)	S. HA. SNYS. Untillered. 112061Variants. HS220.64. Chr.yd	400
35: Select Variants on data 5 and data 18 (log)	2: HA SNPs Filtered 103346Variants WS220.vcf	
36: Homozygous variants VCF (mutant under consideration, lower quality variants for cloning, Hawaiian unflitered SNPs subtracted) 37: Homozygous and heterozygous variants VCF (higher stringency, for downstream subtraction steps) (snpEH)	1: CloudMap TranscriptionFactors wTF2.2.txt	-
37. Homozygous and interrotygous warants vc.r (higher somgency, for downstream substaction steps) (sinplif) 38. Uncovered regions (IED file for downstream subtractions and sinpliff annotation)		

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



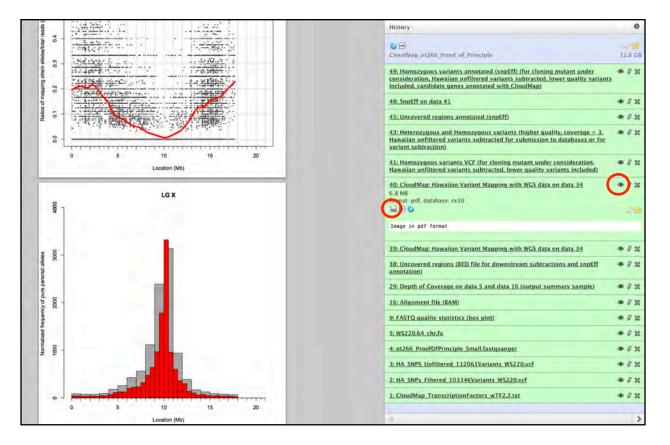
19) You will notice that while over 40 output files were generated during the course of the workflow (output files are sequentially numbered), only some output files remain visible while others are hidden. The visible files are most important for analysis of the mutant under consideration or downstream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

History	HISTORY LISTS	
CloudMap_01266_Proof_of_Principle	Saved Histories Histories Shared with M CURRENT HISTORY	
49: Homozvgous variants annotated (snpEff) (for cloning mu consideration, Hawaiian unfiltered variants subtracted, lowe included, candidate genes annotated with CloudMap)	Create New Clone Copy Datasets	
48: SnpEff on data 41	Share or Publish	
45: Uncovered regions annotated (snpEff)	Extract Workflow Dataset Security	
43: Heterozygous and Homozygous variants (higher guality, Hawaiian unfiltered variants subtracted for submission to d variant subtraction)	Show Deleted Datasets Show Hidden Datasets Purge Deleted Datasets	
41: Homozygous variants VCF (for cloning mutant under cor Hawaiian unfiltered variants subtracted, lower quality variar	Show Structure Export to File	
40: CloudMap: Hawaiian Variant Mapping with WGS data on e	Delete	
39: CloudMap: Hawaiian Variant Mapping with WGS data on (	Delete Permanently OTHER ACTIONS	
38: Uncovered regions (BED file for downstream subtraction annotation)	Import from File	
29: Depth of Coverage on data 5 and data 16 (output summa	ry sample) 👁 0 👷	
16: Alignment file (BAM)	• 0 %	
9: FASTQ quality statistics (box plot)	• 0 %	
5: W5220.64 chr.fa	• 0 %	
4: ot266 ProofOfPrinciple Small.fastqsanger		
3: HA_SNPS_Unfiltered_112061Variants_WS220.vcf		
2: HA_SNPs_Filtered_103346Variants_WS220.vcf	• 0 %	
1: CloudMap_TranscriptionFactors_wTF2.2.txt	• 0 23	

#### 20) You may unhide any files that are hidden:



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



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

CloudMap: Hawaiian Variant Mapping with WGS data (version 1.0.0)	History	0
Please select the species: (C. elegans = 2)	CloudMap_o(265_Proof_of_Principle	00 12.6 GB
WCS Mutant VCF File: [34: (hidden) Variants at mappil: plotting] : WCS Mutant VCF File from peoled F2 mutants that have been crussed to a mapping strain. The VCF should contain data from only mapping strain (e.g.	49: Homozygous variants annotated (snpEff) (for cloning mutant under consideration, Hawaiian unflitered variants subtracted, lower quality variants included, candidate genes annotated with CloudMap)	• 0 35
Hawailan) SNP positions	48: SnpEff on data 41	
0,1	45: Uncovered regions annotated (snpEff)	• 0 2
Parameter that controls the degree of data smoothing. Y-asity upper limit for scatter plot: 0.7	43: Heterozygous and Homozygous variants (higher quality, coverage ≥ 3, Hawaiian unfiltered variants subtracted for submission to databases or for variant subtraction)	• 0 %
V-axis upper limit for frequency plot: 4000	41: Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower guality variants included)	• 0 %
Color for data points: gray27 See below for list of suppored colors	40: Cloud Map: Hawaiian Variant Mapping with WGS data on data 34 6.8 MB format: pdf, database: cel0	• 0 2
Color for losss regression line: red See below for hat of supported colors	Run this job again at	
Standardize X-axis:	39: CloudMap: Hawaiian Variant Mapping with WGS data on data 34	æ Ø %
Scatter plots and frequency plots from separate chromosomes will have uniform X-avis spacing for comparison Normalize frequency plots:	38: Uncovered regions (BED file for downstream subtractions and snpEff annotation)	• 0 %
Frequency plots of oure parental allele counts will be normalized according to the equation in Fig. 28 of the CloudMap paper	29: Depth of Coverage on data 5 and data 16 (output summary sample)	* 0 12
Execute	16: Alignment file (BAM)	
	9: FASTQ quality statistics (box plot)	
What it does:	5: W5220.64_chr.fa	
This tool is part of the CloudMap pipeline for analysis of mutant genome sequences. For further details, please see <u>Gregory Minevich, Danny S. Park, Daniel</u> Blankenberg, Richard J. Poole and Oliver Hobert. CloudMap: A Cloud-based Pipeline for Analysis of Mutant Genome Sequences. (Genetics 2012 in Press)	4: ot266 ProofOfPrinciple Small.fastgsanger	
CloudMap workflows, shared histories and reference datasets are available at the <u>CloudMap Galaxy page</u>	3: HA SNPS Unfiltered 112061Variants WS220.vcf	
This tool improves upon, and automates, the method described in Doitsidou et al., PLoS One 2010 for mapping causal mutations using whole genome sequencing data.	2: HA SNPs Filtered 103346Variants WS220.vcf	
Sample CloudMap output for a linked chromosome:	1: CloudMap_TranscriptionFactors_wTF2.2.1x1	

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/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):

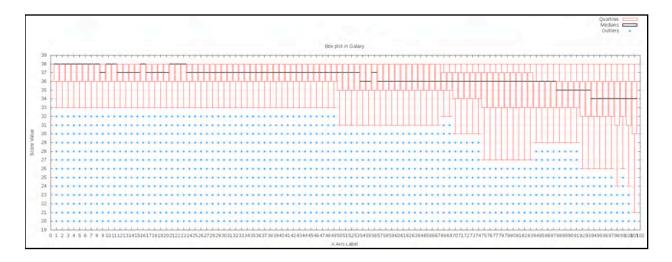
1. A set of homozygous variants (VCF file) in the entire sample that can be further filtered by subtracting variants present in other samples using the *CloudMap Subtract Variants* workflow (GATK). This VCF file is used as input into snpEff to generate the **annotated list** of homozygous variants mentioned 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  $\geq$  30 and coverage  $\geq$  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)

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)

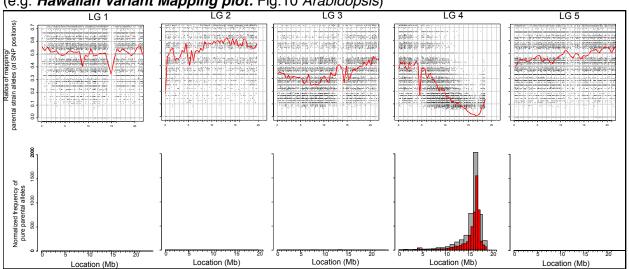
1	A	В	C	D	E	F	G
1	sample_id	total	mean	granular_third_quartile	granular_median	granular_first_quartile	%_bases_above_15
2	rgSM	734789704	7.33	11	7	4	9.7
3	Total	734789704	7.33	N/A	N/A	N/A	

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:

Contractor		Number	Number of effects by type and region						
Contents	Туре			Region					
Change rate by chromosome	Type (alphabetical order)	Count	Percent						
Variants by type	CODON INSERTION	1	0.001%						
Number of variants by impact	DOWNSTREAM	36,909	45.796%						
Number of variants by functional class	FRAME_SHIFT	20	0.025%	Type (alphabetical order)	Count	Percent			
Number of variants by effect	INTERGENIC	22	0.027%	DOWNSTREAM	36,909	45.796%			
Quality histogram	INTRON	-4,139	5.136%	EXON	1,469	1.8239			
Coverage histogram	NON_SYNONYMOUS_CODING	724	0.898%	INTERGENIC	22	0.0279			
Base change table	SPLICE_SITE_ACCEPTOR	3	0.004%	INTRON	4,139	5.1369			
ransition vs transversions (ts/tv)	SPLICE SITE DONOR	1	0.001%	NONE	199	0.2479			
requency of alleles	START_GAINED	13	0.016%	SPLICE SITE ACCEPTOR	3	0.0049			
Codon change table	START_LOST	1	0.001%	SPLICE_SITE_DONOR	1	0.0019			
Amino acid change table	STOP_GAINED	12	0.015%	UPSTREAM	37,618	46.675%			
Chromosome change plots	SYNONYMOUS_CODING	711	0.882%	UTR 3 PRIME	137	0.179			
Details by gene	TRANSCRIPT	199	0.247%	UTR 5 PRIME	98	0.1229			
	UPSTREAM	37,618	46.675%						
	UTR 3 PRIME	137	0.17%						
	UTR 5 PRIME	85	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	1	1222	haw1	4	3	7	0.571429	-21.9682
3	1	3659	haw3	6	7	13	0.461538	-21.9094
4	1	3731	haw4	4	11	15	0.266667	-21.9076
5	1	4101	haw5	9	12	21	0.428571	-21.8987
6	1	4776	haw6	1	8	9	0.111111	-21.8824
7	1	5026	haw7	4	10	14	0.285714	-21.8764
8	1	5868	haw8	0	5	5	0	-21.856

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

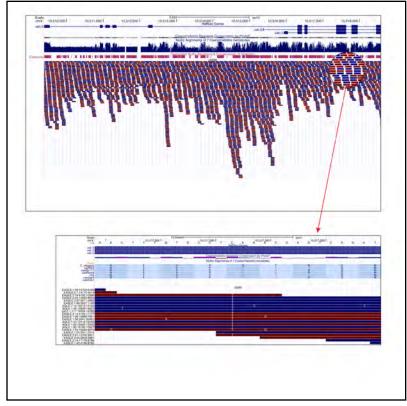
	A			D	· · · ·		6 0	1	Los La	. h.	A. A. Martin	M.		N	0	p.	9
1 100	nromo Positio			Dhange	Change_type		Coverage Gene_ID	Gene_nam			ID Exon Rank Effe			old AA/new	AA Did_codon/New_co	odon Codon_Num(CDS) CI	15_size
3 V.		5472		÷G	1945	299.66	10 ¥43F88.17	¥43F88.17		ne valess I		ISCRIPT: ¥43188					621
F 18		5878		+0	145	2399.2	52 #4889 3	FAILED 3		od/F4889 8		AF SHIFT F4885					585
1 X.		2621		1.	DEL	196.55	25 C04F6.8	CD476.8	IncRNA	1004F6.8	TRA	ISCR/PT: CO4F6.1	8				124
5 IX		3048		C	SMP	37.15	2 (2282.1)	T2282.11	NCRNA	72282.11		SCRIPT: 12282.					148
0.18		3449		T	SNP	157.66	5 \$\$\$D1.1	igem-Z		od(\$5501.1		SYNONYMOUS		G/R	Gag/Agg	TAN	1911
T (K)		7478		*G	INS	210.28	7.80403.12	80403.12	nLRNA	80403.12		ISCRIPT: B0403					300
6 IX		7478		+6	INS .	210 28	7 B0403 11	B040313	ncRNA	60403.13		ISCRIPT: B0403					203
0 X.	731	0138		+6	INS	726.28	26 K03A1.1	KOBALL	pseudoger	ne KOBALI	TRA	SCRIPT: KOBAL	1				410
HIN.		9013		+0	1145	635.6	72 K09F5.11	K09F5-11	ncenia	809F5.11	TRA	SCRIPT: K09F5.	11				137
0. k	773	9013		+C	1145	635.6	22 K09#5.10	K0985.30	INCENA	\$09F5.10	TRA	ISCR:PT: KU9F5.	10				126
10.08	782	3447		+T	1115	300,36	16 HU3G5.8	P0365.8	INCRINA	R0365.8	TRA	ISCRIPT: ROBGS.	ň				141
(EX)	786	6252		-A	DEL	1247.88	50 C5402.16	C5402.16	ncRNA	C54D2.16	TRA	SCRIPT: C54D2.	16				345
H X	802	6796		+T	INS	317.94	10 C34D10.2	C34010.2	prútein_co	odi C34D10.2	.I UTR	3 PRIME: 1423	bases from CDS				2F - CCCH - 2 Idomainy
1 18	829	2734	ε.	T	SNP	1085.02	41 (1389.1	F1389.1	printein co	od F1189.1b	14 NOF	SYNONYMOUS	CODING	5/F	tCt/ITt	1476	4845
× 10	829	2734	0	T	ANP:	1085.02	41 F1389.1	FINB9.1	printera_co	od(#1389.1a	15 Not	SYNONYMOUS	CODING	5/7	tCt/ITI	1448	43090
1.18	829	2734	C	T	SNP	1085.02	41 F1389.1	F1389.1	pratein_co	odi F1389.1c	14 NOT	SYNONYMOUS	CODING	5/E	1Ct/ITt	1426	4830
1.10	840	8774	6	+C	INS	476.87	12 FOBF1 38	F08F1 18	INCENA	F08F1 18	TRA	SCRIPT: FORFI 1	18				283
9.4	863	9239		+C25	1115	775.11	16 #1209.18	61209 18	ncRNA.	F1209-18	TRA	SCRIPT: F12D9	19				81
90 M	863	9239		+05	045	775.11	16 \$1208 15	F1209.15	TRNA	F1209.15	TRA	SCRIPT: F1209.1	15				71
X 1	894	1351		-GATE	DEL	530.28	15 D1073.1	trit-1	orutem_co	odi 01073.1b	15 FRA	ME_SHIFT: 01073	3.10				2523
1 X.	894	1351		-GATC	DEL	\$\$0,28	15 D1075.1	trk-1	protein_cx	odi 01073.1a	12 FRA	AE SHIFT DIOT	3.1a				2112
XLET	934	3610		+A	INS	654.81	50 T2085 B	oga-1	protein co	odi T2085.3a	UTR	I PRIME: 75 ta	ises from CDS				
M X	1045	2433	0	T	<b>UNP</b>	1776:40	42 (3303.1	nit-7	protein co	odi CESD3.1	7 NO	SYNONYMOUS	CODING	3.77	(CI/)TI	111	1102 2F-GATA
<b>X X</b>	1051	7587	6	T	SNP	376-64	16 F14F3.1	vab-3	pratein co	od(F14F3.1b	4.570	GAINED		Q/*	Caa/Tae	152	810 HD - PRD, Paired Domain - FUL
6 8	1051	7587	C	T	SNP	376-64	16 F14F3 1	vab-3	protein_co	od(F14F3 la	9 570	GAINED		0/*	Cas/Tes	338	1368 HD - PRD, Paired Domain - FUL
80x.	1053	7587	2	1	SMP	376,64	16 F14F3.1	Veb-3	protein_co	od F14F3 1c	4 STO	GAINED		Q/*	Cas/Teld	179	891 HO - PRD, Paired Domain - FUL
1.1	1160	0051	6	T	- CMD	572.86	22 TU4F6.1	1905-1.5	protein co	DEL TEAFS 1	5 NO	SYNONYMOUS	CODING	G/R	Gga/Aga	214	975
M	1165	5513	C	1	SAU	427.81	19 644610.4	C14C10.4	protein co	odi C44C10 4	7 NOR	SYNONYMOUS	CODING	U/F.	Ctc/Tti	535	1614
H X.	1249	2661		+G	1115	631.86	18 F45E6.7	F45E6.7	ncRNA	F45E6.7	TRA	SCRIPT P45E6.	7				145
IL X	1406	8660	-	C	SNP	35.86	§ C33G3.13	C33G3.13	ncRNA	C33G3.13	TRA	SCRIPT: C33G3.	13				71
XCE	1430	5870	C	T	SNP	1788.01	46 C11HL2	C11H1 2	protein_co	odi C11H1.2	7 SYN	NYMOUS CODI	INCI	K/K	aaG/au/v.	252	13/0
II IX	1660	8728		-AC	DEL	809.66	24 /59612.8	750C12.8	ncilNA	F59C12.8	TRA	SCRIPT: F59C12	2.8	100			225
HIX.	1725	9200	r i	6	SNP	45.01	14 VIOC78 8	¥40078.3	protein co	odi ¥40C78.3	1.69%	NYMOUS CODI	inici	919	at A/mili	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 (snpEff). This is because read mapping qualities and base qualities are incorporated into which variants are ultimately called. Most alignment viewers have filter settings that can be used to only display reads with mapping quality scores above a certain value. Applying these filters should result in alignments that more closely approximate variant lists.

