CloudMap: A Cloud-Based Pipeline for Analysis of Mutant Genome Sequences

Gregory Minevich,*[,1] Danny S. Park,* Daniel Blankenberg,[†] Richard J. Poole,*[,1,2] and Oliver Hobert*[,1]
*Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical Institute, Columbia University Medical Center, New York, New York 10032, and [†]Center for Comparative Genomics and Bioinformatics, Penn State University, University Park, Pennsylvania 16802

ABSTRACT Whole genome sequencing (WGS) allows researchers to pinpoint genetic differences between individuals and significantly shortcuts the costly and time-consuming part of forward genetic analysis in model organism systems. Currently, the most effort-intensive part of WGS is the bioinformatic analysis of the relatively short reads generated by second generation sequencing platforms. We describe here a novel, easily accessible and cloud-based pipeline, called CloudMap, which greatly simplifies the analysis of mutant genome sequences. Available on the Galaxy web platform, CloudMap requires no software installation when run on the cloud, but it can also be run locally or via Amazon's Elastic Compute Cloud (EC2) service. CloudMap uses a series of predefined workflows to pinpoint sequence variations in animal genomes, such as those of premutagenized and mutagenized *Caenorhabditis elegans* strains. In combination with a variant-based mapping procedure, CloudMap allows users to sharply define genetic map intervals graphically and to retrieve very short lists of candidate variants with a few simple clicks. Automated workflows and extensive video user guides are available to detail the individual analysis steps performed (http://usegalaxy.org/cloudmap). We demonstrate the utility of CloudMap for WGS analysis of *C. elegans* and *Arabidopsis* genomes and describe how other organisms (*e.g.*, *Zebrafish* and *Drosophila*) can easily be accommodated by this software platform. To accommodate rapid analysis of many mutants from large-scale genetic screens, CloudMap contains an *in silico* complementation testing tool that allows users to rapidly identify instances where multiple alleles of the same gene are present in the mutant collection. Lastly, we describe the application of a novel mapping/WGS method ("Variant Discovery Mapping") that does not rely on a defined polymorphic mapping strain, and we integrate the application of this method into CloudMap. CloudMap tools and documentation are continually updated at http://usegalaxy.org/cloudmap.

WHOLE genome sequencing (WGS) represents the fastest and most cost-effective way to map phenotype-causing mutations in model organisms such as *Caenorhabditis elegans* (Hobert 2010). However, analysis of the resulting data is complex and requires specialized bioinformatics knowledge not readily available in most labs. Furthermore, the flood of WGS data has raised new concerns about both computing power needs and data storage capacities. Researchers may be unwilling to commit resources to computers or software in the fear that they may be quickly replaced or will not be interoperable with existing or future systems. As WGS costs continue to plummet and the technology becomes pervasive, all laboratories that use genetic analysis will be faced with these problems.

The basic premise of genetic mapping is simple: out of the millions of base positions in a mutagenized, sequenced genome, we aim to find the region of genome that is linked to the phenotype-causing mutation and identify the causal variant. Our lab has previously developed single-step SNP mapping strategies coupled with whole genome sequencing (Doitsidou *et al.* 2010) as well as software analysis tools (MAQGene) for mutant genome sequence analysis (Bigelow *et al.* 2009). Although MAQGene has been broadly used by labs in the *C. elegans* community (Flowers *et al.* 2010; Sarin

doi: 10.1534/genetics.112.144204
Manuscript received July 21, 2012; accepted for publication September 25, 2012
Supporting information is available online at http://www.genetics.org/content/suppl/2012/10/08/genetics.112.144204.DC1.
[1]Corresponding authors: Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical Institute, Columbia University Medical Center, New York, New York 10032. E-mail: gm2123@columbia.edu; Department of Cell and Developmental Biology, University College London, London WC1E 6BT, United Kingdom. E-mail: r.poole@ucl.ac.uk; and Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical Institute, Columbia University Medical Center, New York, New York 10032. E-mail: or38@columbia.edu
[2]Present address: Department of Cell and Developmental Biology, University College London, London WC1E 6BT, United Kingdom.

et al. 2010; Zhang *et al.* 2011; Kim *et al.* 2012; Labed *et al.* 2012), we no longer support it because the pipeline relies on an outdated aligner (MAQ) and requires technical expertise to install, which inevitably limits its general adoption.

In an effort to take advantage of cloud computing and many freely available open source tools, we have designed a new mutant genome sequence analysis pipeline to run on the Galaxy platform (Afgan *et al.* 2011). Our pipeline uses custom Python scripts to provide greatly improved mutant mapping tools and relies on the Next Generation Sequencing (NGS) Toolbox software suite in Galaxy, along with other software adapted for Galaxy. The pipeline is browser based, requires no software installation (when run on the cloud), and is modular and thus able to accommodate new tools when available. In addition to mapping mutations in *C. elegans* using mapping strains like the polymorphic Hawaiian strain CB4856, CloudMap can be used to support similar mapping strategies for any model organism that can be crossed to a polymorphic mapping strain. We also show how CloudMap can be used to apply a novel, variant-based mapping method ("Variant Discovery Mapping") for WGS-based mutant identification. We demonstrate native CloudMap support for *Arabidopsis* and show how other organisms can easily be accommodated with no changes to the software. We anticipate that as more biologists use these cloud-based tools to process their own WGS data, more intellectual cross-pollination will occur between biologists and bioinformaticians, resulting in many new tools for the Galaxy platform.

Materials and Methods

The reader is referred to the online user guides and videos (http://usegalaxy.org/cloudmap) and the proof-of-principle examples described below. These materials will be continuously updated.

Strains

For the proof-of-principle application of CloudMap, the strains OH4254 *ot266*; *vtIs1[dat-1*::*gfp*; *rol-6(d)]*, OH4240 *ot260*; *vtIs1[dat-1*::*gfp*; *rol-6(d)]*, and OH4247 *ot263*; *vtIs1[dat-1*::*gfp*; *rol-6(d)]* were used. FASTQ files from the mutant strains *fp6* and *fp9* were kindly provided by S. Jarriault and were used for the EMS Variant Density Mapping tool demonstration.

For the Variant Discovery Mapping proof of principle, we used the strain OH4254 *ot266*; *vtIs1[dat-1*::*gfp*; *rol-6(d)]*.

Tool settings

We used default settings for all of the tools except where otherwise noted. Users are strongly encouraged to read documentation for each tool to suit their needs. Custom tool settings are described below, in the user guide, and are also defined in the CloudMap published workflows available at: http://usegalaxy.org/cloudmap.

The mapping quality and base quality thresholds for Hawaiian Variant Mapping, unmapped mutant analysis, submission of variant lists to WormBase (www.wormbase.org), and other community databases follow.

Lenient variant lists: (Parameters applied to the sample being subtracted from, *i.e.*, the mutant strain being analyzed.) To accommodate low-quality/low-coverage WGS data and to ensure the causal variant is not accidentally removed during variant subtraction, the mapping quality and base quality thresholds for the mutant being analyzed are more lenient than for samples used for variant subtraction. WGS alignment data were filtered for reads with PHRED-based mapping score >10 and individual base quality scores at a given variant position were filtered for PHRED-based quality >17. No read depth filter was used.

Stringent variant lists: (Parameters applied to samples used in variant subtraction in both mapping methods, to identify heterozygous and homozygous positions in Variant Discovery Mapping, for Hawaiian variant filtration, and for submission to various databases such as WormBase.) Reads required a PHRED-based mapping score >30 (~1/1000 chances of mismapping to another location of the genome) (Li *et al.* 2008) and individual base quality scores at a given variant position were filtered for PHRED-based quality >30 (~1/1000 chances of being called incorrectly) (Ewing and Green 1998). For a position to be considered, read depth had to be ≥3. For Variant Discovery Mapping heterozygous and homozygous position data were further filtered for a PHRED-based QUAL score of ≥200 assigned to each variant by the Genome Analysis Toolkit (GATK) Unified Genotyper. We empirically determined this value as working best across several samples and several organisms. The CloudMap Variant Discovery workflow produces output files filtered at Q100, Q200, and Q300 and the CloudMap Variant Discovery tool can easily be rerun with any of the three different quality output files.

In silico complementation testing tool: *Liberal subtraction strategy:* Variant call format (VCF) files (Danecek *et al.* 2011) containing both homozygous and heterozygous variants from all samples should be used as input to the Combine Variants tool with the option "Combine variants and output site only if variant is present in at least N input files" set to "2." This VCF file of liberally defined common variants is then subtracted from the VCF for each individual sample using the GATK Select Variants tool with this common variant VCF output file as an input to the parameter "Output variants that were not called in this comparison track."

Conservative subtraction strategy: The GATK Combine Variants tool should be run with the option "Combine variants and output site only if variant is present in at least N input files" set to a number ranging anywhere from half the number of total samples, all the way to the most conservative subtraction strategy, where this parameter would be set to the total number of samples under consideration. Stringently filtered homozygous and heterozygous variants from

all samples can be used as input for the GATK Combine Variants tool. This VCF file of conservatively defined common variants is then subtracted from the VCF for each individual sample using the GATK Select Variants tool with the earlier output file from the Combine Variants step used as an input to the parameter "Output variants that were not called in this comparison track."

Results and Discussion

Overview

While WGS is the preferred method to map and clone mutations from forward genetic screens, there is currently no free, easy-to-use, non–model-organism–specific bioinformatics tool to analyze this type of WGS data. We have therefore developed CloudMap, a Galaxy-based pipeline that allows end users to go from raw sequencing data (from any next-generation sequencing platform) to a specific map position and to a small list of potential candidate variants in a few simple steps. The overall conceptual strategy is schematically depicted in Figure 1.

The CloudMap pipeline is entirely browser- and cloud-based, meaning no software installation is required (as discussed below, CloudMap can also be run locally or via Amazon's Elastic Compute Cloud service). Data are uploaded to secure individual user accounts and analyzed on servers running Galaxy software (hosted at Penn State University among other locations; list of available Galaxy servers in Table 1) and all of the steps required for common mutant analysis functions can be sequentially executed online with a few simple clicks as part of Galaxy workflows (Figure 2). 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. Documentation of common use case scenarios is provided both in the form of pdf user guides and as screen capture videos at http://usegalaxy.org/cloudmap. These serve essentially as a simplified, graphic version of this article that application-oriented end users, who are not interested in the bioinformatic details of the strategy, can use to get started (also see *Proof-of-principle application of CloudMap* below).