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

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

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

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

1	A	В	C	D	E	F	G	Н	and the second second	and the second s	K
1	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	rgSM	
2	chrl	42899	2	G	Α	75.03	PASS	AC=2;AF=1.00;AN=2;DP=3	GT:AD:DP:GQ:PL	1/1:0,3:3:9.03	3:107,9,0
3	chrl	62642	1.1	Т	С	48.77	PASS	AC=2;AF=1.00;AN=2;DP=2	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02	2:80,6,0
4	chrl	341299		TG	Т	181.31	PASS	AC=2;AF=1.00;AN=2;DP=6	GT:AD:DP:GQ:PL	1/1:0,6:6:18.0	06:223,18,0
5	chrl	346149		Т	Α	85.77	PASS	AC=2;AF=1.00;AN=2;DP=3	GT:AD:DP:GQ:PL	1/1:0,3:3:9.03	3:118,9,0
6	chrl	361325		С	Α	232.91	PASS	AC=2;AF=1.00;AN=2;DP=7	GT:AD:DP:GQ:PL	1/1:0,7:7:21.0	7:266,21,0
7	chrl	369870		С	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	2:79,6,0
8	chrl	369871		С	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	2:80,6,0
9	chrl	663697	2	G	С	167.29	PASS	AC=2;AF=1.00;AN=2;DP=5	GT:AD:DP:GQ:PL	1/1:0,5:5:15.0	05:200,15,0
10	chrl	670146		G	Α	36.43	PASS	AC=2;AF=1.00;AN=2;DP=2	GT:AD:DP:GQ:PL	1/1:0,2:2:6.0	1:68,6,0
11	chrl	670173		T	С	36.43	PASS	AC=2;AF=1.00;AN=2;DP=2	GT:AD:DP:GQ:PL	1/1:0,2:2:6.0	1:68,6,0
12	chrl	671425		T	Α	48.77	PASS	AC=2;AF=1.00;AN=2;DP=2	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02	2:80,6,0
13	chrl	687402		T	Α	67.01	PASS	AC=2;AF=1.00;AN=2;DP=3	GT:AD:DP:GQ:PL	1/1:0,3:3:9.0	1:99,9,0

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)

1	A	B C	D	E	F G	Н	1	1
1	#CHROM	POS ID	REF	ALT	QUAL FILTER	INFO FORMA	т	rgSM
2	chrl	962 .	G	т	367.18 .	AC=1;AF=0.50;AN=2;BaseQRankSum=0.403;DP=23GT:AD:D	DP:GQ:PL	0/1:10,13:23:99:397,0,325
3	chrl	991 .	GA	G	100.41 .	AC=1;AF=0.50;AN=2;BaseQRankSum=2.130;DP=14GT:AD:D	DP:GQ:PL	0/1:8,6:14:99:139,0,246
4	chrl	1216 .	A	т	68.96 .	AC=1;AF=0.50;AN=2;BaseQRankSum=1.300;DP=7; GT:AD:E	DP:GQ:PL	0/1:4,3:7:98.95:99,0,138
5	chrl	1222 .	Α	С	109.76 .	AC=1;AF=0.50;AN=2;BaseQRankSum=1.754;DP=7; GT:AD:D	DP:GQ:PL	0/1:3,4:7:57.20:140,0,57
6	chrl	1290 .	Т	A	126.47 .	AC=1;AF=0.50;AN=2;BaseQRankSum=0.933;DP=14GT:AD:D	DP:GQ:PL	0/1:9,5:14:99:156,0,306
7	chrl	1412 .	Т	С	235.12 .	AC=1;AF=0.50;AN=2;BaseQRankSum=-1.203;DP=1 GT:AD:D	DP:GQ:PL	0/1:8,9:17:99:265,0,266
8	chrl	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	chrl	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)

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

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

## CloudMap

Cloud-based Pipeline for Analysis of Mutant Genome Sequences

#### CloudMap UnMapped Mutant Workflow

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

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 <a href="http://usegalaxy.org/cloudmap">http://usegalaxy.org/cloudmap</a>.

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.

- 奇」| + てhttps://main.g2.bs.psu.edu/roo Galaxy Analyze Data Tools • History 큀 search tools Corel. 0. bytes Get Data Galaxy 101 Your history is empty. Click 'Get Data' on the left pane to start Send Data Start small ENCODE Tools Lift-Over The very first tutorial you need Text Manipulation **Convert Formats FASTA** manipulation Filter and Sort Join, Subtract and Group **Extract Features** Live Quickies **Fetch Sequences** Fetch Alignments Get Genomic Scores **Operate on Genomic Intervals** Statistics Graph/Display Data **Regional Variation** Multiple regression Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, **Multivariate Analysis** can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Evolution ments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Computer Science depart Sciences. The Institute for CyberScience at Penn State, and Emory University. Motif Tools
- 1) Navigate to <u>http://usegalaxy.org</u> (URL will resolve to something like <u>https://main.g2.bx.psu.edu</u>)

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:

Analyze Data	Workflow	Shared Data Visualization - Cloud - Help - User -
	R	Data Libraries Published Histories Published Workflows Published Visualizations Published Pages
	Ne	wyou can have a personal Galaxy within the infinite Universe

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

= Galaxy	Analyze Data Workflow Shared Data - Visualization-
Data Libraries	
Close Advanced Search       data library name:       data library description:	
Data library name 4	Data library description
1000 genomes	
100209_HsMtDNA	
anton_test	
bushman	
CloudMap	Contains reference and configuration files for the Cloudmap pipeline
Codon Usage Frequencies	
Dannon's Test Data Library	Testing library for Dannon
FRIK920	
GATK	

-

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

🛄 Name	Message
💷 🌱 📔 CloudMap Candidate Gene Lists –	For CloudMap Check snpEff Candidates tool
CloudMap_C:elegansGenesWithHumanOrthologs.txt =	
CloudMap_ChromatinFactors.txt -	
CloudMap_TranscriptionFactors_wTF2.2.txt =	
CloudMap EMS Variant Density Mapping -	Use this dataset to try out the CloudMap EMS Variant Density Mapping tool
🗉 🍸 🔲 CloudMap ot266 proof of principle dataset –	Use these files to run the CloudMap ot266 proof of principle example
Hawaiian SNP reference files filtered (WS220.64) =	Filtered set of Hawailan SNP variants (used by CloudMap SNP Mapping with WGS tool)
💷 🍸 📴 Hawaiian SNP reference files unfiltered (WS220.64) =	Unfiltered set of Hawaiian SNP variants (used by CloudMap SNP Mapping with WGS tool)
HA_SNPS_Unfiltered_112061Variants_WS220.64_chr.vcf -	
ot266_ProofOfPrinciple_Small.fastqsanger =	None
₩S220.64_chr.fa -	
💷 🕨 🔲 CloudMap user guides =	Detailed guides for using the CloudMap pipeline
Image: Image: Provide the second s	Use these BED files for the CloudMap ot266 proof of principle for uncovered region subtraction
t260 and ot263 VCFs for variant subtraction -	Use these VCF files for the CloudMap ot266 proof of principle variant subtraction

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:



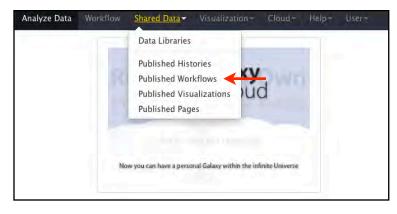
8) You will now see that the data files have been added to an unnamed history:

History	\$
Unnamed history	2.3 Gb
<u>3: WS220.64 chr.fa</u>	• 0 %
2: ot266 ProofOfPrinciple Sm nger	● Ø X nall.fastgsa
1: CloudMap TranscriptionFa 2.txt	● Ø ☆ ctors wTF2.

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

History	0
ان الله الله الله الله الله الله الله ال	2.3 Gb
3: WS220.64 chr.fa	• 0 %
2: ot266 ProofOfPrinciple Sr nger	● Ø X mall.fastqsa
1: CloudMap TranscriptionFa 2.txt	● Ø ☆ actors wTF2.

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



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

Published Workflows	
Cloudmap	IG.
Advanced Search	
Name	
CloudMap Hawaiian Variant Mapping	with WGS and Variant Calling workflow
CloudMap Unmapped Mutant workfl	ow (w/ subtraction of other strains)
CloudMap EMS Variant Density Mapp subtract)	ing workflow (takes VCF of heterozygous and homozygous variants to
CloudMap Unmapped Mutant workfl	

12) You will now have the option to Import workflow

- Galaxy	Analyze Data	Workflow	Shared Data <del>-</del>	Visualization -	Cloud +	Help -	User-	
Published Workflows   gal40   CloudMap Unmapped Mutant workflow	N							Import workflow
Galaxy Workflow ' CloudMap Unmapped Mutant workflo	w'							
Step				Annotation				
Step 1: Input dataset								
Fasta reference genome select at runtime								
Step 2: Input dataset								
FASTQ reads select at runtime								
Step 3: Input dataset								
Candidate gene list (e.g. transcription factors) select at runtime								
Step 4: FASTQ Summary Statistics								
FASTQ File Output dataset 'output' from step 2								
Step 5: Map with BWA for Illumina	1							

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



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



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

our workflows		
Name		
imported: CloudMap Unmapped Muta	nt workflow = ] Edit	
Cloudmap Uncovered Region Subtrac	Run	
CloudMap SNP Mapping with WGS Da	Share or Publish	low
CloudMap Unmapped Mutant workflo	Download or Export Clone	
CloudMap Subtract Variants workflov	Rename	
Norkflows shared with	View 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.

Analyze Data Workflow Shared Data - Visualization - Cloud - Help - User -	Using 71%
Running workflow "imported: CloudMap Unmapped Mutant workflow" Expand All Collapse	History
Step 1: Input dataset	Unnamed history 2.3 Gb
Fasta reference genome 3: WS220.64_chr.fa	<u>3: WS220.64 chr.fa</u> ● Ø 🛠
type to filter	2: ot266 ProofOfPrinciple Small.fastqsa
Step 2: Input dataset	1: O / X
FASTQ reads     Image: Control of the second s	1: ● Ø X CloudMap TranscriptionFactors wTF2. 2.txt
Step 3: Input dataset	
Candidate gene list (e.g. transcription factors)	
Step 4: FASTQ Summary Statistics	
Step 5: Map with BWA for Illumina	
Step 6: Boxplot	
Step 7: Filter SAM	
Step 8: SAM-to-BAM	

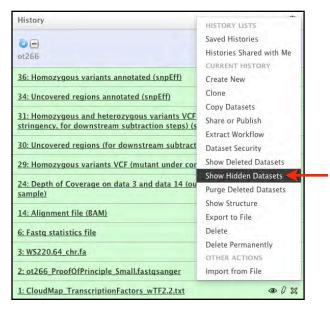
17) All of the automated functions have the appropriate default parameters configured, although experienced users may want to modify these prior to running (see the Analyzing Your Own Data Using CloudMap Workflows section of this user guide). Once you are ready to run the workflow, press Run Workflow 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:

Analyze Data Workflow Shared Data Visualization Cloud+ Help+ User+		Using 71%
Successfully ran workflow "imported: CloudMap Unmapped Mutant workflow". The following datasets have been added to the queue:	History	
3: WS220.64_chr.fa	22: Unified Genotyper on data 3 and data 14 (log)	
2: ot266_ProofOfPrinciple_Small.fastqsanger	21: Unified Genotyper on data 3 and data 14 (metrics)	000
1: CloudMap_TranscriptionFactors_wTF2.2.txt		~ / /
4: FASTQ Summary Statistics on data 2	20: Unified Genotyper on data 3 and data 14 (VCF)	@ 0 2
5: Map with BWA for Illumina on data 2 and data 3: mapped reads	19: Alignment file (BAM) (Genome Coverage BedGraph)	0003
6: Fastq statistics file		
7: Filter SAM on data 5	18: Unified Genotyper on data 3 and data 14 (log)	@ 0 2
8: SAM-to-BAM on data 3 and data 7: converted BAM	17: Unified Genotyper on data 3 and data 14 (metrics)	0003
9: Add or Replace Groups on data 8: bam with read groups replaced		
10: Realigner Target Creator on data 3 and data 9 (GATK intervals)	16: Unified Genotyper on data 3 and data 14 (VCF)	0003
11: Realigner Target Creator on data 3 and data 9 (log)	15: MarkDups Alignment file (BAM).html	.00
12: Indel Realigner on data 3, data 10, and data 9 (BAM)	14: Alignment file (BAM)	-
13: Indel Realigner on data 3, data 10, and data 9 (log)	G 14. Augument me (DAM)	000
14: Alignment file (BAM)	13: Indel Realigner on data 3, data 10, and data 9 (log)	00
15: MarkDups_Alignment file (BAM).html	12: Indel Realigner on data 3, data 10, and data 9 (BAM)	000
16: Unified Genotyper on data 3 and data 14 (VCF)	The main realigner on data 3, data 10, and data 3 (bring)	- 1 A
17: Unified Genotyper on data 3 and data 14 (metrics)	11: Realigner Target Creator on data 3 and data 9 (log)	000
18: Unified Genotyper on data 3 and data 14 (log)	10: Realigner Target Creator on data 3 and data 9 (GATK	000
19: Alignment file (BAM) (Genome Coverage BedGraph)	intervals)	
20: Unified Genotyper on data 3 and data 14 (VCF)	3 9: Add or Replace Groups on data 8: bam with read groups	
21: Unified Genotyper on data 3 and data 14 (metrics)	replaced	
22: Unified Genotyper on data 3 and data 14 (log)		
23: Depth of Coverage on data 3 and data 14 (per locus coverage)	8: SAM-to-BAM on data 3 and data 7: converted BAM	
24: Depth of Coverage on data 3 and data 14 (output summary sample)	T: Filter SAM on data 5	00
25: Depth of Coverage on data 3 and data 14 (output statistics sample)	🖀 6: Fastg statistics file	
26: Depth of Coverage on data 3 and data 14 (output cumulative coverage counts sample)		
27: Depth of Coverage on data 3 and data 14 (output cumulative coverage proportions sample)	5: Map with BWA for Illumina on data 2 and data 3: mapped	
28: Depth of Coverage on data 3 and data 14 (log)	reads	
29: Homozygous variants VCF (mutant under consideration)	4: FASTQ Summary Statistics on data 2	
30: Uncovered regions (for downstream subtractions)	3: WS220.64 chr.fa	
31: Homozygous and heterozygous variants VCF (higher stringency, for downstream subtraction steps) (snpEff)		
32: Homozygous variants not annotated (snpEff )	2: ot266 ProofOfPrinciple Small.fastgsanger	
33: SnpEff on data 29 34: Uncovered regions annotated (snpEff)	1: CloudMap TranscriptionFactors wTF2.2.txt	

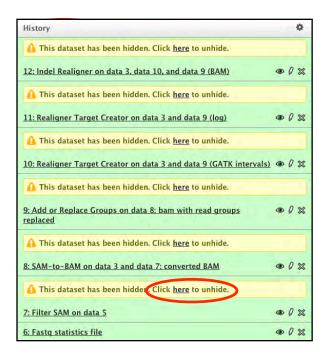
18) Once the workflow has finished running, you can view the resulting output:

	Analyze Data Workflow Mared Data - Visualizations Chinds Help - Users		Using 86%	
~	Hello world! This is galaxy test.	History		0
ľ	This Galaxy server has the latest and greatest features. It is designed for testing these features and may explode and/or implode without warning. It tracks the palaxy-central repo	<b>U</b> = ot266		
	Impose would warning, it overs the <u>sensity ventual</u> (cpu-	36: Homozygous variants annotated (snpEff)	•	2
		34: Uncovered regions annotated (snpEff)		2
0	Galaxy Test has <u>usage quotas</u> .	31: Homozygous and heterozygous variants VCF (higher stringency, for downstream subtraction steps) (snpEff)	• 0	3
	The number of concurrent jobs and the amount of disk you can use on this server is limited by guotas.	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)	•0	
	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	6: Fastg statistics file	• 0	
	2	3: WS220.64_chr.fa		9
		2: ot266 ProofOfPrinciple Small.fastqsanger	• 0	1
	the test is for breaking	1: CloudMap_TranscriptionFactors_wTF2.2.txt	- (D	
alaxy i	axy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Project is supported in part by NSE. NHGRI. The Huck Institutes of the Life Sciences. The Institute for CyberScience at Penn State, and Iniversity.			
te, inc	a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your h data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public fuding this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup in local Galaxy instance or run Galaxy on the cloud.			

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:



20) You may unhide any files that are hidden:



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

- Galaxy					Anal	vze Data Wurl	illin Sanat	Nation N	vuolio utiviti -	Eloud-	Belo- Dance		Using 8	6%
Tools	0	# Snpäll vers	sion 2.1a (b	uild 2012-04	-20), by Pa	blo Cingolani		-				- 7	History	0
		# Command	line: SnoEH	eff -c /galax	y/home/g2	test/galaxy_test	/tool-data/snpel	H/snpEff.cd	onlig -ivef -a t	xt-upDow	nStreamLen 1000			
search tools		# Chroma	Position	Reference	Change	Change_type	Homozygous	Quality	Coverage 1	Warnings	Gene_ID			
Get Data		1	42899	G	A	SNP	Hom	75.03	3		Y48G1C.12	¥48G	0265	67 Glr
BEDtools		T	42.899	G	A	SNP	Hoim	75.03	3		¥48G1C.4	pgs-	36: Homozygous variants annotated (snpEff)	1 40
SNP Eff		1	42899	C	A	SNP	Hom	75,03	3		Y74C9A.1	Y74C	80,595 lines, 3 comments	200
Send Data		1	62642	т	C	SNP	Hom	48.77	2		¥48G1C.4	pgs-	on at tabular, database ce10	
ENCODE Tools		1	62642	T	C	SNP	Hom	48.77	2		¥48G1C.5	Y48C	100	
Lift-Over		3	62642	T	C	SNP	Hom	48,77	2		Y48GLC.Z	esk-1		_
Text Manipulation		4	62642	т	C	SNP	Hom	48.77	2		¥48G1C.Z	csk-1	3 2 3 4 3 6 7 8	
		1	62642	т	C	SNP	Hom	48.77	2		Y48G1C.2	CSR-1	# SnpEff version 2.1a (build 2012-04-20), by Pablo Cingolani	
Filter and Sort			72355	A	G	SNP	Hom	48.77	2		¥48G1C.2	tsk-1	# Command Line: SmpEff off -t /galaxy/home/g2test/galaxy_test/tool-data/s	inpett.
Join, Subtract and Group		1	72355	A	6	SNP	Hom	48.77	2		¥48G1C.10	¥48G	st_pool/pool1/files/000/304/dataset_304498.dat	
Convert Formata		1	72355	A	G	SNP	Hom	48.77	2		¥48G1C-2	csk-1	# Chrono Positizon Reference Change Change_type H	tomozyr
Extract Features		1	72355	A	G	SNP	Hom	48.77	2		¥48G1C.2	csk-1	interval ID	
Fetch Sequences		1	72355	A	6	SNP	Hom	48.77	2		¥48G1C.11	Y48G		
Fetch Alignments		4	72355	A	G	SNP	Hom	48.77	2		¥48G1C.5	¥480		
Get Genomic Scores		1	200948	A	- E	SNP	Hoim	317.33	9		Y48G1BL7	¥48G		
Operate on Genomic Intervals		1	200948	A	T	SNP	Hom	317.33	9		K10F9,1		34: Uncovered regions annotated (sngEff)	0 32
Statistics		1	200948	A	τ	SNP	Hom	317.33	9		¥48G18L2	atm-	31: Homozygous and heterozygous variants VCF (higher	0 22
Wavelet Analysis			200949	C.	Ŧ	SNP	Hom	309.05	9		Y48G18L7	¥48C	stringency, for downstream subtraction steps) (snpEff)	1 44
Graph/Display Data		1	200949	C	T	SNP	Hom	309,05	9		KIDE9.1	K10E		
Regional Variation		4	200949	C	τ.	SNP	Hom	309.05	9		¥46G18L2	atm-	30: Uncovered regions (for downstream subtractions)	0 22
		4	341300	ð	-6	DEL.	Hom	181.31	6		Y48G1A.6	mbtr	29: Homozygous variants VCE (mutant under consideration)	0 st
Multiple regression		1	341300	3 A.	-6	DEL	Hom	181,31	6		Y48G1A.6	mbtr		0 H
Multivariate Analysis		1	\$41300	3	-6	DEL	Hom	181.31	6		Y48GLA.3	¥48G		22 0
Evolution			341300		-G	DEI.	Hom	181.31	6		¥48G1A-1	¥48G		
Motif Tools		4	341300	25	-G	DEL	Hom	181.31	6		Y48G1A.2	¥48G		é
Multiple Alignments			341300	a	-6	DEL	Hom	181.31	6		¥48G1A.2		14: Alignment file (BAM)	0#
Metagenomic analyses		1	346149	T	A	SNP	Hom	85.77	3		¥48G1A.3	Y48G	6: Fasto statistics file	0 22
FASTA manipulation		5 I I I	346149	τ.	A	SNP	Hom	85.77			¥48G1A.1	Y48C		
NCBI BLAST+		1	346149	T	A	SINP	Hoin	85.77	3		¥48G1A.6	mbtr	3: WS220.64 chr.fa	自該
NGS: OC and manipulation		2 C	346149	T		SNP	Hom	85.77	3		Y48G1A.6	mbtr	2 - 1277 Bar (DIB) - 12 Facility - 12 - 12 - 12	0 22
NGS: Picard (beta)		5 C	546149	T	A	SNP	Hom	85.77	3		Y48G1A.Z		and the provide state of the st	0 24
NGS: Mapping		1.1	346149	1	N.	5NP	Hom	85.77			Y48GLA.2	Y48G	1: CloudMap TranscriptionFactors wTF2.2.txt	0 32
NGS: Indel Analysis			361325	C		SNP	Hom	232.91	- Z		R119.7	rnp-L		
NGS: RNA Analysis			361325	C.	A	SNP	Hom	232.91			Y48G1A.1	Y48C		
		<b>T</b>	361325	S	2	SNP	Hom	232,91	1		R119.1	R119		
NGS: SAM Tools		1	\$61325	£	A	SNP	Hom	232.91	- Z -		R119.2	R119		



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)	History	¢
SnpEff File: [32: Homozygous variand (snpEff ) = #]	6 () 01266	12.7 Gb
tabular autput file from snpEff Candidate List:	36: Homozvoous variants annotated (snpE(f) Run this joh again minents atlabase: ce10	
Execute	1 2 3 4 5 6 7 # SnpEff version 2.1a (build 2012-04-20), by Pable Cingolani	8 9
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: <u>http://snpeff.sourceforge.net</u> .	# Command Line: SngEff eff -c /galaxy/home/gites//galaxy_test//tool-d st_pool/pooli/files/000/304/daiase_304408.dat # Chromo Position Reference Change Change_type interval_ID I 42809 G A 90P Hon 75,03	ata/snpeff, Homozyg 3
Input: The candidate list should be in a tabular format with two columns: Gene ID and Gene Description (e.g. C5587.12 and transcription_factor). The file should contain no headers.	34: Uncovered regions annotated (snpEff) 32: Homozygous variants not annotated (snpEff.)	• 0 x • 0 x
Useful candidate lists (e.g. transcription factors, genes expressed in neurons, transgene silencers, chromatin factors) are available on the CloudMap Galaxy page.	31: Homozygous and heterozygous variants VCF (higher stringency, for downstream subtraction steps) (snpEff)	• 0 %
https://test.g2.bx.psu.edu/u/gal40/p/cloudmap	30: Uncovered regions (for downstream subtractions)	
Citation:	29: Homozygous variants VCF (mutant under consideration)	
This tool is part of the CloudMap pipeline for analysis of mutant genome sequences. For further details, please see <u>Gregory Minewich, Danny</u> Park, Richard J. Poole, Daniel Blankenberg, Anton Nekrutenko, and Oliver Hobert. CloudMap. A Cloud-based Pipeline for Analysis of Mutant	24: Depth of Coverage on data 3 and data 14 (output summary sample)	• 0 2
Genome Sequences. (2012 In Preparation)	14: Alignment file (BAM)	
Correspondence to am2123@columbia.edu (G.M.) or ar38@columbia.edu (O.H.)	6: Fastq statistics file	00 D 22
	3: W\$220,64_chr.fa	
	2: 01266. ProofOfPrinciple_Small.fastgsanger	
	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):