Strategies and workflows

When running the CloudMap pipeline using the workflows we provide, several different bioinformatic processing steps are performed automatically using a standardized set of tools. Figure 3 illustrates all of these steps and additionally indicates branch decision points where more experienced users can choose among different software applications to perform desired operations (detailed step-by-step instructions are available in provided workflows, user guides, and videos). For instance, users may choose from several aligners including the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2010), Bowtie (Langmead *et al.* 2009), or other aligners. Users may also choose from several alignment viewers such as the UCSC Genome Browser (Kent *et al.* 2002), the Integrated Genome Browser (IGB) (Nicol *et al.* 2009), WormBase (Harris *et al.* 2004), or other alignment viewers depending on their preferences. Additionally, we have written several tools, which are incorporated into the automated workflows, designed to extract mapping positions for model organisms where a mapping cross has been performed, to perform *in silico* complementation test analysis on mapped or unmapped strains, and to further annotate candidate variants. The modular nature of CloudMap allows the latest bioinformatics tools to be seamlessly incorporated into the data analysis pipeline as they become available. Likewise, the latest data stores, including releases of genome reference files for most common model organisms, are updated regularly within Galaxy. Alternatively, they can be uploaded by users and immediately used.

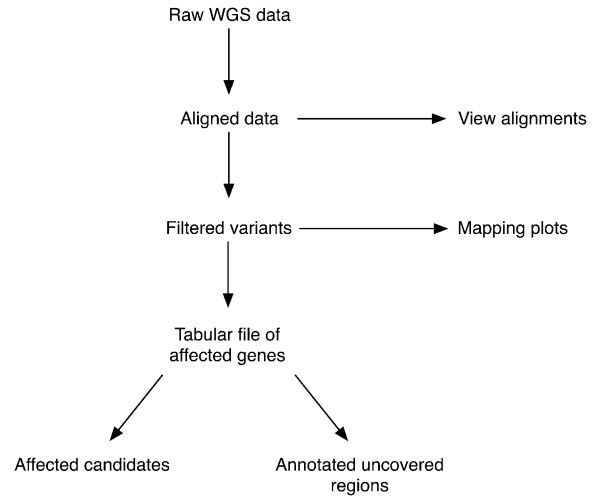


Figure 1 CloudMap overall conceptual strategy for mutant genome analysis. This high-level summary depicts the main CloudMap processes and outputs. Detailed overview of all the CloudMap functions is provided in Figure 3, in the user guides, and published workflows available at http://usegalaxy.org/cloudmap.

All workflows begin with users generating an account and uploading sequencing data to the Galaxy website (http://usegalaxy.org) or loading such data in their local Galaxy installation (see below). Galaxy is also tightly integrated with many popular databases such as WormBase, modENCODE, and Biomart, and transferring data from these websites into Galaxy is a straightforward process. Galaxy accepts FASTQ format data files from all of the major next-generation sequencing platforms (*e.g.*, Illumina, ABI, 454) and files may be compressed for quicker upload times. Current user quotas on the Galaxy main site (250 Gb) allow for analysis of data from an entire Illumina flow cell to be performed in ~1 day with relatively minimal user

Table 1 Summary of useful Galaxy links

Function	Link
Galaxy Wiki	http://wiki.g2.bx.psu.edu/
Learn Galaxy	http://wiki.g2.bx.psu.edu/Learn
List of hosted Galaxy servers	http://wiki.g2.bx.psu.edu/Public%20Galaxy%20Servers
Instructions to configure Galaxy locally	http://wiki.g2.bx.psu.edu/Admin/Get%20Galaxy
Other options for running Galaxy	http://wiki.g2.bx.psu.edu/Big%20Picture/Choices
Instructions on running Galaxy using Amazon Elastic Compute Cloud (EC2)	http://wiki.g2.bx.psu.edu/CloudMan
List of common Galaxy tool errors	http://wiki.g2.bx.psu.edu/Support#Error_from_tools
Toolshed	http://toolshed.g2.bx.psu.edu

This table will continually be updated at http://usegalaxy.org/cloudmap.

involvement. Users requiring more data storage or more powerful computing needs are encouraged to configure their own local install of Galaxy or to configure an instance of Galaxy using Amazon Elastic Compute Cloud (EC2), which offers computing services on a payment-per-usage basis (for details about these configurations see Table 1).

Below we describe in more detail the individual steps and specific operations that can be performed as part of the CloudMap pipeline. Importantly, all of these steps are automated when using the workflows we provide. Our general assumption in the description below is that users will employ CloudMap to identify mutations in a mutagenized model system strain, such as *C. elegans*.

Alignment, variant calling, and annotation

Following upload of WGS sequencing data into Galaxy, several preprocessing steps may be run on FASTQ files if necessary. First, if the same sample was run on different sequencing lanes, users may concatenate FASTQ files using the Concatenate Datasets tool in Galaxy (Figure 3). Next, if FASTQ quality scores are not in the recommended Sanger format, users may convert the quality encoding in their FASTQ files using the Galaxy FASTQ Groomer tool (Blankenberg *et al.* 2010). Once the quality score encoding of the FASTQ file is resolved, the Galaxy FASTQ Summary Statistics and Boxplot tools can be used to perform quality control checks on the raw data (Blankenberg *et al.* 2010).

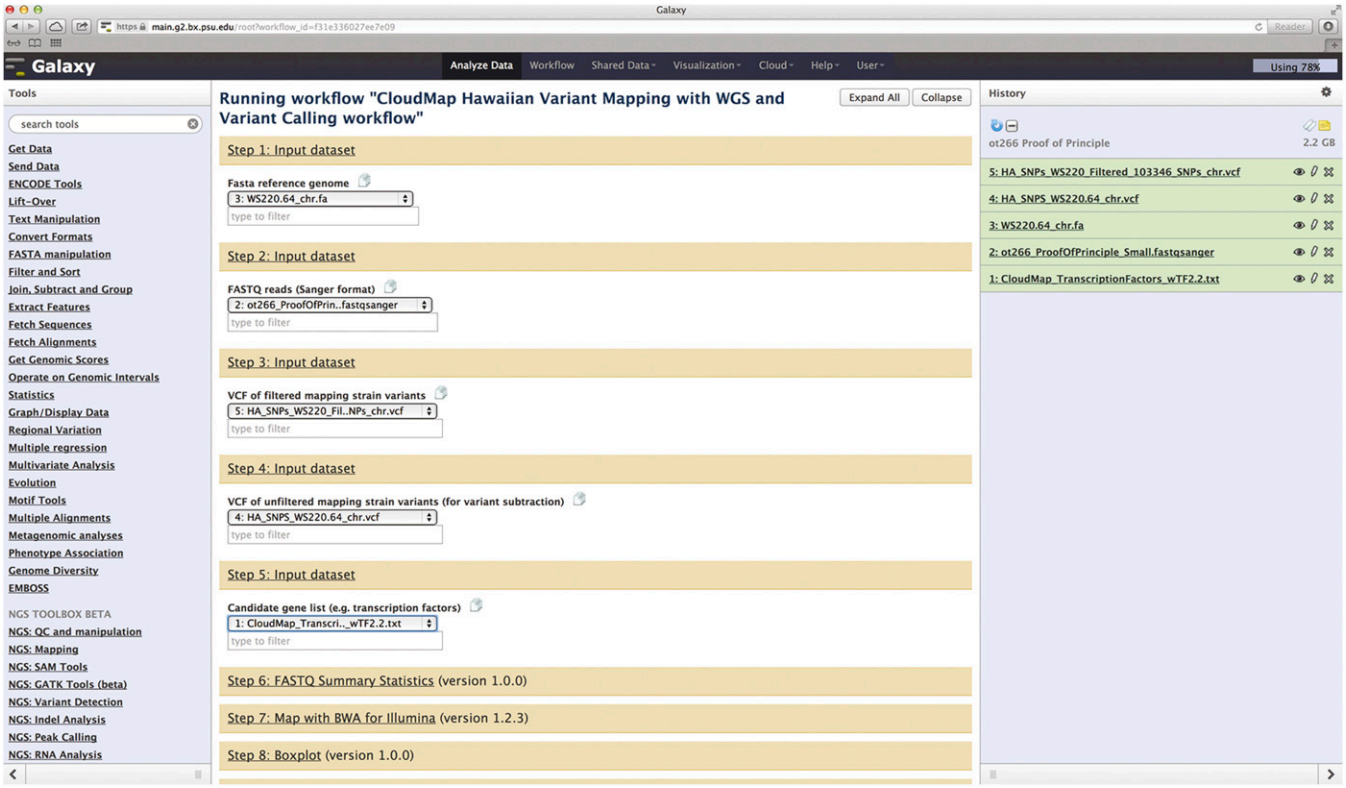


Figure 2 Screenshot of Galaxy workflow using the *ot266* example discussed in *Proof-of-principle application of CloudMap*. Users may run this workflow as well as others at http://usegalaxy.org/cloudmap. The output of the *ot266* workflow is also available as a shared history at the URL mentioned above. Here we see a Galaxy history with the FASTQ raw data file for *ot266* along with various reference files used as input into the CloudMap Hawaiian Variant Mapping With WGS Data and Variant Calling workflow. The reader is referred to user guides and videos for step-by-step instructions.