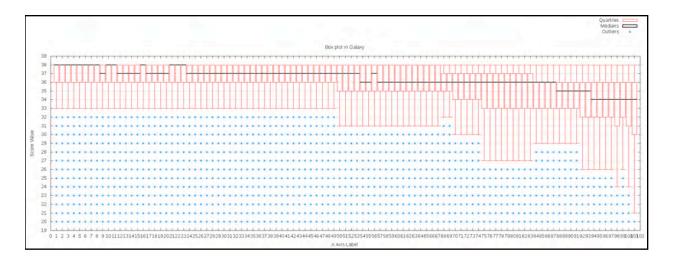
1. A set of homozygous variants (VCF file) in the entire sample that can be further filtered by subtracting variants present in other samples using the *CloudMap Subtract Variants* workflow (GATK). This VCF file is used as input into snpEff to generate the **annotated list** of homozygous variants mentioned 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  $\geq$  30 and coverage  $\geq$  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)

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)

12	A	В	С	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

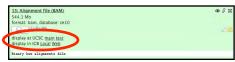
Туре		Region					
Type (alphabetical order)	Count	Percent					
CODON_INSERTION	1	0.001%					
DOWNSTREAM	36,909	45.796%					
FRAME_SHIFT	20	0.025%	Type (alphabetical order)	Count	Percent		
INTERGENIC	22	0.027%	DOWNSTREAM	36,909	45.796%		
INTRON	4,139	5.136%	EXON	1,469	1.823%		
NON_SYNONYMOUS_CODING	724	0.898%	INTERGENIC	22	0.027%		
SPLICE_SITE_ACCEPTOR	3	0.004%	INTRON	4,139	5.136%		
SPLICE_SITE_DONOR	1	0.001%	NONE	199	0.247%		
START_GAINED	13	0.016%	SPLICE_SITE_ACCEPTOR	3	0.004%		
START_LOST	1	0.001%	SPLICE_SITE_DONOR	1	0.001%		
STOP_GAINED	12	0.015%	UPSTREAM	37,618	46.675%		
SYNONYMOUS_CODING	711	0.882%	UTR_3_PRIME	137	0.17%		
TRANSCRIPT	199	0.247%	UTR_5_PRIME	98	0.122%		
UPSTREAM	37,618	46.675%					
UTR_3_PRIME	137	0.17%					
UTR_5_PRIME	85	0.105%					

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

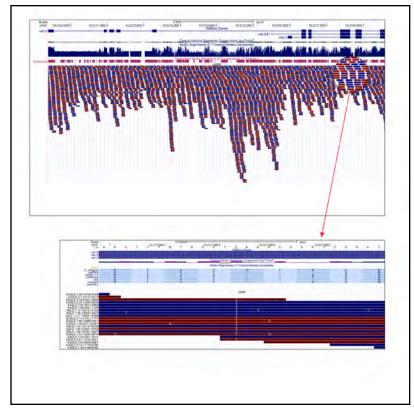
	A	D			6 11	1		all he all	L M	1	0		9 .
I I Chris	imo Position Referen	ce Diange	Change type	Quality	Coverage Gene_ID	Gene nam	e Bio_type	Trancript ID E	kon Rank Effect	old AA/r	new AA Did codon/New codon Co	don Num(CDS) CI	DS size
5 V	19485472 *	÷G	1145	299.66	10 ¥43F88.17	¥43588 17	plaudoge	ne ¥41F8617	TRANSCRIPT: ¥43188.17				621
× 18	2165878 *	+6	1145	2399.2	52 (4889 3	FAUB9 3	pràtéin o	od/ F4889 8	5 FRAME SHIFT F4889 3				585
1 18	3412021 *	-2	DEL	196.55	25 C04F6.8	C0476.8	IncRNA.	1004F6.8	TRANSCRIPT: C04F6.8				124
S 1X	3903048 T	c	SNP	37.15	2 (2282.1)	T2292.11	INCRINA.	72282.11	TRANSCRIPT: 72282.11				148
0.18	6333449 C	T	SNP	157.66	5 \$\$\$D1.1	igcm-Z	protein c	od/\$5501.1	5 NON SYNONYMOUS CODING	G/R	Gast/Assr	1.8.8	1911
TX	7037478 *	a.G	INS	210.28	7.00403.12	80403.12	ncRNA	00403.12	TRANSCRIPT: 80403.12				200
n 1X	7037478 *	*G	145	210 28	7 80403 11	8040313	nckNA	60403.13	TRANSCRIPT: 80403 15				203
0 8	7310138 *	+6	INS	726.28	26 K03A1.1	KOBALL	pacudoge	ne klobAli	TRANSCRIPT: K03A1.1				410
111 X	7719013 *	+C	1145	635.6	22 K09F5.11	K09F5-11	nestia	809F5.11	TRANSCRIPT: K09F5.11				137
LL R	7719013 *	+0	1145	635.6	22 K0945.10	80985-10	neRNA	¥09F5.10	TRANSCRIPT: 80945.10				126
AT 18	7823447 *	+T	1145	300,35	16 H03G5.8	P0365.8	ncRNA	R0365.8	TRANSCRIPT: R03GS.#				141
LL X	7866252 *	-A	DEL	1247.88	50 CS402.16	CS407.16	ncRNA	C54D2.16	TRANSCRIPT: C5402.16				345
til X	8026796 *	+T	INS	317.94	10 (34010.2	C34010.2	prutoin_c	odi C34D10.2.1	UTR 3 PRIME: 1423 bases from CDS				2F - CCCH - 2 domainv
NI X	#292734 C	T	SNP	1085.02	41 F1389.1	F1387.1	protein c	odi F1189.1h	14 NON SYNONYMOUS CODING	5/F	tCt/TT	1426	4845
X IX	8292734 C	T	KNP .	1085.02	41 F1389.1	FINB9.1	printers, o	od(#1389.1a	15 NON SYNONYMOUS CODING	5/1	tCt/ITi	1448	49.99
17.3	8292734 C	T	SNP	1085.02	41 F1389.1	F1389.1	pratein_c	odi F1389.1c	14 NON SYNONYMOUS CODING	S/E	10(/17)	1426	4830
10.0	8408774 *	+0	145	476.87	12.F08F1.18	F08F1 18	INCRINA	F08F1 18	TRANSCRIPT: FOBF1 18				283
19.4	8639239 *	+CI3	1945	775.11	16 \$1209.18	F1209.18	ncRNA	F1209-18	TRANSCRIPT: F12D9 18				81
20.18	8639239 *	+05	045	775.11	16 #1208 15	F1209.15	TRNA	F1209.15	TRANSCRIPT: #1209.15				75
28.8	8941351 *	-GATE	DEL	530.28	15 D1073.1	trit-1	onstein o	odi D1073.1b	15 FRAME_SHIFT: 01073.1b				2523
22 X	8941351 *	-GATC	DEL	\$30.28	15 D1075.1	ok-1	protein_c	od/ D1073.1a	12 FRAME_SHIFT: D1073.1a				2112
21 X	9343610 *	+A	INS	654.81	30 T2085 3	oga-1	protein p	odi T2085.3a	UTR 1 PRIME: 75 tasses from CD5				
24 X	10482433 C	T	SNP	1776.49	42 (3303.1	nit-7	protein_c	odi CESD3.1	7 NON_SYNONYMOUS_CODING	3/1	1CI/ITI	881	1802 2F-GATA
X III	10517587 C	T	SNP	376.64	16 F14F3.1	vab-3	pratein_c	od(F14F3.1b	4 STOP_GAINED	Q/*	Caa/Tae	152	810 HD - PRD, Paired Domain - FUI
26 X	10517587 C	1	SNIP	376-64	16 F14F3 1	vab-3	protein, c	od(F14F3.la	9 STOP_GAINED	0/*	Caa/Tee	338	1368 HD - PRD, Paired Domain - FUI
A138	10517587 C	1	SMP	376,64	16 F14F3.1	Veb-3	protein c	od F14F3.1c	4 STOP GAINED	Q/*	Eas/Tel	179	891 HO - PRD, Pained Domain - FUI
A 16	11660051 C	1	- SMP	572.86	22 10416.1	1905-1.5	protein c	DEL TOAFE 1	5 NON SYNONYMOUS CODING	G/R	Gga/Age	214	975
ALL INC.	11695513 C	1	SMP	427.81	19 644610.4	C14C10.4	protein o	odi C44C10 4	7 NON SYNONYMOUS CODING	U/F.	Cte/Tti	535	1514
AL IN	12492661 *	+G	1145	631.86	18 F45E6.7	F45E6.7	ncRNA	F45E6.7	TRANSCRIPT: 14516.7				145
IL X	14060338 T	c	SNP	35.86	\$ C33G3.13	(3363.13	ncRNA	C33G3.13	TRANSCRIPT: 03363.13				71
XI.IX	14305870 C	T	SNP	1758.01	46 C11HL2	C11H1.2	protein_p	odi C11H1.2	7 SYNONYMOUS CODING	K/K	aaG/aaA.	252	1383
III X	15508728 *	-AC	DEL	809.66	24 759612.8	F50C12.8	<b>ncRNA</b>	F59C12.8	TRANSCRIPT: F59C12.8	12			225
HOX .	17259200 T	6	SNP	45.01	14 Y40C78.3	¥40C75.3	pratein ci	odi ¥40C78.3	1 SYNONYMOUS CODING	W/W	atA/atū	104	1251

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

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

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

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

1	A	В	C	D	E	F	G	Н	and the second second	A CONTRACTOR OF THE	K
1	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	rgSM	
2	chrl	42899	2	G	Α	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	chrl	62642	1.1	Т	С	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	chrl	341299		TG	Т	181.31	PASS	AC=2;AF=1.00;AN=2;DP=6;	GT:AD:DP:GQ:PL	1/1:0,6:6:18.0	6:223,18,0
5	chrl	346149		Т	Α	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	chrl	361325	2.	С	Α	232.91	PASS	AC=2;AF=1.00;AN=2;DP=7;	GT:AD:DP:GQ:PL	1/1:0,7:7:21.0	7:266,21,0
7	chrl	369870		С	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	chrl	369871		С	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	chrl	663697	2.	G	С	167.29	PASS	AC=2;AF=1.00;AN=2;DP=5;	GT:AD:DP:GQ:PL	1/1:0,5:5:15.0	5:200,15,0
10	chrl	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	chrl	670173		Т	С	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	chrl	671425		T	Α	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	chrl	687402		Т	Α	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)

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

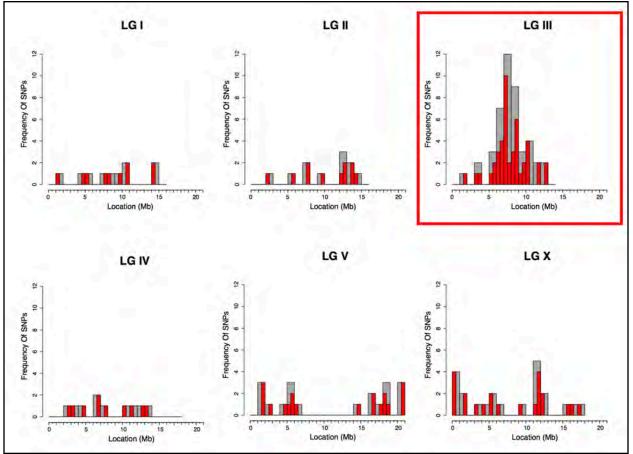
25.3) Set of uncovered regions (BED file) (BEDtools). (For more information on file format, see: http://genome.ucsc.edu/FAQ/FAQformat)

	A	В	С	D
1	chrl	2645	2664	0
2	chrl	3467	3482	0
3	chrl	3925	4014	0
4	chrl	8673	8703	0
5	chrl	8835	8995	0
6	chrl	9774	9787	0
7	chrl	11219	11317	0
8	chrl	11450	11469	0
9	chrl	15107	15117	0
10	chrl	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 quide 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.

e /) //	lu/root		Galaxy.		c lig-tes		0
Go ID III Nema Calegoria							
Galaxy	~ 1_	Analyze Data WorkTow	Shared Data + Visioniza	non- Cloud+ Help+ U	iet-		Using 0%
Tools	0	111 Mar 1				History	0
search tools						-	
Get Data			Galaxy 1	01		0 bytes	and the second second
Send Data			Galaxy			1 Your history is e	
ENCODE Tools			Start sma			Data' on the left	pane to start
Lift-Over			The very first tutorial	you need			
Text Manipulation			The tery mat totolla	Journeed		1	
Convert Formats							
FASTA manipulation							
Filter and Sort							
Join, Subtract and Group							
Extract Features			Live Ouickies				
Fetch Sequences			Live Quickles				
Fetch Alignments			1.00				
Get Genomic Scores	Basic fastQ	Advanced fastQ	454 Mapping:	Uploading Data	Managing		
Operate on Genomic Intervals	manipulation	manipulation:	Single End	using FTP	account histories		
Statistics			00000000		Canal ( gan and a		
Graph/Display Data							
Regional Variation							
Multiple regression	Galaxy is an open, w	eb-based platform for data in	tensive biomedical research.	Whether on this free public ser	ver or your own instance, you		
Multivariate Analysis				t of BX at Penn State, and the B			
Evolution		partments at Emory Universit e for CyberScience at Penn St		irted in part by NSE, NHGRI, Th	e Huck Institutes of the Life		
Motif Tools	stances, the institut	a no speciality at renit a	are, and carry ourselarly.	111111 - La 20			

1) Navigate to http://usegalaxy.org

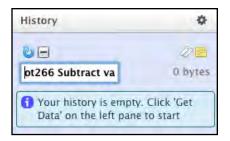
2) You should already have a Galaxy account at this point because you have run earlier workflows:



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:

History	\$				
ot266 Subtract variants example	0 bytes				
Your history is empty. Click 'Get Data' on the left pane to start					

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

Analyze Data	Workflow	Shared Data Visualization Cloud Help User
	R	Data Libraries Published Histories Published Workflows Published Visualizations Published Pages
	No	wy you can have a personal Galaxy within the infinite Universe