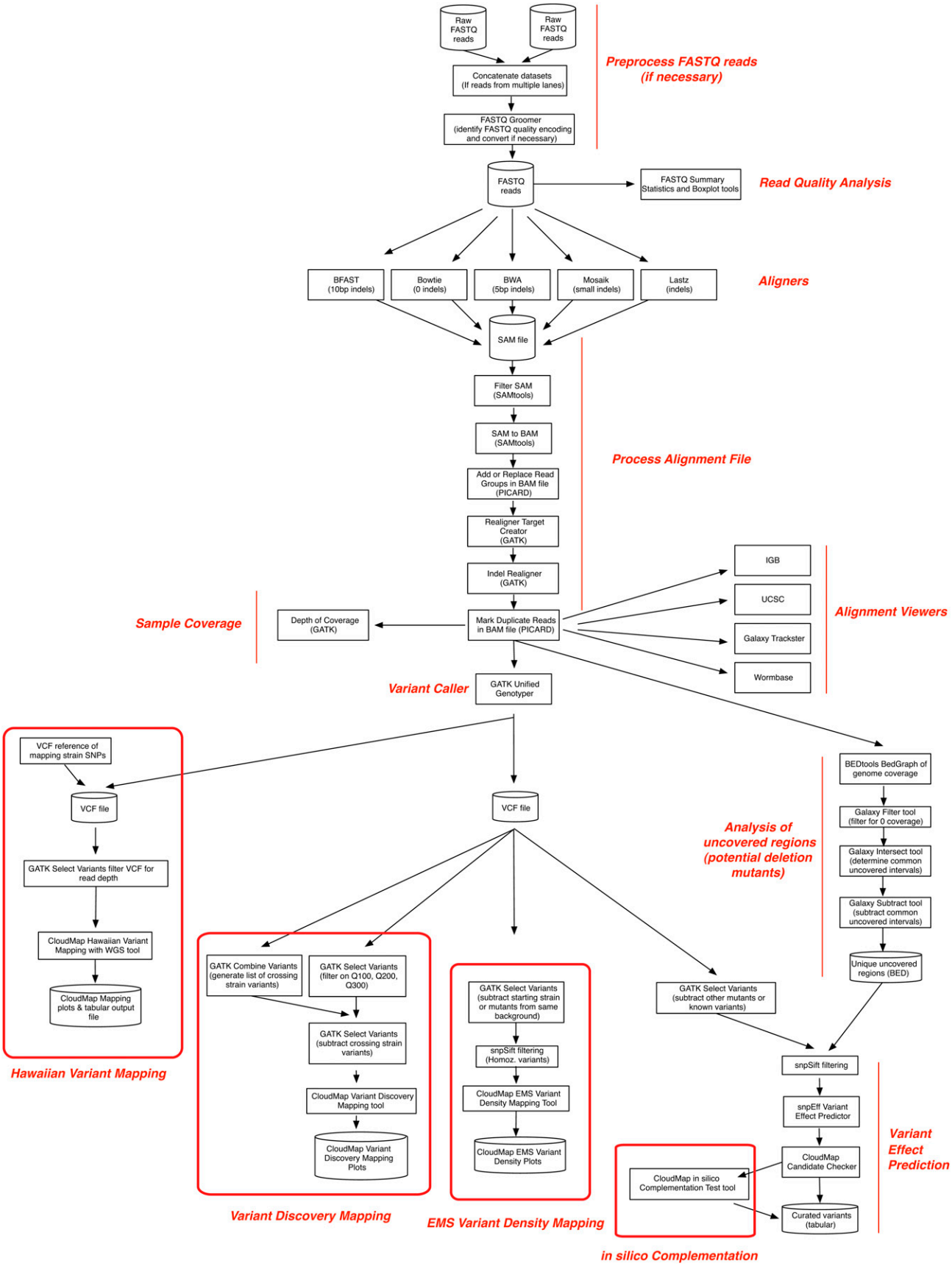
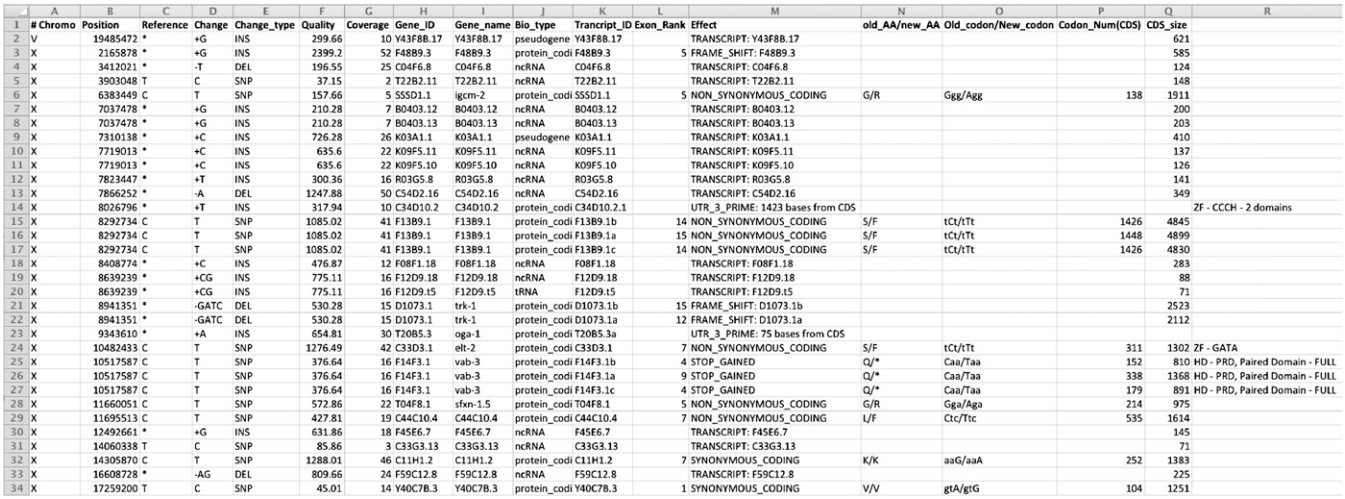


Figure 3 Summary flowchart illustrating all functions used in the CloudMap pipeline. More experienced users may choose among different software tools to perform desired operations at marked decision points in the flowchart. Detailed step-by-step instructions are available in user guides and videos.



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

Figure 4 Sample screenshot of snpEff output following markup of affected transcription factors by CloudMap Check snpEff Candidates tool. Tabular output of mutated genes and transcripts from snpEff together with lists of candidate loci can be used as input into the CloudMap Check snpEff Candidates tool. In the example shown here, the output of the analysis of *ot266* is displayed with the causal lesion in the *vab-3* gene labeled as a homeodomain transcription factor. The "Quality" column reflects the GATK-assigned, PHRED-based QUAL score from the VCF file input into snpEff (Danecek *et al.* 2011).

Users can then choose from the latest open source aligners to align their samples to their reference genome (Figure 3). Readers are referred to a useful survey of sequencing alignment algorithms to learn more about their respective strengths (Li and Homer 2010). Our automated workflow uses BWA (see *Proof-of-principle application of CloudMap*). Following alignment, we use several tools to process the alignment file to improve insertion/deletion (indel) calling and remove duplicate reads (reads with identical start and end points) (see *Proof-of-principle application of CloudMap*).

Several alignment viewers are available for viewing alignments, together with any type of additional track-based data such as gene structure and conservation (via the UCSC Genome Browser), protein domains, available alleles, and modENCODE transcriptome data (via WormBase for *C. elegans* datasets). More experienced users may also upload custom tracks of their own into any supported alignment viewer. In the cases where alignment viewers are integrated into Galaxy (UCSC Genome Browser, IGB, WormBase, Galaxy Trackster, etc.) the alignment file [binary alignment/map (BAM) file] resides on the Galaxy server and is dynamically streamed to the location hosting the alignment viewer.

While alignment viewers are useful for assessing read coverage and sequencing quality, a list of variants is necessary to determine how sequence variants affect genomic function. Variant callers from GATK (DePristo *et al.* 2011) and SAMtools (Li *et al.* 2009) calculate comprehensive lists of all the genomic variants (SNPs and indels of varying sizes depending on the aligner used) with respect to a reference genome in a given sample and output these in the VCF (Danecek *et al.* 2011). Users have the choice to use either variant caller. Readers are referred to a review of SNP calling from next-generation sequencing data that also discusses respective features of GATK and SAMtools (Nielsen *et al.* 2011). Our automated workflow uses GATK (see *Proof-of-principle application of CloudMap*). Then, the variant effect caller snpEff (Cingolani *et al.* 2012) predicts and annotates the effects of these variants on genes (amino acid changes, splice site variants, upstream/downstream potential regions, etc.) and provides PHRED-based quality scores (Ewing and Green 1998) and coverage statistics for each variant. A sample output is shown in Figure 4. Users can sort these tabular files for specific types of variants, such as those that fall within a mapping interval (see below) and are likely to have high impact (for example, premature stops, frameshift mutations, etc.). The user can then easily navigate in an alignment viewer, to the region that contains the mutation of interest for detailed inspection of the reads that were used to call the variant.

Variant subtraction and filtration

The process of variant subtraction is motivated by the desire to remove all but the phenotype-causing variant(s) in a given sample. Therefore, prior to performing analysis on the variants in a given sample, variants that are present in the premutagenesis starting strain ("background variants"), and thus not responsible for the mutant phenotype, should be subtracted. As previously shown in various WGS studies (*e.g.*, Flowers *et al.* 2010), such variant subtraction greatly reduces the number of variants to consider. This can be achieved in several ways, as illustrated in Figure 5. If the premutagenesis, starting strain has been sequenced (our preferred approach, given the low cost of WGS), users may use the GATK Select Variants tool to subtract common variants in this starting strain from the variants in their mutant sample (Figure 5A).

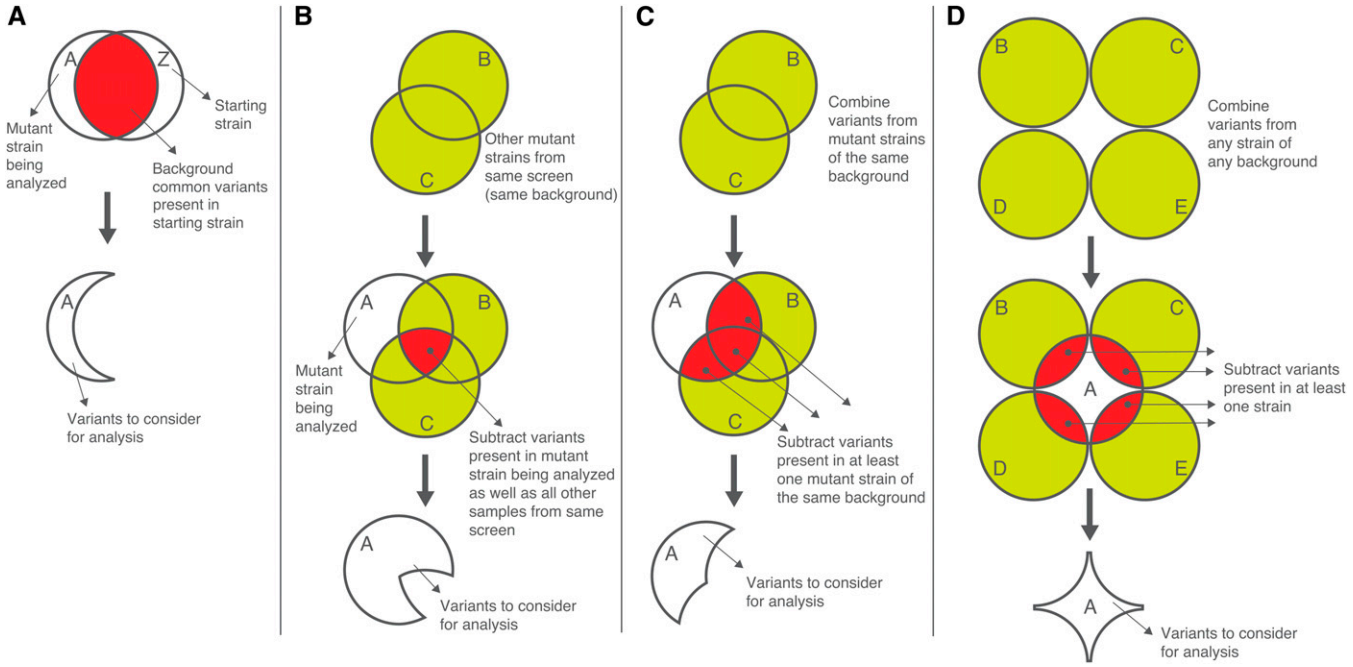


Figure 5 Variant subtraction and filtration. Only a subset of variants in a sample are legitimate candidates that might be responsible for the mutant phenotype of interest. In addition to the ability to map potential mutant lesions to a small region (~1 Mb), the CloudMap pipeline allows users to subtract nonphenotype-inducing variants from consideration. (A) Subtracting variants present in the background strain. If the premutagenesis, starting strain has been sequenced, users may use the GATK Select Variants tool to subtract starting strain variants ("background variants") from consideration. (B) Subtracting variants present in other mutant strains from the same screen. If the premutagenesis strain has not been sequenced, then fewer variants can be subtracted from the mutant under consideration. If other mutant strains from the same screen have been sequenced, common variants present in the premutagenesis strain can be deduced from sequence analysis of such mutants. Employing a fairly conservative approach, we can choose to subtract variants only if they are present in only two or more mutants that have been derived from forward genetic screens on the same starting strain. (C) Subtracting variants present in at least one mutant strain of the same background. A less conservative variant subtraction strategy than mentioned in B involves subtracting all variants that are present in the mutant strain of interest and at least one additional strain from the same screen. (D) Subtracting variants present in at least one strain of any background. A more liberal variant subtraction strategy can be performed by subtracting variants present in at least one strain of any background. The same caveats for this strategy apply as for the strategy described above in C. As variant information from more whole genome sequenced strains becomes available, more variants will be available for this subtraction strategy.