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

= Galaxy	Analyze Data Workflow Shared Data - Visualization-				
Data Libraries					
Close Advanced Search       data library name:     Cloudmap       data library description:     Q					
Data library name 4	Data library description				
1000 genomes					
100209 HsMtDNA					
anton_test					
bushman					
CloudMap	Contains reference and configuration files for the Cloudmap pipeline				
Codon Usage Frequencies					
Dannon's Test Data Library	Testing library for Dannon				
FRIK920					
GATK					

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

0	Name
0	n 🔁 CloudMap Candidate Gene Lists -
	CloudMap_C.elegansGenesWithHumanOrthologs.txt =
	CloudMap_ChromatinFactors.txt =
>	CloudMap_TranscriptionFactors_wTF2.2.txt -
a	CloudMap EMS Variant Density Mapping =
	🖣 🔚 CloudMap ot266 proof of principle dataset -
5	🗇 🖗 💼 Hawaiian SNP reference files filtered (WS220.64) =
1	🕨 💼 Hawaiian SNP reference files unfiltered (WS220.64) –
	ot266_ProofOfPrinciple_Small.fastqsanger -
>	₩S220.64_chr.fa =
0	CloudMap user guides
	of260 and of263 BEDs for uncovered subtraction -
0	🔊 📴 ot260 and ot263 VCFs for variant subtraction –
>	ot260_Heterozygous_and_Homozygous_variants_(For_subtracting_from_other_strains_(higher_stringency)).vcf
$\rightarrow$	ot263 Heterozygous and Homozygous variants (For subtracting from other strains (higher stringency)).vcf

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

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

Data Library "CloudMap"
4 datasets imported into 1 history: ot266 Subtract variants example

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

Analyze Data Shared Data -

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

History	\$			
ot266 Subtract variants example	2 🖻 97.6 Mb			
<u>4: WS220.64 chr.fa</u>	• 0 ×			
3:				
2: ot260 Heterozygous and us variants (For subtracti ther strains (higher string	ng from o			
<u>1:</u> CloudMap TranscriptionFa 2.2.txt	⊕ Ø X actors wTF			

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:

Analyze Data	Workflow	Shared Data - Visualization	- Claud-	Heip-	Usei –
		Data Libraries			
		Published Histories			
		Published Workflows			
Da	tasets	Published Visualizations Published Pages		Size o	n Disk

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

<b>Published Histories</b>	
CloudMap x	9
Advanced Search	
Name	
CloudMap_ot266_Proof_of_Princip	le (with hidden data)
CloudMap_ot266_Proof_of_Princip	le (with unhidden data)

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

Published Histories   gm2123   CloudMap_ot266_Proof_of_Principle (with his	O Import		
Galaxy History ' CloudMap_ot266_Proof_of_Principle (with hidden data)'			
Dataset	_	Annotation	
1: CloudMap_TranscriptionFactors_wTF2.2.txt	æ		
2: HA_SNPs_Filtered_103346Variants_WS220.vcf	æ		
3: HA SNPS Unfiltered 112061Variants W5220.vcf	æ		
4: ot266_ProofOfPrinciple_Small.fastqsanger	٠		
5: W\$220.64 chr.fa	æ		
9: FASTQ quality statistics (box plot)	æ		
16: Alignment file (BAM)	æ		
29: Depth of Coverage on data 5 and data 16 (output summary sample)	٠		
38: Uncovered regions (BED file for downstream subtractions and snpEff annotation)	- 20-		
39: CloudMap: Hawalian Variant Mapping with WGS data on data 34			
40: CloudMap: Hawaiian Variant Mapping with WGS data on data 34	•		
41: Homozyajous variants VCE (for cloning mutant under consideration, Hawaian unfiltered variants subtracted, lower quality variants included) 3,213 lmes, 36 comments format: vef, database: ec10 info: Picked up: JAVA.OPTIONS ~Djava.io.tmpdir=/spacer/g2main (Sat Riv c 24 23:19 05 ST 2012) net sf. pictraf. sam. CreateSequenceDictionary REFERENCE /space/g2main/tmp-gatk-3D9Rm/dict4827351121460120347.tmp up (i) display at UCSC main	*		
I.Chron 2:100 3:10 4:Mof 5:Alt 6:Qual 7:Filter #fileFormatwCF4.1 #fileFormatwCF4.1 #fileRomatwCF4.1 #f00MAT=0:0A0,Numbers1, Type:Stoper, Description="Altelic depths for the #f00MAT=0:0A0,Numbers1, Type:Thesp. Operation="Antelic feed topit #f00MAT=0:0A0,Numbers1, Type:Thesp. Operation="Genotype Quality"> #f00MAT=0:0A0,Numbers1, Type:Thesp. Operation="Genotype"> #f00MAT=0:0A0,Numbers1, Type:Thesp. Operation="Genotype")	oni G e ref		
43: Heterozygous and Homozygous variants (higher quality, coverage $>3$ , Hawaiian unfiltered variants subtracted for submission to databases or for variant subtraction)			
45: Uncovered replons apportated (spoFff)			

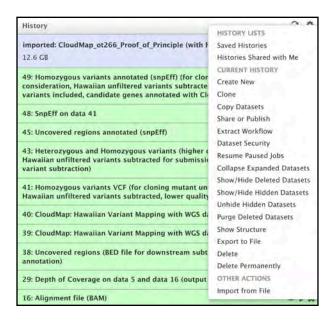
#### 15) Click the Start using this history link.

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	0	7	۰
imported: CloudMap_ot266_Proof_of_Principle (with hidden data) 12.6 GB		2	
49: Homozygous variants annotated (snpEff) (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included, candidate genes annotated with CloudMap)	۲	0	×
48: SnpEff on data 41	۲	0	×
45: Uncovered regions annotated (snpEff)	۲	0	×
43: Heterozygous and Homozygous variants (higher quality, coverage $>$ 3, Hawaiian unfiltered variants subtracted for submission to databases or for variant subtraction)		0	×
41: Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)	۲	0	×
40: CloudMap: Hawaiian Variant Mapping with WGS data on data 34	۲	0	×
39: CloudMap: Hawaiian Variant Mapping with WGS data on data 34	۲	0	×
38: Uncovered regions (BED file for downstream subtractions and snpEff annotation)	۲	0	×
29: Depth of Coverage on data 5 and data 16 (output summary sample)	۲	0	×
16: Alignment file (BAM)	۲	0	×
9: FASTQ quality statistics (box plot)	۲	0	×
5: WS220.64_chr.fa	۲	0	×
4: ot266_ProofOfPrinciple_Small.fastqsanger	۲	0	×
3: HA_SNPS_Unfiltered_112061Variants_WS220.vcf	۲	0	×
2: HA_SNPs_Filtered_103346Variants_WS220.vcf	۲	0	×
1: CloudMap_TranscriptionFactors_wTF2.2.txt	۲	0	*

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



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

Saved Histories						
search history names and tags						
Name Name	Datasets	Tags Sharing	Size on Disk	Created	Last Updated 1	Status
ot266 Subtract variants example 1~	4	D Taqs	97.6 Mb	– 1 hour ago	29 minutes ago	
imported CloudMap View	17	0 Tags	12.7 Gb	- 2 hours ago	~ 2 hours ago	current

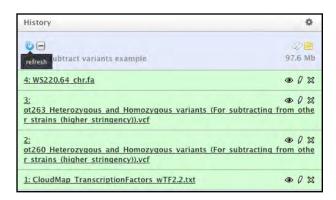
19) To copy the ot266 homozygous variants into this history, click Copy Datasets in your history options:



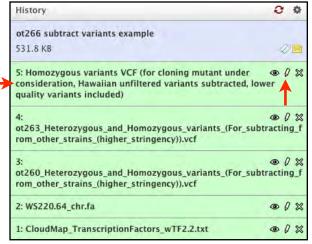
20) Copy the ot266 Homozygous variants VCF from the newly imported ot266 history:

Source History:	Destination History
1: imported 2	2: ot266 subtract variants +
1: CloudMap_TranscriptionFactors_wTF2.2.txt	
2: HA_SNPs_Filtered_103346Variants_WS220.vcf	Choose multiple histories
3: HA_SNPS_Unfiltered_112061Variants_WS220.vcf	
4: ot266_ProofOfPrinciple_Small.fastqsanger	New history named:
5: WS220.64_chr.fa	New history hamed.
9: FASTQ quality statistics (box plot)	
16: Alignment file (BAM)	
29: Depth of Coverage on data 5 and data 16 (output summary sample)	
<ul> <li>38: Uncovered regions (BED file for downstream subtractions and snpEff annotation)</li> </ul>	
39: CloudMap, Hawaiian Variant Mapping with WGS data on data 34	
40: CloudMap: Hawaiian Variant Mapping with WGS data on data 34	
✓ 41: Homozygous variants VCF (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)	
43: Heterozygous and Homozygous variants (higher quality, coverage > 3 Hawaiian unfiltered variants subtracted for submission to databases or for var subtraction)	
45: Uncovered regions annotated (snpEff)	
48: SnpEff on data 41	
49: Homozygous variants annotated (snpEff) (for cloning mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included, candidate genes annotated with CloudMap)	

21) Hit refresh in your history:



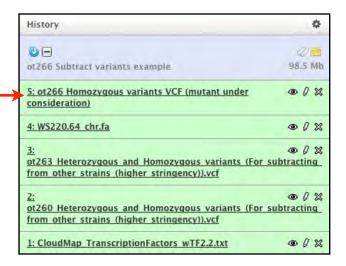
22) You will now see the ot266 Homozygous Variants (VCF) in your history. Click on the pencil icon to change the name of the file to add the ot266 prefix.



23) Add the ot266 prefix to the file name:

Attributes Convert Format Datatype Permissions	History O Ø			
Edit Attributes	ot266 subtract variants example 531.8 KB			
Hame: ot266 Homozygous variants VCF (for c	5: Homozygous variants VCF (for cloning mutant under db $\notD$ 22 consideration, Hawailan unfiltered variants subtracted, lower quality variants included)			
Info: Picked up_JAVA_OPTIONS: Djava.io.tmpdir=/space/g2main	4: ot263_Heterozygous_and_Homozygous_variants_(For_subtracting_fro m_other_strains_(higher_stringency)).vcf			
Annotation / Notes:	3: Or Ø ot260_Heterozygous_and_Homozygous_variants_(For_subtracting_ m_other_strains_(higher_stringency)).vcf			
Add an annotation or notes to a dataset: annotations are available when a history is viewed.	2: W\$220.64_chr.fa 🐵 Ø 38			
Database/Build: C. elegans Oct. 2010 (WS220/ce10) (ce	1: CloudMap_TranscriptionFactors_wTF2.2.txt 🔹 Ø Ø 🕱			
Number of comment lines:				
Score column for visualization: 2 3 4 5				
Save				
Auto-detect This will inspect the dataset and attempt to correct the above column values if they are not accurate.				

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



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:

Analyze Data	Workflow	Shared Data Visualization Cloud Help User
	R	Data Libraries Published Histories Published Workflows Published Visualizations Published Pages
	No	aw you can have a personal Galaxy within the infinite Universe

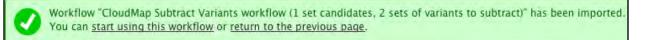
26) Select the CloudMap Subtract Variants workflow:

Published Workflows	
cloudmap 🐹	a.
Advanced Search	
Name	Annotation
CloudMap Subtract Variants workflor sets of variants to subtract)	w (1 set candidates, 2

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

Published Workflows   gm2123   CloudMap Subtract Variants workflow (1 set candidates, 2	2 sets of variants to subtract)
Galaxy Workflow ' CloudMap Subtract Variants workflow (1 set candidates, 2 s	ets of variants to subtract)'
Step	Annotation
Step 1: Input dataset	
Fasta reference select at runtime	
Step 2: Input dataset	
Candidate gene list select at runtime	
Step 3: Input dataset	
Variants for mutant under analysis (VCF file) (e.g. ot266 in CloudMap paper) select at runtime	
Step 4: Input dataset	
Variants to subtract 1 (VCF file) (e.g. ot260 or ot263 in CloudMap paper) select at runtime	
Step 5: Input dataset	
Variants to subtract 2 (VCF file) (e.g. ot260 or ot263 in CloudMap paper) select at runtime	
Step 6: Combine Variants	Merges variant files to be used for subtraction (Uniquify)
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	

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



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 cand	idates, 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.