Alternatively, if the premutagenesis strain has not been sequenced, common variants present in the premutagenesis strain can be deduced from sequence analysis of two or more mutants that have been derived from forward genetic screens on the same starting strain. The GATK Combine Variants and Select Variants tools can be used to query mutant WGS datasets for variants present in the mutant strain of interest as well as in all other samples from the same screen (Figure 5B, *i.e.*, the logical intersection of all sets of variants, $A \cap B \cap C$).

Researchers may employ a less conservative variant subtraction strategy by subtracting all variants that are present in the mutant strain of interest and at least one additional strain from the same screen (Figure 5C). As this approach involves subtracting more than only background mutations in the starting strain, it carries with it the increased risk that potentially important phenotype-causing or modifier variants may be eliminated from consideration (*e.g.*, if the identical phenotype-inducing variant is present in one of the subtraction strains, something that is unlikely within a small set of mutant strains but which should be taken into account as the number of datasets grows).

The most liberal variant subtraction strategy involves subtracting from the mutant strain of interest all variants present in additional strains of any background (from any screen), even if present in only one of these additional strains (Figure 5D). The same caveats for this strategy apply as for the strategy described above in Figure 5C; namely, the potential exists for phenotype-causing mutations to be unintentionally subtracted from the mutant strain of interest.

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 from outcrossed samples using either the Hawaiian Variant Mapping or the Variant Discovery Mapping approaches (see *Hawaiian Variant Mapping with WGS Data tool* and *Variant Discovery Mapping tool implements a novel mapping method*), since background variants will be present in a heterozygous state in these pooled samples as a consequence of the mapping cross.

Because it is easy to perform variant subtraction and downstream analysis as well as to document settings for each analysis performed (via Galaxy histories), we recommend starting with the most liberal variant subtraction strategy to quickly see whether prime candidate variants are present in the data. In parallel, users can run more conservative subtraction strategies and analyze those results, as they deem appropriate.

Our lab maintains a list of background variants found in various different screens and routinely compares them to new WGS datasets. We submit premature stops, frame-shift and splice variants to WormBase, and strains carrying these variants to the Caenorhabditis Genetics Center (CGC). We have set our thresholds for variant quality analysis rather high to err on the side of caution (see *Materials and Methods*).

CloudMap Candidate List Checker and useful gene lists

Depending on the type of mutant sequenced or the genetic screen performed, users may be especially interested in variants that affect certain classes of genes. For instance, forward genetic screens for *C. elegans* mutants in which cellular fates are not appropriately executed often yield mutations in transcription factors (*e.g.*, Doitsidou *et al.* 2008). To quickly check whether classes of genes (such as transcription factors) have been affected in a mutant, CloudMap provides the Candidate List Checker tool for snpEff output. This tool accepts a two column, tab-delimited list of gene names and their respective annotation information together with the tabular output from snpEff.

We have generated five such lists that can be used with the Candidate List Checker; the lists are available at http://usegalaxy.org/cloudmap. These lists consist of: (1) all predicted transcription factors (Reece-Hoyes *et al.* 2011); (2) a list of predicted chromatin factors (Tursun *et al.* 2011); (3) a list of all *C. elegans* genes with human orthologs (Shaye and Greenwald 2011); (4) a list of genes associated with neuronal function (Hobert 2012); and (5) a list of genes commonly involved in transgene silencing; such mutants are often retrieved from screens in which transgenes are used (Kim *et al.* 2005; Wang *et al.* 2005; Vastenhouw *et al.* 2006). Notably, some transgene silencer mutations may affect one transgene but not another (even within the same cell), thus falsely encouraging researchers to pursue what they may think is a variant that affects cell fate.

We encourage users to share similar lists by uploading them to shared Galaxy libraries (http://wiki.g2.bx.psu.edu/Admin/Data%20Libraries/Libraries).

In silico complementation testing

If performed on a large scale, forward genetic screens usually yield multiple alleles of individual loci, which define specific complementation groups. The traditional way to identify such complementation groups is via complementation tests performed by genetic crosses. If screens have revealed dozens of mutants, comprehensive complementation testing can be time consuming and labor intensive. Moreover, complementation tests are impossible to perform with dominant alleles and are sometimes subject to misleading results (such as allelic complementation or non-allelic noncomplementation). With the decreasing costs of WGS, it is now possible to simply sequence many mutants that result from a screen and determine *in silico* which mutants carry variants in the same locus (user-defined upstream/downstream boundaries of a gene can also be considered as part of a locus by modifying snpEff output—our pipline sets the upstream/downstream gene boundary to 0 bp as default). To allow such analysis, we developed the CloudMap In Silico Complementation Test tool to compare tabular snpEff outputs [which have been filtered for quality and had common variants subtracted for shared gene hits (alleles) — see *Materials and Methods*] for shared gene hits (alleles). This tool creates two output files: (1) a summary file of the number of shared gene hits among the sequenced mutants sorted from most to fewest (an abbreviated example is shown as Supporting Information, Figure S1A) and (2) a corresponding file of the snpEff annotated alleles from each sample also sorted from most to fewest (Figure S1B). Below we describe the general steps involved in using the tool (see user guide for detailed examples and the CloudMap in Silico Complementation Test workflow: http://usegalaxy.org/cloudmap).

As a first pass at analysis, to remove background variants present in the premutagenized strain, we recommend a liberal subtraction strategy be applied where the most possible variants are subtracted from each strain prior to *in silico* complementation analysis (see *Materials and Methods* for liberal subtraction strategy tool settings). The GATK Combine Variants tool should first be used to create a single VCF file that contains only variants present in at least two samples from the same genetic background (Figure 5C). This VCF file of liberally defined common variants will next be subtracted from the VCF for each individual sample using the GATK Select Variants tool. snpEff will then annotate each of these subtracted VCF files and it is these annotated, tabular snpEff output files that will be used as input to the In Silico Complementation Test tool. When this liberal subtraction strategy is used, the tool will only return results where allelic genetic loci contain nonidentical hits in more than one sample.

It is possible that two independent alleles of the same locus carry the exact same genetic variant. In this case, if a liberal subtraction strategy were applied, the causal variant would be subtracted at a step before running the In Silico Complementation Test tool. Therefore, if the phenotype-inducing mutation is not identified with the liberal subtraction approach, we recommend that a more conservative background variant subtraction be employed to not exclude identical mutagenesis-induced variants from different datasets (see *Materials and Methods* for conservative subtraction strategy tool settings). Users should be aware that by erring on the side of caution with the conservative

subtraction approach, many nonphenotype-causing background variants will remain in the output files and will appear as "false positive" identical variants present in multiple samples. Most of these will simply be background variants present in the starting strain premutagenesis. To deal with this problem in the context of the conservative subtraction strategy, users may lower the Combine Variants parameter to subtract variants shared by <100% of strains, but more than two strains (see *Materials and Methods*).

Candidate allelic mutants identified by such *in silico* complementation tests can then be confirmed by classic genetic complementation and therefore can be reduced to a few simple crosses of candidate alleles (rather than an entire mutant collection).

Identifying deletions in WGS datasets

Various mutagens, including EMS, induce small (a few base pairs) to large (many kilobases) deletions at appreciable frequency (Janssen *et al.* 2010). Current aligners have difficulty in identifying deletions using short read length sequencing approaches such as those favored by Illumina. While BWA is capable of identifying indels in short reads, the size of such indels is limited to 5 bp according to default settings. We have included in the automated workflow a realignment-around-indels step (using GATK), which improves the reliability of indel calling. However, it is still often difficult to distinguish regions of the genome that have no sequencing coverage from genomic regions that are genuinely deleted in the sample (Sarin *et al.* 2010).

CloudMap contains a workflow that helps identify deletions based on lack of coverage. Using the BedGraph tool that is part of the BEDtools package (Quinlan and Hall 2010), users can obtain a file of genome coverage at every position in the genome. This file can be filtered for positions with zero coverage using the Galaxy Filter tool, and the data in the resulting Browser Extensible Data (BED) file can then be subtracted (using the Galaxy Intersect and Galaxy Subtract tools) from other uncovered regions in samples from the same screen in much the same way that normal variant subtraction is performed. We provide a predefined workflow that automates these steps and a user guide that explains the process in detail. The uncovered regions that are unique to a strain can then be used as input into snpEff and a resulting tabular annotated file can be sorted for those unique uncovered regions that fall in the mapping region and affect protein coding genes.

While the list of annotated uncovered regions is a useful guide, in terms of data confidence, it does not compare to having high read depth for a variant and high quality scores for the bases at the variant position. In other words, uncovered regions are the absence of evidence, not evidence that a genomic region is truly missing. To ascertain with greater confidence which uncovered regions are real, researchers should view these regions in an alignment viewer. We find that true deletions tend to have a "cliff-like" pattern of high coverage followed by a steep drop in coverage down to zero (Figure S2). Furthermore, the regions of high coverage flanking the putative deletion often have SNPs or insertions present in many of the reads—reflecting the fact that regions of the genome that were previously physically separated are now adjacent (short NGS reads do not span large deletions as BWA by default only recognizes indels of up to 5 bp). We confirmed that the putative deletion in Figure S2 was real by amplifying this genomic region by PCR and Sanger sequencing of the amplicon.

CloudMap Hawaiian Variant Mapping With WGS Data tool

As mutagenized strains contain thousands of variants (SNPs, indels, etc.), a list of variants and their effects alone is usually not adequate for identifying the phenotype-causing mutation. We have previously described a one-step strategy for whole genome sequencing and mapping (Doitsidou *et al.* 2010) modeled on a similar strategy in plants (Schneeberger *et al.* 2009) that allows for fine mapping of phenotype-causing mutations in a time- and cost-efficient manner. Our strategy requires crossing a mutant parental strain from a genetic screen (derived from the N2 Bristol strain) to the polymorphic mapping strain, Hawaiian CB4856, and picking a number of F_2 mutant progeny (Figure 6A). The progeny of these F_2's (F_3's and F_4's) are then pooled and sequenced. Genomic regions that are linked to the causal mutation (and thus selected for) will recombine less frequently during meiosis and thus sequencing will reveal a stretch of pure parental sequence in the region of the mutation. In contrast, nonlinked genomic regions will recombine in a roughly 50/50 ratio of mapping *vs.* parental sequences (Figure 6A). We can discriminate between stretches of parental *vs.* mapping strain sequence in pooled F_2 progeny by examining the presence/absence of mapping strain SNPs relative to the parental Bristol strain, that are distributed at ~1 every 1000 bp (Hillier *et al.* 2008). The ratio of Hawaiian reads/total reads at each of the ~112,000 Hawaiian SNP positions is then plotted according to physical position in the genome. Genomic regions devoid of Hawaiian SNPs represent loci that are linked to the causal mutation (Doitsidou *et al.* 2010).

CloudMap's Hawaiian Variant Mapping With WGS Data tool offers an improved plotting method for representing variant mapping data derived from crosses such as described above. [Note: while the Hawaiian Variant Mapping tool accepts variants (SNPs and indels) as input, we use only Hawaiian SNPs for mapping here because this list of SNPs was already provided in WormBase. Researchers who wish to use the mapping tools with known indel positions as well as with SNPs—for instance if they have sequenced their crossing strain—may do so with no modifications to the tool.] Prior to running the plotting tool, we feed an alignment (BAM) file and a VCF reference file of provided Hawaiian SNP positions compiled from WormBase (http://usegalaxy.org/cloudmap) as input into the GATK Unified Genotyper tool. The output of this tool is a VCF file containing

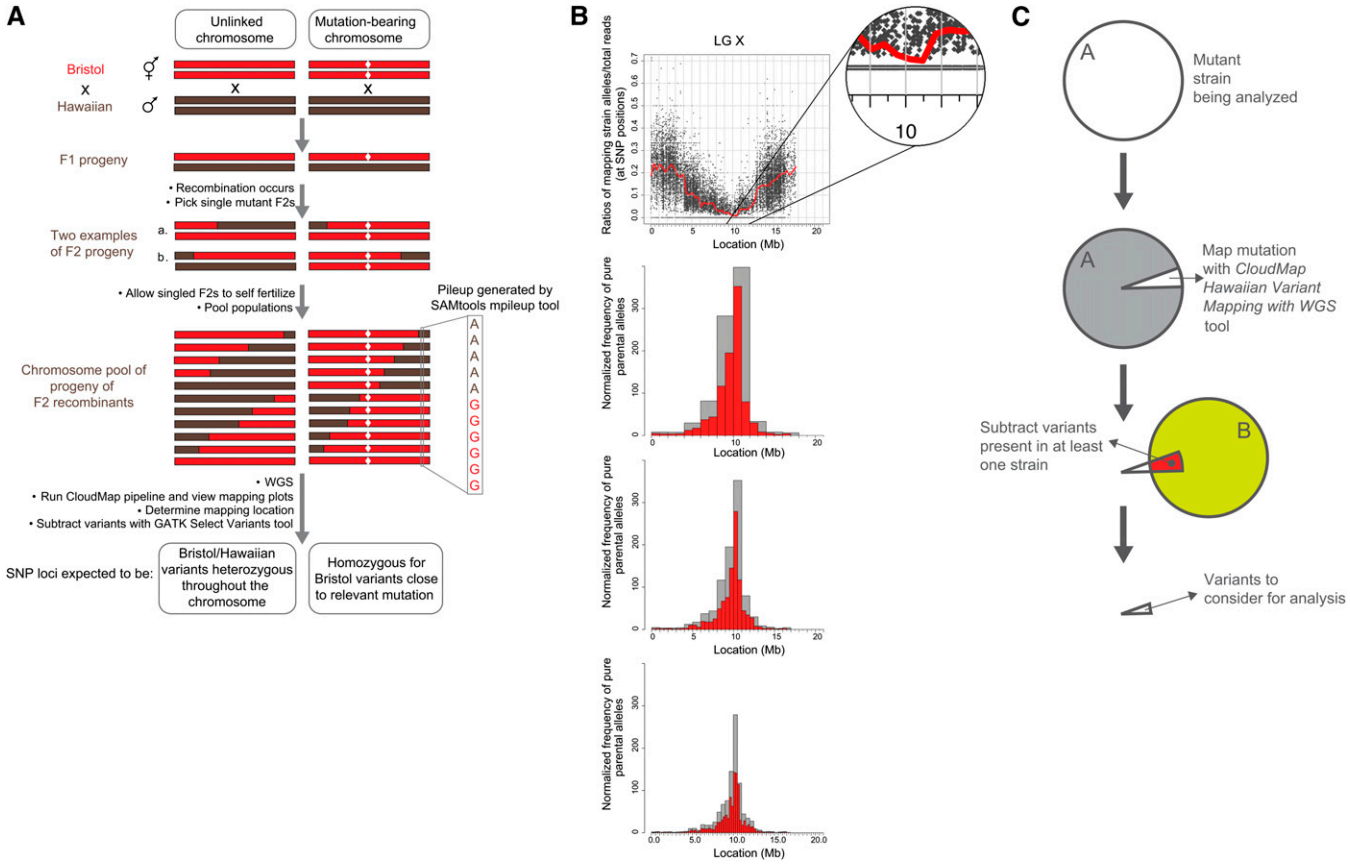


Figure 6 CloudMap Hawaiian Variant Mapping with WGS Data strategy. (A) Schematic presentation of a previously described one-step strategy for whole genome sequencing and mapping (Doitsidou *et al.* 2010) modeled on a similar strategy in plants (Schneeberger *et al.* 2009). (B) The CloudMap Hawaiian Variant Mapping With WGS tool plots the ratio of mapping strain alleles/total reads at each of the mapping strain SNP positions in the genome, as exemplified with the *ot266* dataset. To better visualize trends in the scatter plots of the SNP ratios, we plot a LOESS regression line (red) through all the points on each chromosome. Each scatter plot also has a corresponding frequency plot that displays regions of linked chromosomes where pure parental allele SNP positions are concentrated. The same genomic region that shows linkage in the LOESS scatter plots also shows a matching peak in the frequency plots of pure parental alleles. These frequency plots are binned by default into 1-Mb (gray) and 0.5-Mb (red) bins although these bin sizes are adjustable. The figure also shows 2-Mb (gray) and 1-Mb (red) bin sizes (top frequency plot) and 0.5-Mb and 0.25-Mb bin sizes (bottom frequency plot). Data in these plots can also be normalized to improve the mapping signal (details in text, Figure 7, and Table S1). (C) CloudMap Hawaiian Variant Mapping with WGS Data variant subtraction. As described in the text and in Figure 5, subtracting variants present in other samples can reduce the number of variants that are considered candidates for causing the phenotype of interest.

reference (Bristol) and alternate (Hawaiian) allele calls at each of the ~100,000 Hawaiian SNP positions in the pooled sample. The reader is referred to the user guide and online video for direction on this procedure. The Hawaiian Variant Mapping with WGS Data plotting tool accepts as input this single VCF file (the list of base calls at all Hawaiian SNP positions in the pooled sample). As output, the tool plots the ratio of Hawaiian strain reads/total read depth at each SNP position of the mapping strain. Chromosomes that contain regions of linkage to the causal mutation(s) display regions enriched for SNP positions that have a Hawaiian reads/total reads ratio equal to 0. The scatter plots of these ratios plotted against chromosome position for such linked regions will have a cluster of data points lying exactly on the *x*-axis.

The Hawaiian Variant Mapping With WGS Data tool then plots a local regression (LOESS) regression line through all the points on a given chromosome thus giving far greater accuracy to the mapping region (Figure 6B). LOESS regression is a locally weighted polynomial regression that gives weight to points near the position being estimated as well as to points further away on the chromosome (Cleveland *et al.* 1992). We use the LOESS smooth function in R (available as dloess at http://www.netlib.org/a/) and provide an adjustable span parameter that allows users to modify the weight given to points other than the point being estimated. Based on our testing, we have settled on 0.1 as a LOESS span default for *C. elegans*. Larger span values result in smoothing of the regression line to reflect trends at a more macro level while smaller values result in regression lines that more closely reflect local data fluctuations. While adjusting the LOESS span is a useful method for achieving a tighter mapping region, the largest gains in mapping accuracy come when greater numbers of F_2 progeny are pooled and sequenced.

For each scatter plot, the CloudMap Hawaiian Variant Mapping With WGS Data tool also plots a corresponding frequency plot that displays regions of linked chromosomes where pure parental allele variant positions are concentrated (Figure 6B). For a given linked chromosome scatter plot, the same genomic region that shows a dip in the LOESS line should show a matching peak in the pure parental allele frequency plot. Parental allele variant position frequencies are calculated on a binned basis with bins of 1 Mb and 0.5 Mb by default. Users wishing to decrease their bin sizes to achieve finer mapping (in the case where many F_2 progeny were sequenced) or increase their bin sizes to accommodate data from fewer pooled F_2 progeny can easily do so (details in user guides and videos). We find that the causal variant typically resides in the largest 1-Mb frequency bin (unpublished data; 5/5 cases where single gene mutants have been mapped using CloudMap and rescued by expressing wild-type (WT) copies of the gene, ≥50 recombinants used).

In cases where two or more mutations reside on a single chromosome, or a mutation and a transgene on the same chromosome are selected during picking of F_2's, users should expect to see a LOESS line that approximates a zero ratio for a long physical distance and a corresponding broad peak for the frequency plots (G. Minevich, R. J. Poole, and O. Hobert, unpublished data). Such plots indicate that two or more linked genetic elements have been selected for, with little to no recombination in between.

Both the scatter plots and frequency plots of pure parental alleles can be customized via the CloudMap Hawaiian Variant Mapping With WGS tool to output publication-ready mapping figures. *Y*-axis limits, and the colors of the dots and LOESS line, can be adjusted. Users also have the option to standardize the *x*-axis of all plots to the size of the largest chromosome so that plots for each chromosome can be easily compared. Accurate mapping of this nature allows users to further filter previously subtracted variant lists for a specific region, dramatically reducing the number of possible causal variants (Figure 6C; see Figure 8 further below for an example).