Running workflow "imported: CloudMap Subtract Variants workflow Expand All Collapse	History 2 🌣
(1 set candidates, 2 sets of variants to subtract)" Step 1: Input dataset	ot266 subtract variants example 531.8 KB 🖉 😬
Step 1: input dataset       Fasta reference       4: WS20.64_chr.fa	5: ot266 Homozygous variants VCF (for cloning $\textcircled{O}$ (2) mutant under consideration, Hawaiian unfiltered variants subtracted, lower quality variants included)
type to filter	4: WS220.64_chr.fa 🐵 🖉 🕱
Step 2: Input dataset	3: ● ℓ ≈ ot263_Homozygous_and_Heterozygous_variants_(for_subtract ing_from_other_strains_(higher_stringency)).vcf
Candidate gene list 1: CloudMag_TranscriwTF2.2.txt + type to filter	2: ● Ø ≈ ot260_Homozygous_and_Heterozygous_variants_(for_subtract ing_from_other_strains_(higher_stringency)).vcf
Step 3: Input dataset	1: CloudMap_TranscriptionFactors_wTF2.2.txt
Variants for mutant under analysis (VCF file) (e.g. ot266 in CloudMap paper)	
Step 5: Input dataset	
Variants to subtract 2 (VCF file) (e.g. ot260 or ot263 in CloudMap paper)	
<u>Step 6: Combine Variants</u> (version 0.0.4) Merges variant files to be used for subtraction (Uniquify)	
<u>Step 7: Select Variants</u> (version 0.0.2) Subtracted variants (liberal, variants present in either subtraction strain removed)	
Step 8: Select Variants (version 0.0.2)	

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:

uccessfully ran workflow "imported: CloudMap Subtract Variants workflow (1 set candidates, 2 sets of variants to	History
ubtract)". The following datasets have been added to the queue: 4: WS220.64. chr.fa	ot266 subtract variants example
4: WS220.04_Chr.fa 1: CloudMap_TranscriptionFactors_WTF2.2.txt	531.8 KB
<ol> <li>Clouwag Traincipionractors wrz.c.kt</li> <li>cot266 Homozygous variants VCF (For cloning mutant under consideration, Hawaiian unfiltered variants subtracted, hower quality variants included)</li> </ol>	$\textcircled{0}$ 19: Annotated subtracted variants (conservative, only variants present in both $\emph{p}$ subtraction strains removed)
2: ot260_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strains_(higher_stringency)).vcf	🕲 18: SnpEff on data 14 🛛 🖉
3: ot263_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strains_(higher_stringency)).vcf	17: SnpEff on data 14
6: Merge of variants that will be used for subtraction 7: Combine Variants on data 2, data 4, and data 3 (log)	$\otimes$ 16: Annotated subtracted variants (liberal, variants present in either $g$ subtraction strain removed)
Subtracted variants (liberal, variants present in either subtraction strain removed)     Select Variants on data 4, data 5, and data 6 (log)	15: Select Variants on data 4, data 5, and data 10 (log)
10: Select Variants on data 4 and data 6 (Variant File) 11: Select Variants on data 4 and data 6 (log)	$\otimes$ 14: Subtracted variants (conservative, only variants present in both $$\mathcal{Q}$$ subtraction strains removed)
12: SnpEff on data 8	13: SnpEff on data 8
13: SnpEff on data 8 14: Subtracted variants (conservative, only variants present in both subtraction strains removed)	🕲 12: SnpEff on data 8 🕖
15: Select Variants on data 4, data 5, and data 10 (log)	11: Select Variants on data 4 and data 6 (log)
16: Annotated subtracted variants (liberal, variants present in either subtraction strain removed) 17: SnoEff on data 14	🕲 10: Select Variants on data 4 and data 6 (Variant File) 🛛 🖉
18: SnpEff on data 14	(3) 9: Select Variants on data 4, data 5, and data 6 (log)
19: Annotated subtracted variants (conservative, only variants present in both subtraction strains removed)	$\bigotimes$ 8: Subtracted variants (liberal, variants present in either subtraction strain $\rho$ removed)
	T: Combine Variants on data 2, data 4, and data 3 (log)
	sign 6: Merge of variants that will be used for subtraction
	5: ot266 Homozygous variants VCF (for cloning mutant under consideration, $ $
	4: WS220.64_chr.fa 🔹 Ø
	3: • • • • • • • • • •
	2: ot260_Homozygous_and_Heterozygous_variants_(for_subtracting_from_other_strain (higher_stringency)).vcf
	1: CloudMap_TranscriptionFactors_wTF2.2.txt

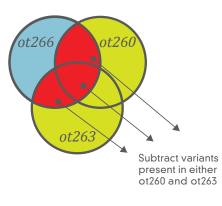
33) The workflow has finished running and you can view the resulting output:



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

History	0 0
ot266 subtract variants example 15.1 MB	HISTORY LISTS Saved Histories Histories Shared with Me
19: Annotated subtracted variants (conservative, only variants present strains removed)	CURRENT HISTORY Create New
16: Annotated subtracted variants (liberal, variants present in either si removed)	Clone Copy Datasets Share or Publish
8: Subtracted variants (liberal, variants present in either subtraction st	Extract Workflow
5: ot266 Homozygous variants VCF (for cloning mutant under conside unfiltered variants subtracted, lower quality variants included)	Dataset Security Resume Paused Jobs
4: WS220.64_chr.fa	Collapse Expanded Datasets Show/Hide Deleted Datasets
3: ot263_Homozygous_and_Heterozygous_variants_(for_subtracting_fron ency)).vcf	Show/Hide Hidden Datasets Unhide Hidden Datasets Purge Deleted Datasets
2: ot260_Homozygous_and_Heterozygous_variants_(for_subtracting_fron ency)).vcf	Show Structure Export to File
1: CloudMap_TranscriptionFactors_wTF2.2.txt	Delete Permanently OTHER ACTIONS Import from File

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.



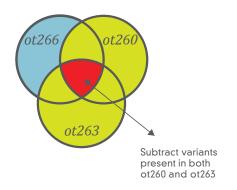
This file contains ot266 homozygous variants after both homozygous and heterozygous variants present in either ot260 or ot263 have been subtracted. This file should be downloaded to be easily viewed in its entirety. The first several lines in any VCF file are header lines starting with "#" so users who 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	В	C	D	E	F	G	Н	1	
1	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	rgSM
2	chri	62642		T	C	48.7	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	chrl	346149		T	A	85.7	7 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	chrl	369870	Q	C	T	48.0	B 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	chrl	369871	2	c	T	48.7	7 PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:80,6,0
6.	chri	663697	1	G	C	167.2	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	chrl	670146	a	G	А	36.43	B 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	chrl	670173		Т	С	36.4	B 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	chrl	671425	-	T	A	48.7	7 PASS	AC=2;AF=1.00;AN=2;DP=2;	GT:AD:DP:GQ:PL	1/1:0,2:2:6.02:80,6,0
10	chri	687402	6	Т	A	67.0	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	chrl	714649		С	G	67.7	B 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.

1 # Chromin	Position Reference	Change	Change	type Homozygous Qu	ality Coverage	Warnings	Gene ID	Gene name	e Bin_type Trancript_ID Exon_ID	Exon_Rank 8	ffect nld_AA/new_Old_codor/PCodo	on_Num Codon_Dege COS_size
21	62642 T	C	SNO	Hom	08.77	2	¥48G1C.4	PH3-1	protein_codi Y4861C.4.	0	OWNSTREAM: 8216 bases	
3.1	62642 1	C	(SNP)	Hom	48,77	2	'Y48G1C,5	¥48G1C.5	protein_codi Y48G1C.5		NTRON	3496
-1 1	62642 T.	C	SNP	Hom	48.77	2	Y48G1C.2	C58-1	protein_codi ¥48G1C.2.1		IPSTREAM: 9216 bases	
51	62642 1	C	SNP	Hom	48,77	2	V48G1C 2	¢38-1	protein_codi Y48GIC 2.2		PSTREAM: 9236 bases	
10 I	62642 1	C	SNP.	Hom	48.77	2	'Y48G1C.2	tsk-1	protein_codi Y48G1C Z.3	U	IPSTREAM: 9869 bases	
7.1	346149 T	A	SNP	Hom	85.77	3	Y48G1A.3	¥48G1A.3	protein_codi ¥48G1A.3	0	OWNSTREAM: 8304 bases	
8.1	346149 T	ă.	SNP	Hom	85.77	3	V48G1A1	¥48G1A.1	protein_codi Y48G1A.1		PSTREAM: 2389 bases	
M 4	346149 T	A	SNP	Hom	85.77	3	¥48G1A.6	mbtr-1	protein_codi Y48/51A.6b		NTRON	1656
101	346149 T.	A	SNP	Hom	\$5.77	3	Y48G1A-6	mbir-1	protein_cod/¥48G1A.6a	10	NTRON	1695
11.1	346149 T	A	SNP	Hom	85.77	3	Y48G1A.Z	¥48G1A.2	protein_codi ¥48G1A.2.2	0	PSTREAM: 1316 bases	
121 4	346149 T	A	SNP	Hom	85,77	3	'Y48G1A.2	¥48G1A.2	protein_codi ¥48G1A.2.1		IPSTREAM: 1323 bases	
18.1	369870 €	T	SNP	Hom	-18-05	2	R119.3	R119.3	protein_codi /119.3.1	0	OWNSTREAM: 3480 bases	
8.4.1	369870 C	T	SNP	Hom	48.08	2	R119,3	R119.3	protein_codi R119.3.2	0	IOWNSTREAM: 3704 bases	
1.7 4	369870 C	т	ISNP	Hom	45.08	2	R119.1	R119.1	protein_codi R119.1		PSTREAM: 5966 bases	
30 1	369870 C	T	SNP	Hom	48.08	2	R119.4	pgn-59	protein_codi 8119.4.1	0	OWNSTREAM: 7608 bases	
6.2. 1	369870 C	T	SNP	Hom	48.08	2	R119.2	R119,2	protein_codi R119.2		NTRON	1055
TR I	369870 C	T	SNP	Hom	48.08	2	B119.7	mp-8	pratein_codi R115.7	0	OWNSTREAM: 1359 bases	
19.1	369870 C	T	SNP	Hom	48.08	2	\$119.4	pgn-59	protein_codi R119.4.2	0	OWNSTREAM: 9377 bases	
20 1	369871 C	1	SNP	Hom	48.77	2	R119.3	R119,3	protein_codi R119.3.1	0	OWNSTREAM: 3479 binies	
10.1	369871 C	T	SNP	Hom	48.77	2	R119.3	R119.3	protein_codi #119.3.7	0	OWNSTREAM: 3703 bases	
22 1	369871 C	T	SNIP	Hom	48.77	2	R119.1	R119.1	protein_codi 8119.1	1	PSTREAM: 5967 bases	
23 1	369871 C	3	SNP	Hom	48.77	2	R119.4	pan-59	protein_codi R119.4.1	0	OWNSTREAM: 7607 bases	
23411	369871 C	T	15/04/	Hom	01.77	2	B110.2	#110.2	protein_codi 6110.2	10	NTRON	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.





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.

9.0.0			Galaxy				1.2
● 12 田 Henry Calendry Gen 12 田 Henry Calendry	/root				6 Q-64		0
Galaxy		Analyze Data Workflow	Shared Data - Visualiza				Using 0%
Tools	0					History	0
search tools						-	
Get Data			Galaxy 1	01		D bytes	and the second
Send Data			Galaxy			1 Your history is er	npty. Click 'Get
ENCODE Tools			Start sma			Data' on the left	pane to starf
Lift-Over			The very first tutorial	you need			
Text Manipulation			The resy mattered	Journeed			
Convert Formats							
FASTA manipulation							
Filter and Sort							
Join, Subtract and Group							
Extract Features			Live Quickies				
Fetch Sequences			Live Quickies				
Fetch Alignments							
Get Genomic Scores	Basic fastQ	Advanced fastQ	454 Mapping:	Uploading Data	Managing		
Operate on Genomic Intervals	manipulations	manipulation:	Single End Genetic subset (3)	Using FTP	account histories		
Statistics			Contra Galler 1 G				
Graph/Display Data							
Regional Variation							
Multiple regression	Galaxy is an open, w	eb-based platform for data int	ensive biomedical research.	Whether on this free public ser	rver or your own instance, you		
Multivariate Analysis		ice, and share complete analys					
Evolution		partments at Emory University te for CyberScience at Penn Sta		irted in part by NSE, NHGRI, TR	he Huck Institutes of the Life		
Motif Tools	sciences, the institut	te tor cyperscience at renn sta	te, and entitle university.				

### 1) Navigate to http://usegalaxy.org

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



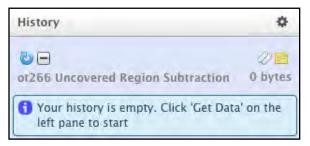
3) Once you are logged in using your email address, create a new history:



### 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).

Analyze Data	Workflow	Shared Data Visualization Cloud Help User
	R	Data Libraries Published Histories Published Workflows Published Visualizations Published Pages
	No	aw you can have a personal Galaxy within the infinite Universe

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

= Galaxy	Analyze Data Workflow Shared Data - Visualization-
Data Libraries	
Close Advanced Search       data library name:       Cloudmap       data library description:	
Data library name ↓	Data library description
1000 genomes	
100209 HsMtDNA	
anton_test	
bushman	
CloudMap	Contains reference and configuration files for the Cloudmap pipeline
Codon Usage Frequencies	
Dannon's Test Data Library	Testing library for Dannon
FRIK920	
GATK	

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

Galaxy	Analise Dilla Workflow Shared Data+ Vinustication+ Element and com			Using 86
Data Library "CloudMap"		Add datasets	Add folder	ibrary Acti
Contains reference and configuration files for the Cloudmap pipeline				1
Name	Message	Data type	Date uploaded	Files
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 ClaudMap EMS Variant Density Mapping tool			
🕈 🦲 at266 proof of principle dataset *	Use these files to run the CloudMap ot266 proof of principle example			
Hawailan SNP reference files unfiltered (WS220.64)				
T state of the	Use these BEDs for the CloudMap 0266 proof of principle for uncovered region subtraction			
ot260_Uncovered_regions.bed =		bed	2012-07-17	1.01
otZ63_Uncovered_regions.bed -		bed	2012-07-17	1.71
ot265_Uncovered_regions.bed =		bed	2012-07-17	54.0
🔹 🖭 🚞 ot260 and ot263 VCFs for variant subtraction —	Use these VCFs for the CloudMap ot266 proof of principle variant subtraction			
ot266_ProofOfPrinciple_Small.fastqsanger +	Sample FASTQ file for ot266 Proof of principle	fastqsanger	2012-06-27	2.2
HA_SNPs_WS220_Filtered_103626_SNPs_chr.bed =	Filtered set of Hawailan SNP positions (used by mpileup tool)	bed	2012-06-11	2.5
HA_SNPs_WS220_Filtered_103626_SNPs_chr.vcf -	Filtered set of Hawaiian SNP variants (used by CloudMap SNP Mapping with WCS tool)	vet	2012-06-11	431
- WS220.64_chr.fa -	WS220.64 genamic reference file	fasta	2012-06-11	97.6
SNP Mapping with WCS Data Other Species Config Files	Use these config files if you want to use the SNP Mapping with WGS Data for any species other than C elegans and Arabadopsis			
For selected datasets Import to current history Go				

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



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



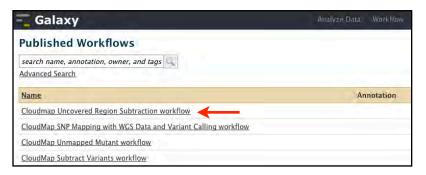
11) You will now see these files in your history:

History	\$
ot266 Uncovered Region 0 Subtraction	0 😑 bytes
3: ot266 Uncovered regions.bed	0 \$
2: ot263 Uncovered regions.bed	0 %
1: ot260_Uncovered_regions.bed	0 \$

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:



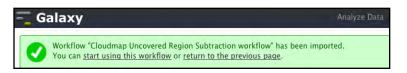
13) Select the Uncovered Region Subtraction workflow:



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

T Galaxy		Shared Data - Viscolization - Choud + Helpi-	
Published Workflows   gal40   Cloudmap Uncovered Region Subtraction	ı workflow		Import workflow
Galaxy Workflow ' Cloudmap Uncovered Region Subtraction	n workflow'		
Step		Annotation	
Step 1 Input dataset			
Uncovered regions for mutant under analysis (BED file) select at runtime			
Step 2: Input dataset			
Uncovered regions for subtraction 1 (BED file) select at runtime			
Step 3: Input dataset			
Uncovered regions for subtraction 2 (BED file) select at sustime			
Step 4: Intersect			
Return Overlapping pieces of Intervals			
of Output dataset 'output' from step 2			
that intersect Output dataset 'output' from step 3			
for at least 1			
Step 5: Subtract			
Subtract Output dataset 'output' from step 4			
from Output dataset "output" from step 1			

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



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*:

- Galaxy	Analyze Data
Your workflows	
Name	
imported: Cloudmap Uncovered Region Subtract	Edit
Cloudmap Uncovered Region Subtraction workfl	Run
CloudMap SNP Mapping with WGS Data and Vari CloudMap Unmapped Mutant workflow = CloudMap Subtract Variants workflow =	Share or Publish Download or Export Clone Rename
Workflows shared with you by No workflows have been shared with you. Other options	View Delete
Configure your workflow menu	

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.