Hawaiian SNP normalization

Analysis of the CloudMap SNP Mapping with WGS Data tool output from several mutant strains reveals previously described patterns of genetic incompatibility between the Bristol and Hawaiian strains (Seidel *et al.* 2008). For instance Figure 7 shows that at the ~2-Mb physical locations on chromosomes I and II, there are sharp dips in the LOESS regression line with corresponding peaks in the frequency plots of pure parental alleles. These regions, enriched for parental alleles, are consistent in all mapping plots examined thus far (>10 strains, G. Minevich, R. J. Poole, and O. Hobert, unpublished data) and in terms of locating the correct mapping region, contribute to noise in both the scatter plots and frequency plots of pure parental alleles. To normalize for these incompatibilities and also to correct for divergence of either our Bristol or Hawaiian strains from their respective published reference strains, we subtracted SNPs where the ratio of Hawaiian alleles/total read depth was <0.05 or >0.95 in at least two mutant strains (Table S1). For the purposes of Hawaiian SNP normalization alone, SNPs that have Hawaiian alleles/total read depth ratios of <0.05 are considered Bristol at that position (allowing for sequencing error or picking of WT animals) and SNPs with Hawaiian alleles/total read depth ratios of >0.95 are considered Hawaiian at that position (allowing for sequencing and picking error as described above). It is important to note that these ratios of <0.05 or >0.95 reflect the presence/absence of Hawaiian mapping strain alleles in the progeny of individually picked F_2's (mostly F_3 and F_4 generations) that are pooled and sequenced. The purpose of filtering out those SNP positions that consistently show ratios of <0.05 or >0.95 across several samples is simply to normalize the plots, not to assess potential incompatibility regions.

We were able to remove from consideration 8715 Hawaiian SNP positions that consistently had a ratio of Hawaiian alleles/total read depth <0.05 or >0.95 in at least two mutant strains. We provide this filtered list of 103,346 Hawaiian SNPs together with the unfiltered list of 112,061 Hawaiian SNPs from the WS220 release of WormBase (http://usegalaxy.org/cloudmap). It is important to note that researchers who sequence their own Hawaiian strain may use their personal derived list of Hawaiian variants instead of the one provided here.

If the CloudMap Hawaiian Variant Mapping With WGS Data tool is run using the filtered list of Hawaiian SNPs, unlinked chromosomes no longer display the same levels of linkage as measured by the frequency plots of pure parental allele positions, while linked chromosomes continue to appear linked (Figure 7A). The effect of subtracting these 8715 Hawaiian SNPs from the plots is to dramatically clean up the linkage signal while dampening linkage resulting from incompatibility regions. In principle, any mapping strain whose SNP positions and variants are known can be used with this tool. We encourage researchers to share their curated lists of mapping variants so other members of the genetics community may use them.

The Hawaiian Variant Mapping With WGS Data tool further normalizes the frequency plots by adjusting for the density of Hawaiian SNPs on each chromosome according to the equation in Figure 7B. We find that this equation sharpens the signal of pure parental allele frequencies on the linked chromosomes (Figure 7A). Users have the option of running the Hawaiian Variant Mapping With WGS Data tool without the normalization option by leaving the "normalize" checkbox unchecked and using the list of 112,061 unfiltered Hawaiian SNPs (the tool is set to normalize by default).

Hawaiian Variant Mapping with WGS Data support for other organisms

The CloudMap Hawaiian Variant Mapping With WGS Data tool supports data from any organism that has been crossed

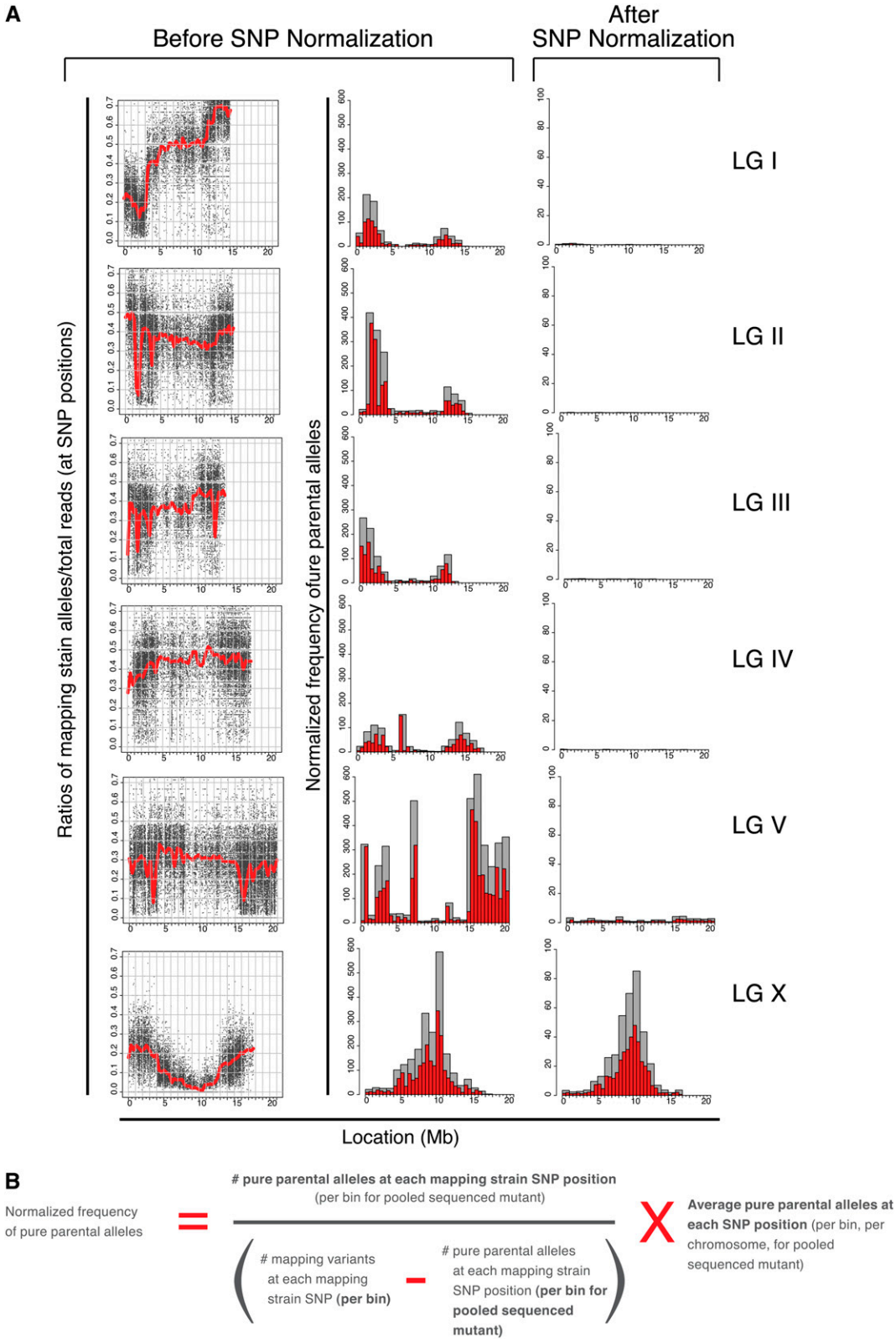


Figure 7 Hawaiian SNP normalization. (A) All mapping plots examined thus far contain similar regions of pure parental alleles. To normalize for these and improve the mapping signal, we removed those Hawaiian SNPs from consideration where the ratio of Hawaiian alleles/total read depth was either <0.05 or >0.95 in at least two mutant strains (Table S1 and details in text). (B) Equation of mapping strain SNP normalization procedure. Users have the option of applying this normalization when using the CloudMap Hawaiian Variant Mapping With WGS Data tool.

to a mapping strain for which variant information is available. *C. elegans* and *Arabidopsis* are natively supported. For all other organisms, users must provide a simple tab-delimited configuration file containing chromosome numbers and respective lengths (example configuration files for most major organisms provided at http://usegalaxy.org/cloudmap). Additional files required for other organisms are the same as described for *C. elegans*: a VCF file consisting of pooled F_2 mutant progeny sequencing data, and a VCF file of the mapping strain variants. To demonstrate support for organisms other than *C. elegans*, we show that CloudMap can be used to map mutant WGS data from *Arabidopsis*.

Our one-step strategy for whole genome sequencing and mapping (Doitsidou *et al.* 2010) was originally modeled on a similar strategy in plants (Schneeberger *et al.* 2009) so we decided to test whether the CloudMap Hawaiian Variant Mapping With WGS Data tool would work on the publically available data from that publication. We find that the LOESS regression line and frequency plots very closely resemble the plots from Schneeberger *et al.* (2009) (Figure 9). Furthermore, our method has the additional advantage of being available on the cloud and not requiring installation or configuration of any software.

Proof-of-principle application of CloudMap

We used the CloudMap Hawaiian Variant Mapping with WGS Data tool together with tools within Galaxy to map and identify the *vab-3(ot266)* allele that we have previously analyzed with another WGS sequence analysis tool developed in our lab, MAQGene (Bigelow *et al.* 2009). This direct comparison with MAQGene using the same reads analyzed in Doitsidou *et al.* (2010) allows us to illustrate some of the advantages inherent in the new pipeline. For this proof of principle, we focused exclusively on the analysis of SNPs and indels in the mapping region but readers are reminded that tools exist in the pipeline for analysis of putative uncovered regions (see *Identifying deletions in WGS datasets*).

vab-3(ot266) is a recessive mutation that we identified in a screen for loss of dopamine neuron specification as evaluated by loss of *dat-1*::*gfp* expression (contained on transgene *vtIs1*) (Doitsidou *et al.* 2008, 2010). *ot266*; *vtIs1* animals were crossed with the Hawaiian CB4856 strain and 50 singled F_2 progeny were picked according to the defective dopamine neuron specification phenotype. Only F_2 progeny that were heterozygous for the *vtIs1* reporter were sequenced to avoid linkage to the *vtIs1* transgene. Progeny of the 50 singled F_2's (F_3's and F_4's) were pooled and sequenced using 100-bp reads on an Illumina GA2 sequencer, as previously described (Doitsidou *et al.* 2010).

Data were processed according to the CloudMap Hawaiian Variant Mapping with WGS Data workflow. We describe here each of the individual steps this automated workflow performs, with links to the relevant sections of the user guide and videos. Raw data files and automated workflows used for this analysis are available at (http://usegalaxy.org/cloudmap). The analysis of *ot266* was performed using a workflow that consists of two main branches: one that leads to a list of annotated variants and another that results in mapping plots. The initial data analysis steps are shared between the two branches: raw FASTQ reads are loaded into Galaxy from the CloudMap *ot266* library and the FASTQ Summary Statistics and Boxplot tools immediately begin to calculate quality metrics on the raw FASTQ reads prior to alignment. While this is happening, those same FASTQ reads are aligned using BWA. The BWA aligner is preferred over MAQ, the aligner used by MAQGene, because it is faster, more accurate, and can also detect indels of up to 5 bp using default settings. Aligned sequence alignment map (SAM) files then had unmapped reads removed and were converted to their binary equivalent BAM files using SAMtools. We then added read groups to the BAM file (sample annotations required for downstream processing) using PICARD (http://picard.sourceforge.net), realigned the BAM file around indels using GATK, and removed duplicate reads using PICARD (this step keeps the best quality read out of a set of identical reads, thus reducing a potential source of noise). This BAM alignment file is used as input into several downstream functions, but it can also be viewed directly using one of the available alignment viewers (Figure 3). Sample coverage is automatically calculated on this BAM file using the GATK Depth of Coverage tool as part of the provided workflow. At this point, the two alternative branches in the analysis diverge but these steps are run in parallel using the provided workflow. Users receive an E-mail when the separate branches finish analysis.

The analysis branch that ultimately results in a tabular list of annotated variants uses one of two analysis paths to arrive at an intermediate step list of variants in the form of a VCF file (Danecek *et al.* 2011). Our preferred variant caller (used in the automated workflow) is the GATK Unified Genotyper, which is used to quickly and directly generate a VCF from a BAM file. Alternatively a combination of SAMtools "mpileup" and BCFtools "view" (Li *et al.* 2009) (which we provide a Galaxy wrapper for) can also be used to generate a VCF. Provided user guides detail how either GATK's Unified Genotyper or SAMtools mpileup followed by BCFtools view can be used to generate VCF files from BAM files.

As described earlier, the ability to easily subtract variants between samples is one of the advantages of CloudMap over the MAQGene pipeline. VCF files contain all variant information provided by the variant caller and VCF files can be merged or subtracted from one another using the GATK Combine Variants and GATK Select Variants tools as part of the CloudMap Subtract Variants workflow (details in the provided user guides and videos; see Supporting Information). In Figure 8, we schematically summarize the variant filtration and subtraction steps we followed to subtract non-causal mutations from the *ot266* variant list.

While the above workflow processed the list of annotated variants, we simultaneously ran the mapping branch of the pipeline (readers are reminded that both branches of the

analysis occur in parallel using automated workflows). Using the BAM file we generated earlier, we selectively ran the GATK Unified Genotyper tool on Hawaiian SNP positions only (using a supplied reference VCF file of 103,346 filtered Hawaiian SNP positions, http://usegalaxy.org/cloudmap). The resulting VCF file of reference (Bristol) and alternate (Hawaiian) alleles at all Hawaiian SNP positions was then used as input into our CloudMap Hawaiian Variant Mapping with WGS Data tool. From the resulting plots (Figures 6B and 7A) we narrowed down the mapping interval for our causal mutation to a 1-Mb interval on linkage group (LG) X.

We next used the *ot266* postsubtraction VCF with 1085 remaining variants that we generated earlier (Figure 8) as input into snpEff to predict the effect of the homozygous variants in the ot266 VCF file and to annotate these variants (Cingolani *et al.* 2012). To check what transcription factors might be affected in the sample, we then ran the CloudMap Candidate List Checker with our provided list of transcription factors (http://usegalaxy.org/cloudmap). When we filtered for the 10- to 11-Mb mapping region on LG X in the tabular output file in Excel, we observed 10 variants in total (Table S2). Of these 10 variants, 2 were protein coding and 1 was a premature stop in the *vab-3* transcription factor previously identified as the causal variant (Doitsidou *et al.* 2010). This variant can then easily be viewed in the UCSC Genome Browser to examine the individual reads covering the mutation and also to view nematode conservation (or any other track-based data) at that position (Figure 10).

EMS Variant Density Mapping tool

Due to the vast number and frequency of Hawaiian SNPs relative to *C. elegans* and the speed with which one can pick individual F_2 mutants and sequence their progeny, we recommend the Hawaiian SNP mapping approach for most mapping applications. However, there remain certain scenarios where alternate mapping approaches are useful. For instance, introducing tens of thousands of Hawaiian variants into a mutant strain may not be desirable for individuals concerned with the possibility that some of these Hawaiian variants may act as modifiers of a given phenotype. Behavioral mutants may be especially vulnerable in this regard. Furthermore, in the case of suppressor screens or other screens that have been performed in a mutant background, it is complicated to recover both the suppressor variant and the starting mutation when picking the F_2 progeny required for the Hawaiian SNP mapping technique. In these scenarios, it is useful to follow the approach detailed in Zuryn *et al.* (2010) that involves plotting frequencies of variant density in a mutant *C. elegans* strain that has been backcrossed to its (premutagenesis) starting strain (Zuryn *et al.* 2010). We therefore developed the CloudMap EMS Variant Density Mapping tool to plot data acquired from the Zuryn *et al.* (2010) backcrossing approach.

Common (*i.e.*, nonphenotype causing) variants present in multiple WGS strains with the same background should first be subtracted from the variants in the sequenced mutant using the GATK Select Variants tool. The EMS Variant Density Mapping tool accepts this list of filtered variants in the form of a VCF file and offers the option of further filtering for the most common EMS-induced mutations *i.e.*, G/C → A/T).

We analyzed the *fp6* mutant from Zuryn *et al.* (2010) using the EMS Variant Density Mapping tool and the results are shown in Figure S3. While Zuryn *et al.* (2010) subtracted common variants in the *fp9* and *fp12* strains from *fp6*, we only used the *fp9* strain for common variant subtraction. Nonetheless, we were able to clearly show linkage to the causal mutation on approximately the same region of LG III as in Zuryn *et al.* (2010).

As we used the more sensitive BWA aligner (which can pick up indels ≤5 using default settings) for our CloudMap approach, we were able to retrieve many more variants than Zuryn *et al.* (2010). In theory, these additional variants should be useful for achieving greater mapping accuracy. Subtracting common (nonphenotype causing) variants from more whole genome sequenced strains (using the GATK Select Variants tool) should result in less noise and a tighter mapping region. Additional backcrosses will, of course, also result in a smaller mapping region.

Variant Discovery Mapping tool implements a novel mapping method

As mentioned above, mapping based on the segregation of linked EMS-induced mutations as described by Zuryn *et al.* (2010) has the important advantage of not relying on a polymorphic mapping strain like the Hawaiian strain. A recent study in plants has extended this conceptual approach, using not only EMS-induced variants to map a phenotype-causing mutation, but also using bulk segregant analysis to increase mapping accuracy (Abe *et al.* 2012). We have developed a similar method, which we call Variant Discovery Mapping. Our method makes use of background variants in addition to EMS-induced variants (including indels as well as SNPs), and it also uses the bulk segregant approach. We have integrated the method into CloudMap and undertook a proof-of-principle analysis, as described below.

The conceptual strategy of Variant Discovery Mapping is to perform *in silico* bulk segregant linkage analysis using variants that are already present in the mutant strain of interest, rather than examining those introduced by a cross to a polymorphic strain. Any individual mutant strain will contain a certain number of homozygous variants compared to the reference genome. These homozygous variants are of two types: (1) those directly induced during mutagenesis (one or more of which are responsible for the mutant phenotype) (Figure 11A, red diamonds) and (2) those already present in the background of the parental strain, either because of genetic drift or because of the parental strain containing, for example, a transgene that was integrated into the genome by irradiation (Figure 11A, pale blue diamonds).

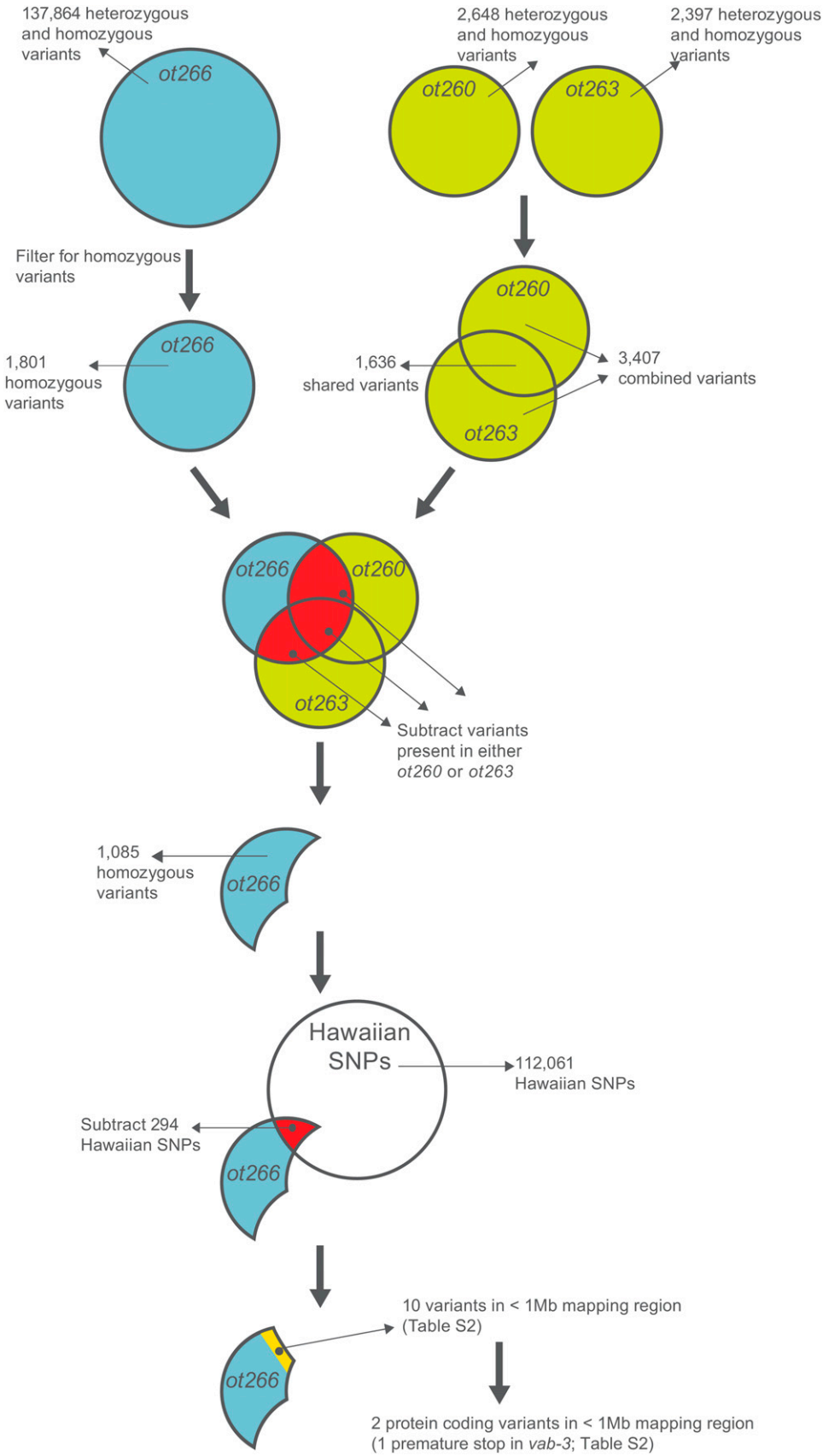


Figure 8 Proof-of-Principle Variant Subtraction strategy. A step-by-step proof-of-principle analysis using the *vab-3 (ot266)* allele. Strains *ot260* and *ot263* are mutants retrieved from the same screen for loss of dopamine neuron specification as *ot266*. *ot266* was crossed to the highly polymorphic Hawaiian strain so it contains many more variants than the *ot260* and *ot263* strains, which were sequenced without outcrossing. Automated workflows for this analysis, raw datasets, and a shared history of the analysis are all available at http://usegalaxy.org/cloudmap.