Analyze Data Workflow Shared Data - Visualization - Cloud - Help - User-		Using 86%
Running workflow "imported: Cloudmap Uncovered Region Subtraction workflow"	Expand All Collapse	History Ø
Step 1: Input dataset Uncovered regions for mutant under analysis (BED file)		ot266 Uncovered Region D bytes Subtraction
3. ot266_Uncovered_regions.bed 2		3: at 266 Uncovered regions.bed
Step 2: Input dataset		Z: ot263_Uncovered_regions.bed
Uncovered regions for subtraction 1 (BED file) 1: ot260_Lincovered_regions.bed		1: It at 260_Uncovered_regions.bed
Step 3: Input dataset		
Uncovered regions for subtraction 2 (BED file)		
Step 4: Intersect		
Step 5: Subtract		
Step 6: SnpEff		
Send results to a new history.		
Run workflow		



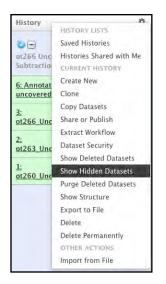
18) All of the automated functions have the appropriate default parameters configured, although experienced users may want to modify these prior to running. Once you are ready to run the workflow, press Run Workflow and the workflow will start (this step takes a minute or two to begin, be patient and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

Successfully ran workflow "imported: Cloudmap Uncovered Region Subtraction workflow". The following datasets have been added to the queue:	History
Succession y an worknow imported. Coolump Uncovered region Subtraction worknow - The following bacasets have been added to the guede:  Coolumn and the subtraction worknow in the subtraction subtraction worknow - The following bacasets have been added to the guede:  Coolumn and the subtraction subtraction subtraction subtraction so that subtrac	Contraction Subtraction Contr
6: Annotated subtracted uncovered regions 7: SnpEff on data 5	S. Subtracted subtracted with a set of a se
	<ul> <li>4: Common regions</li> <li>0 2</li> <li>uncovered in strains used for subtraction</li> </ul>
	3: ot266 Uncovered regions.bed
	2:
	1: ot260 Uncovered regions.bed

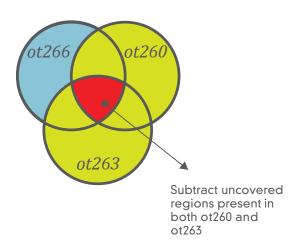
19) The workflow has finished running and you can view the resulting output:

1. ot260_Uncovered_regions.bed     at266_Uncovered Region_2       2: ot263_Uncovered_regions.bed     Subtraction       4: Common regions uncovered in strains used for subtraction     6: Annotated subtracted uncovered regions (not annotated)       5: Subtracted uncovered regions     3:	successfully ran workflow "imported: Cloudmap Uncovered Region Subtraction workflow". The following datasets have been added to the queue: 3: ot266. Uncovered, regions.bed	00
2: 0253 Uncovered regions.bed 4: Common regions uncovered in strains used for subtraction 5: Subtracted uncovered regions (not annotated) 6: Annotated subtracted uncovered regions 7: SnpEff on data 5 2: ••••••••••••••••••••••••••••••••••••	1: ot260_Uncovered_regions.bed	at266 Uncovered Region 2.7
5: Subtracted uncovered regions (not annotated) 6: Annotated subtracted uncovered regions 7: SnpEff on data 5 2: @	2: ot263_Uncovered_regions.bed	Subtraction
Subtracted uncovered regions into annotated     Subtracted uncovered regions     Annotated subtracted uncovered regions     SpipEff on data 5     2:	4: Common regions uncovered in strains used for subtraction	
7: SnpEff on data 5	5: Subtracted uncovered regions (not annotated)	uncovered regions
2: •	6: Annotated subtracted uncovered regions	
	7: SnpEff on data 5	ot266_Uncovered_regions.bed
ot263 Uncovered regions.bed		
		ot263 Uncovered regions.bed

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



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



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

12	A	Ц	C I	D	E	F	6		1.		ĸ	L	Contract A Desire	- 14	a	.p.	Q
L	# Chromo	Position	Reference	Change	Change_ty	pe Homozygia	as Quality	Coverage	Warnings	Gene_ID	Gene_name	a Bio_type	Trancipt_I	Exon_ID	Exon_Rank	Effect	old_AA/new_AA
2	E.	2646	766	1		Interval			0	0	Y74C9A.2	nlp-40	protein_con	1 Y74C9A.2.4			UPSTREAM: 8859 bases
3	E.	2646	2664	1		Interval			0	0	Y74C9A.2	ntp-40	protein_con	1 Y74C9A.2.6			UPSTREAM 8972 bases
÷.	1 I I I I I I I I I I I I I I I I I I I	2646	2664	1		Interval			U	0	Y74C9A.2	nlp-40	protein_coo	II Y74C9A.2.3			UPSTREAM: 7767 bases
10	í	2646	2664	2		interval			0	0	Y74C9A.2	nlp-40	protein_coo	1 Y74C9A.2.2			UPSTREAM: 8849 bases
0	1	2646	2664	1		Interval			0	0	¥74C9A.2	nlp-40	protein_coo	1 Y74C9A.2.1			UPSTREAM: 8853 bases
1.1	1	2646	2664	1		interval			0	0	Y74C9A.2	nlp-40	protein_cos	ii Y74C9A.2.5			UPSTREAM: 8853 bases
ak .	1	2646	2664	1		interval			0	0	V74C9A.3	¥74C9A.3	protein cod	1 Y74C9A.3.1			DOWNSTREAM: 1473 bases
9	1	2646	2664	1		interval			0	0	Y74C9A.3	V74C9A.3	protein_cod	ii Y74C9A 3.2			DOWNSTREAM: 1575 bases
10	1	2646	2664	1		Interval			0	0	Y74C9A.6	¥74C9A.6	MORNA	Y74C9A.6			DOWNSTREAM: 1101 bases
01	1	3468	348;	2		Interval			0	0	Y74C9A.2	nlp-40	protein_coo	I Y74C9A.2.4			UPSTREAM: 8037 bases
12.5	1	3468	348	2		Interval			0	0	¥74C9A.2	nlp-40	protein_cod	1 Y74C9A.2.6			UPSTREAM: 8150 bases
12	1	3468	3483	2		Interval			0	0	¥74C9A.2	01-qin	protein_cos	1 Y74C9A.2.3			UPSTREAM: 6945 bases
14	1	3468	3483	1		Interval			0	0	¥74C9A.2	nlp-40	protein_cod	1 Y74C9A.2.2			UPSTREAM: 8077 bases
152	t	3468	3482	1		Interval			0	0	¥74C9A.2	nlp-40	protein_cod	1 Y74C9A.2.1			UPSTREAM 8031 bases
16	1	3468	3482	2		Interval			0	0	¥74C9A.2	nip-40	protein_cod	1 Y74C9A 2.5			UPSTREAM: 8031 bases
19	1	3468	3483	1		Interval			0	0	¥74C9A.3	Y74C9A.3	protein_cod	1 Y74C9A.3.1			DOWNSTREAM: 651 bases
1.6.	1	3468	3482	2		Interval			0	0	¥74C9A.3	Y74C9A.3	protein_cod	1 Y74C9A.3.2			DOWNSTREAM: 753 bases
19	1	3468	348	2		Interval			0	0	Y74C9A.6	¥74C9A.6	snoRNA	¥74C9A.6			DOWNSTREAM: 279 bases
20		3926	401/	1		interval			0	0	¥74C9A.2	nip-40	protein cod	1. 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 quide.

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.

🔤 Galaxy	Ar	nalyze Data Workflow Sh	ared Data ~ Visualization ~ Cloud ~ Help ~ User ~		Usic	ng 87%
Tools.	Upload File (version 1.1.3)			1	History	0
search tools Search tools Upload File from your computer UCSC Test table browser UCSC Archaea table browser Bit Mail browser Eiti SRA ENA SRA Get Microbial Data BibMan Central server BibMan Cest server	URL/Text:		To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).		Urnamed history  Your history is empty. Click Data' on the left pane to sta	0 byte
<u>CBI Rice Mart</u> rice mart	Files uploaded via FTP: File	Size	Date			
GrameneMart Central server	ot260_lane5.fastg	3.8 Gb	07/15/2012 09:44:58 AM			
<ul> <li>modENCODE fly server</li> </ul>	ot260_lane6.fastq	3.9 Gb	07/15/2012 10:14:22 AM			
Ehvning server     Elvning tests server     Elvning tests server     Mathing server     YeastMing server     mathabiltMing server     mathACCODE worm server     mathACCODE worm server     WormBase server	This Galaxy server allows you to upload files via FTP. T password). Ornert spaces to tabs: Via Use this option if you are entering intervals by hand. Genome: C. elegans Oct. 2010 (VS220/ce10) (ce10) V Execute	a upload some files, log in to	the FTP server at test.g2.lax.psu.edu using your Galaxy credentials (email address and			

4) The files will be added to your history:

The following job has been successfully added to the queue:	History	0
1: ot260_lane5.fastq	🔁 🖃 Unnamed history	0 bytes
2: ot260_lane6.fastq You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from	n 2: ot260 lane6.fastq	@ 0 X
'running' to 'finished' if completed successfully or 'error' if problems were encountered.	n 1: ot260 lane5.fastq	• 1 %

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:

History	\$
Unnamed history	0 bytes
2: ot260 Jane6.fastq	
1: ot260 lane5.fastq 3.8 Cb format: fastq, ostabase: ce10	
mfo: uploaded fastq file	4 <b>1</b>
@EAGLE:5:1:0:92/1 A.	c.a
@EAGLE:5:1:0:126/1	*****
A	

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

Analyze Data Workflow Shared Data Visualization Cloud Help User-		Using 87%
Edit Attributes	History	0
Name:	Unnamed history	0 bytes
ot260_lane5.fastq	2: ot260 lane6.fastq	
uploaded fastq file	1: ot260 lane5.fastg 3.8 Gb	
Annotation / Notes: None	format: fastq, database: ce10 Info: uploaded fastq file	
Add an annotation or notes to a dataset; annotations are available when a history is viewed.	@EAGLE:5:1:0:92/1	
Database/Suild: C. elegans Oct. 2010 (WS220/ce10) (ce:	A+	G
(Save)	*************************	*****
Auto-detect	@EAGLE:5:1:0:126/1	
This will inspect the dataset and attempt to correct the above column values if they are not accurate.	C	<b>)</b> () ()
Convert to new format		
Convert FASTQ files to seek locatic + This will create a new dataset with the contents of this dataset converted to a new format.		
(Convert)		
Change data type		
New Type:	<b>U</b>	
fastqsa V		
fastqsanger         ting dataset but not modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.           (Save)         Save		
Jave		

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

History	\$
C - Unnamed history	2 📑 7.7 Gb
2: ot260 lane6.fastq	• 0 %
1: ot260 lane5.fastq 3.8 Cb	• 0 %
format: fastqsanger, latabase: ce10 Into: upioaded tastq file	
	20
@EAGLE:5:1:0:92/1	
A	G
######################################	*******
(	)4(F)

### **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:

- Galaxy		Analyze Data	Workflow Sl	ihared Data~	Visualization	Cloud	Help+ User	ir -			Using 91%
Tools 🔅	Concatenate datasets (version 1.0.0)							-	1	History	¢
Conca	Concatenate Dataset: 1: ot260_lane5.fastq 🛟									Unnamed history	<b>2.7 G</b> b
<u>Concatenate datasets</u> tail-to-head	Datasets									2: ot260 lane6.fastg	002
Join, Subtract and Group Group data by a column and perform aggregate operation on other columns.	Dataset 1 Select: 2: ot260_lane6.fastq +									1: ot260 lane5.fastg	• 0 z
Convert Formats AXT to concatenated FASTA Converts an AXT formatted file to a concatenated FASTA alignment	Remove Dataset 1 Add new Dataset										
MAF to FASTA Converts a MAF formatted file to FASTA format	Execute										

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 singleend 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

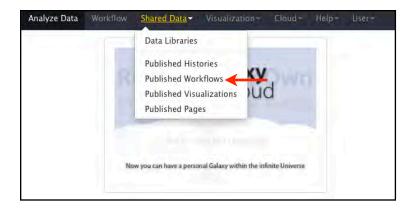
1) Let's assume that you haven't yet imported any CloudMap workflows. Navigate to http:// usegalaxy.org/



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:



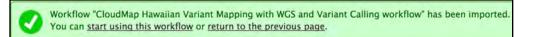
4) Click Published Workflows on the menu bar to access the automated workflow. Select the CloudMap Hawaiian Variant Mapping with WGS Data and Variant Calling workflow.

Published Workflows	
Cloudmap	Q.
Advanced Search	
Name	
CloudMap Hawaiian Variant Mapping with WG	S and Variant Calling workflow
CloudMap Unmapped Mutant workflow (w/ su	ubtraction of other strains)
CloudMap EMS Variant Density Mapping work subtract)	flow (takes VCF of heterozygous and homozygous variants to
CloudMap Unmapped Mutant workflow	

5) You will now have the option to Import workflow

- Galaxy	Analyze Data	Workflow	Shared Data -	Visualization	Claud +	Help - Liser+
Published Workflows   gm2123   Cloud!	tap Hawailan Variant I	Mapping with	WGS and Varian	t Calling work	flow	O Import workflow
Galaxy Workflow ' CloudMap Haw	aiian Variant Map	ping with \	WGS and Varia	nt Calling w	orkflow'	
Step				Annotation		
Step 1: Input dataset						
Filtered mapping strain VCF (e.g. 103, select at runtime	46 Hawailan SNPs)					
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 select at runtime	factors)					
Step 5: Input dataset			100			
Unfiltered mapping strain VCF (e.g. 11 select at runtime	2,061 Hawaiian SNPs)		2			
Step 6: Map with BWA for Illumi	na					

6) You will see this message:



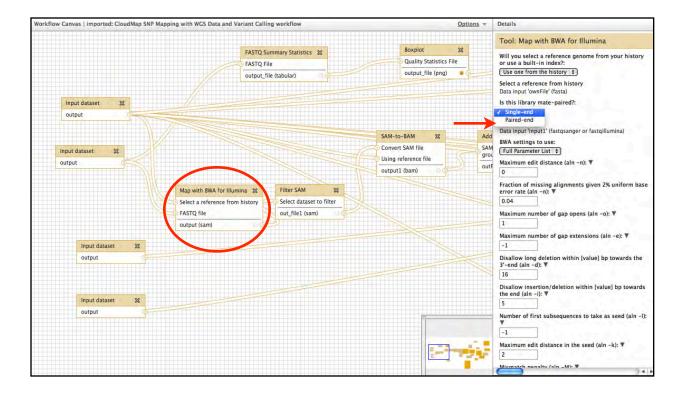
7) Click Start using this workflow and you will see that the workflow has been imported. From now on, you can easily access this workflow under the Workflow tab or in the Galaxy tools section (left frame of the browser window) under Workflows .

Your workflows	Create new workflow	TUpload or import workflow	
Name		# of Steps	
imported: CloudMap Hawaiian Variant Mapping with WCS and Variant Calling workflow $\neg$		29	

8) Click on the workflow and select Edit:



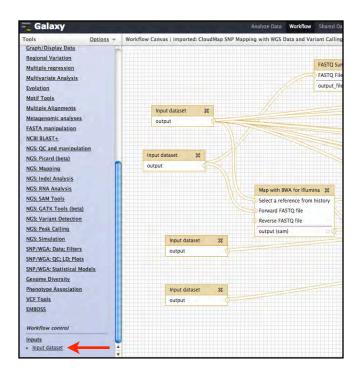
9) You will now see the workflow canvas that displays all the tools and input datasets in the workflow. By clicking on a given tool, you can change its parameters in the right frame of your browser window. We want to change the BWA mapping tool to accept paired-end data so we select the mapping tool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



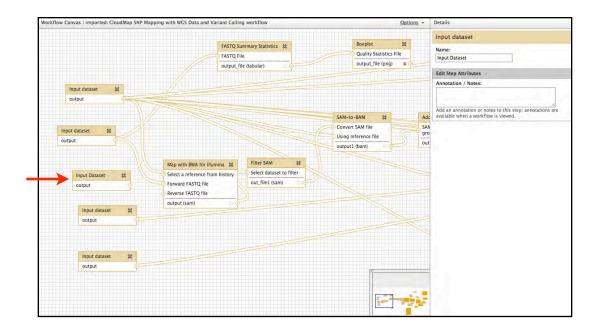
10) Once you select *paired-end* as the data type, the BWA mapping tool will now expect another input dataset.

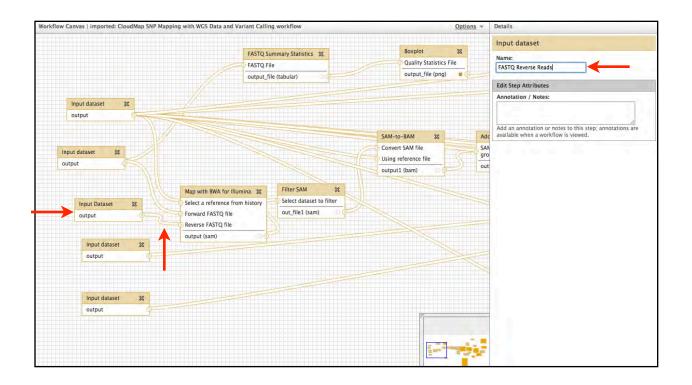
w Canvas   imported: CloudMap SNP Mapping v	ith WGS Data and Variant Calling workflow	Options
		Boxplot 💥
	FASTQ Summary Statistics 🕱	
	FASTQ File	Quality Statistics File
	output_file (tabular)	output_rile (phg)
Input dataset 🛛 🕱		
output		
London 1		
		SAM-to-BAM 💥
		Convert SAM file
Input dataset 🗱	E CONTRACTOR CONT	Using reference file
output		output1 (bam)
NI.		outputs (outing
	ap with BWA for Illumina 💥 Enter SAM 🛛 💥	
(1×65	elect a reference from history Select dataset to filter	
	orward FASTQ file out_ile1 (sam)	
	everse FASTQ file	
	utput (sam)	
Input dataset 🛛 🕱		
output		
Catpor		
Input dataset 🛛 🕱		
output		

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*.

	Analyze Data Workflow Shared Da	ta- Visualization- Cloud- Help	a÷ User≠		Using C
ow Canvas   imported: CloudMap SNP Map	oping with WGS Data and Variant Calling	workflow		Save	
	FASTQ Sun PASTQ File output, file		Boxplot Quality Statisti output_file (pn	Run Edit Attributes Auto Re-layout	dataset 2 Reverse Reads
					Edit Step Attributes
Input dataset 🛛 💥					Annotation / Notes:
Input dataset 22 output			SAM-to-BAM & Convert SAM file Using reference file	Adc SAN gro	Add an annotation or notes to this step; annotations an available when a workflow is viewed.
output		The second secon	output1 (bam)	out	
	Map with BWA for Illumina 🕺	Filter SAM 88			
Input Dataset & output	Select a reference from history Forward FASTQ file Reverse FASTQ file output (sam)	Select dataset to filter out_file1 (sam)			
output					
Input dataset 🛛 🕱					
output			\$*		

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>

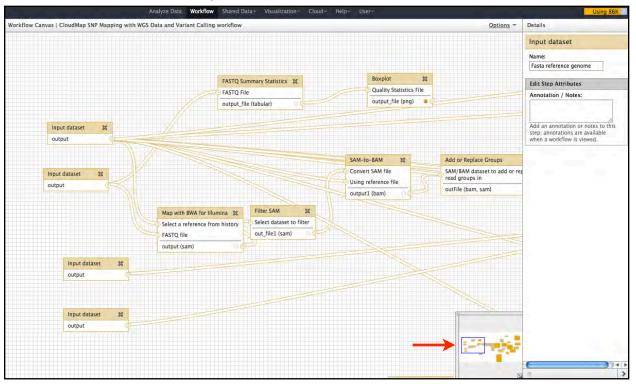
><number>

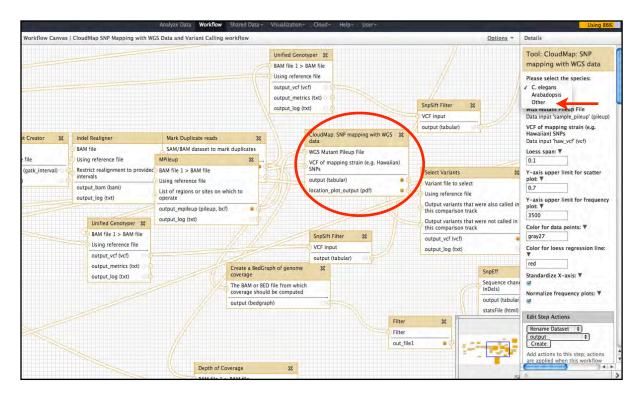
i.e.: >CHROMOSOME\_1 >CHROM\_1 >1

Sample D.rerio configuration file:

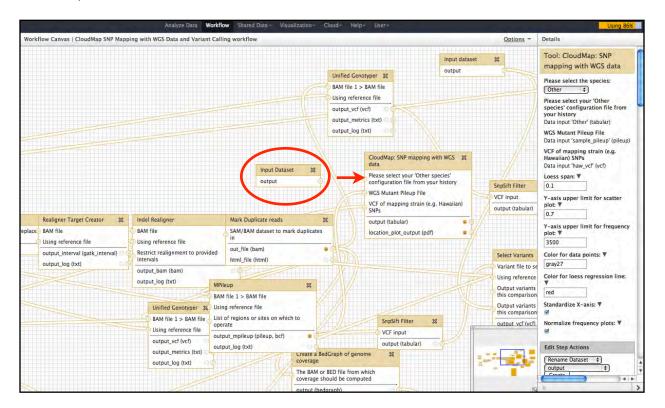
Please see more sample **Other species** configuration files in the CloudMap data library in the Hawaiian Variant Mapping with WGS Data Other Species Config Files folder.

- 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.

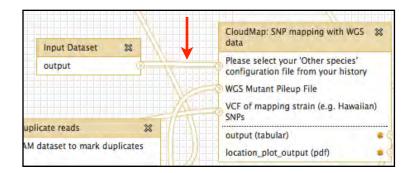




4) Select the CloudMap Hawaiian Variant Mapping with WGS Data tool, then select Other from species list.



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.

Shared Data - Visualization - Cloud - H	elp+ User+			Using 86%
orkflow		Input dataset	Save Run Edit Attributes	s : CloudMap: SNP ping with WGS data
BAM file Using ref output_v	netrics (txt)	output	Auto Re-layout Close	e select the species: Other = Please select your 'Other species' configuration file from your history Data input 'Other' (tabular) WGS Mutant Pileup File Data input 'sample_pileup' (pileup) VCF of mapping strain (e.g.
Input Dataset 28 output rk Duplicate reads 28	CloudMap: SNP mappi data Please select your 'Ott configuration file from WGS Mutant Pileup File VCF of mapping strain SNPs output (tabular)	ner species' n your history e	SnpSift Filter VCF input output (tabular)	Hawaiian) SNP5 Data input 'haw_vcf' (vcf) Loess span: V 0.1 V-axis upper limit for scatter plot: V 0.7
M/BAM dataset to mark duplicates	location_plot_output (	pdf)		Y-axis upper limit for frequency plot:

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://seganswers.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.

Galaxy	Ahalyze Data Workflow	Shared Data - Visioninzation-	Elaad - Help- User+		
Published Workflows	_	Data Libraries			
search name, annotation, owner, and tags		Published Histories			
Advanced Search		Published Workflows			
Name	Annotation	Published Visualizations Published Pages	Community Rating	Community Tags	Last Updated 1
Cloudmap Uncovered Region Subtraction workflow		ga140	****		- 21 hours ago
CloudMap SNP Mapping with WGS Data and Variant Calling workflow		gal40			2 days ago
CloudMap Unmapped Mutant workflow		ga140	43333		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?

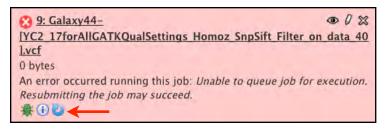
You can email us with your request at <u>gm2123@columbia.edu</u> or <u>or38@columbia.edu</u>. If you want to make the change yourself and run the tool locally, you can download the source code from the Galaxy Tool Shed at: http://toolshed.g2.bx.psu.edu/

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.

8 9: Galaxy44-	
[YC2 17forAllGATKQualSettings Homoz SnpSift Filter on	data 40
l.vcf	
0 bytes	
An error occurred running this job: Unable to queue job for e	xecution.
Resubmitting the job may succeed.	
* 🔁 🕑	

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

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

Edit Attributes	History				
Lun Attributes	# Command line: SnpEff eff -c /galaxy/home/g2test/galaxy_test/tool-d				
Name:	st_pool/pool1/files/000/304/dataset_304498.dat				
ot266_ProofOfPrinciple_Small.fastqsan	# Chromo Position Reference Change Change_type				
Info:	interval_ID I 42899 G A SMP Hom 75.03				
uploaded fastq file	7 45933 V A Mr. Hun 13103				
Annotation / Notes:	34: Uncovered regions annotated (snpEff) 🔹 Ø 🕫				
None	32: Homozygous variants not annotated (snpEff.) @ 0 22				
Add an annotation or notes to a dataset; annotations are available when a history is viewed. Database/Build:	31: Homozygous and heterozygous variants VCE (higher $\oplus 0$ 32 stringency, for downstream subtraction steps) (snpEff)				
C, elegans Oct. 2010 (WS220/ce10) (c	30: Uncovered regions (for downstream subtractions) @ Ø 2				
Save	29: Homozygous variants VCE (mutant under 🔹 🖉 🗴				
Auto-detect . This will inspect the dataset and attempt to correct the above column values if they are not accurate.	24: Depth of Coverage on data 3 and data 14 (output @ 0 20 summary sample)				
Convert to new format	14: Alignment file (BAM) 🐵 🖉 😫				
Convert FASTQ files to seek locatic +	6: Fastq statistics file and 0 2				
This will create a new dataset with the contents of this dataset converted to a new format.	3: WS220.64_chr.fa				
Convert	2: ot266 ProofOfPrinciple Small.fastqsanger @ 0 22				
Change data type	2,2 Cb format: fastgsanger, database: ce10				
New Type:	Info: uploaded fastq file				
fastqsanger					
This will change the datatype of the existing dataset but not modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.	@EAGLE:1:26:1248:15526/1 ATTTTTCCCGTATTCGCACAACACTTCTCATGTCTCAACCCTCACTGCCACATTTGACGCAATTCAACA				
Save					

#### 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.