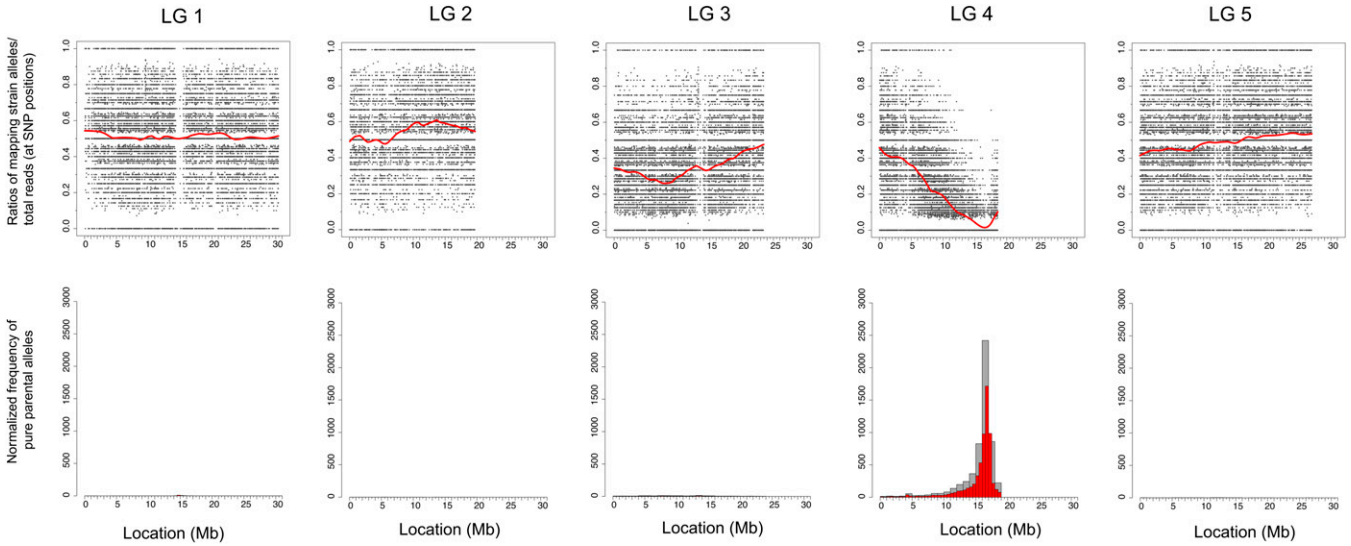


Figure 9 Hawaiian Variant Mapping With WGS Data tool support for other organisms. To demonstrate support for organisms other than *C. elegans*, we show that CloudMap can be used to map mutant WGS data from *Arabidopsis* (Schneeberger *et al.* 2009). Users must provide a simple configuration file for organisms other than *C. elegans* and *Arabidopsis*. Configuration files for most organisms and instructions for other organism support are provided at http://usegalaxy.org/cloudmap.

Following an outcross to a nonparental strain and selection of a pool of F_2 mutant recombinants, these homozygous variants will segregate according to their degree of linkage to the phenotype-inducing locus. Just as with previously described polymorphic mapping (Figure 6A), the degree of linkage will be directly reflected in the allele frequency among the pool of recombinants and this can be represented as scatter plots of the ratio of variant reads/total reads present in the pool of sequenced recombinants (Figure 11A). We then plot a LOESS regression line through all the points on a given chromosome to give greater accuracy to the mapping region (Figure 11B) (as discussed in the CloudMap SNP Mapping With WGS Data tool section). While the concept of the scatter plots is similar to that of the polymorphic mapping approach, the LOESS lines on scatter plots for linked chromosomes now approach 1 instead of 0, indicating retention of the original homozygous variants in the linked region.

We also draw corresponding frequency plots that display regions of linked chromosomes where pure parental allele variant positions are concentrated (positions where the ratio of variant reads/total reads are equal to 1) (Figure 11B). These pure parental allele frequency plots are by default normalized in a manner similar to that for polymorphic mapping, although this normalization option can be turned off (see Figure S4 for the normalization equation). In addition, we provide an option for users to draw frequency plots of positions where the ratio of variant reads/total reads are $\geq$0.9. This option is useful in cases where there are few variants remaining after subtraction, as it effectively increases the number of variants and may therefore provide a stronger mapping signal. Using this 0.9 option should also handle cases where one or a few of the pooled mutants was incorrectly picked (*i.e.*, it was not a mutant) or where sequencing errors have resulted in less than the perfect 1 ratio for pure parental alleles at a given position. We find that using this option in some cases increases the strength of the linkage signal in the frequency plots. However, the scatter plots with LOESS regression remain the more robust and accurate of the two plots and we recommend adhering to them in any cases where the scatter plots and frequency plots disagree.

During the outcrossing to a nonparental strain ("crossing strain"), variants that are present in the crossing strain (relative to the reference genome) will be introduced into the F_2 mutant recombinants (Figure 11, A and B, lime green diamonds). This set of crossing strain variants will cause false linkage or false nonlinkage to the causative mutation and must be identified and removed to assess only those mutant strain variants that were homozygous (relative to the reference genome), before outcrossing. If we consider the two most extreme examples, any variant that is homozygous in the crossing strain and also by chance homozygous in the mutant strain will remain homozygous after outcrossing (see unlinked chromosome in Figure 11A). In this case, the allele frequency, as assessed by the variant read/total read ratio, in the pool of outcrossed F_2 mutant recombinants will be close or equal to 1. This will generate regions of false linkage on ratio plots of unlinked chromosomes (Figure 11B, top left scatter plot). In the second case where crossing strain variants pose a problem, variants that are homozygous in the crossing strain, and positionally close to, but not present in the linked region of the mutant, will be picked up by meiotic recombination at a low frequency. This frequency depends on the physical distance of the crossing-strain variant from the phenotype-inducing mutation

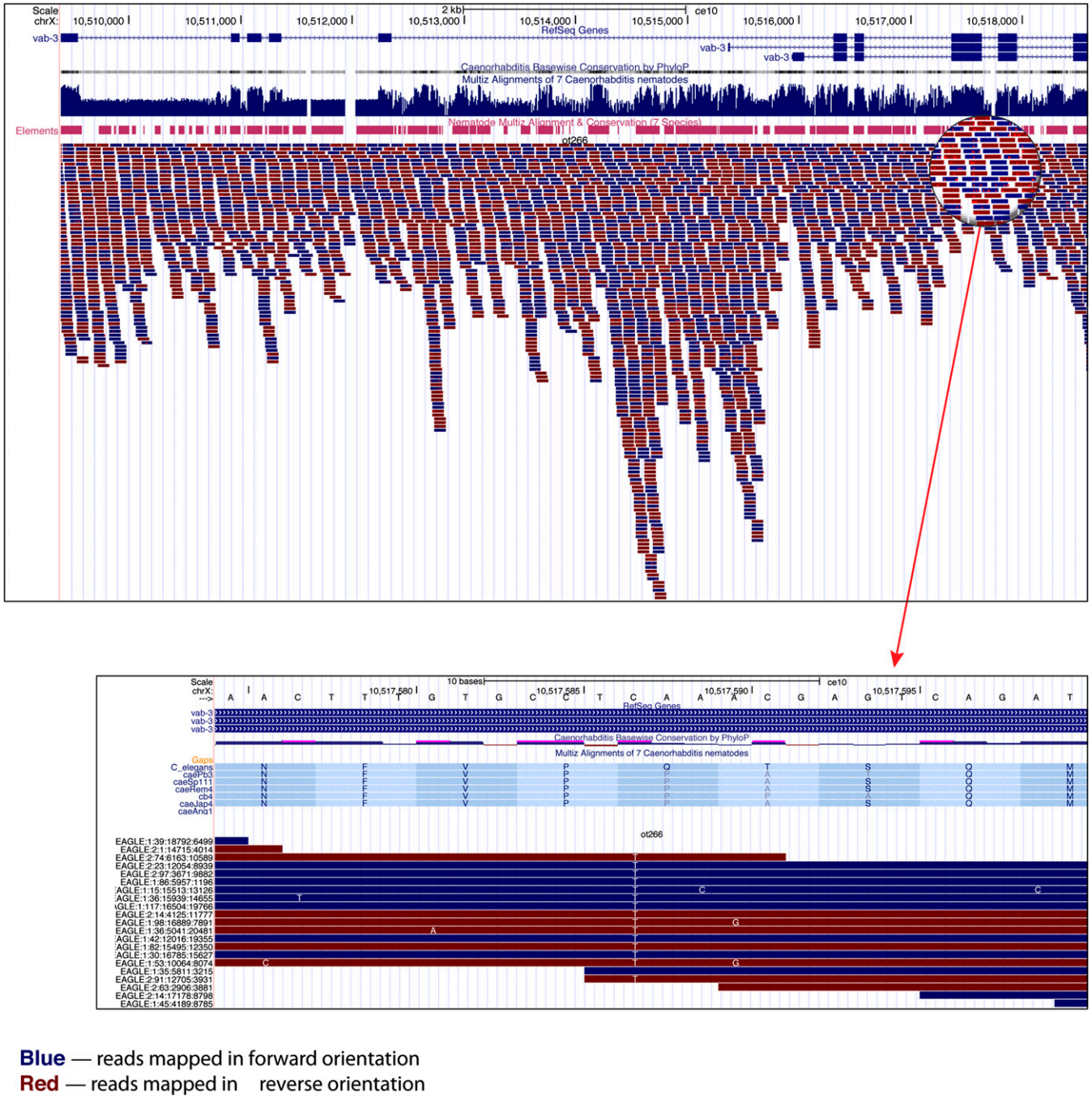


Figure 10 The *vab-3(ot266)* allele as displayed in the UCSC Genome Browser. Users can view their WGS alignments and any other track-based data in their choice of genome browser (UCSC, WormBase, IGB, or Galaxy Trackster). Here we show the *vab-3* locus and a zoom-in view of the C—>T SNP that leads to a premature stop mutation.

(Figure 11A). These variants will therefore have low allele frequencies in the pool of recombinants, despite being close to the phenotype-inducing mutation and will appear as false nonlinkage in the scatter plots (Figure 11B, top middle scatter plot). These crossing strain variants must therefore be identified and subtracted from the list of variants in the pool of outcrossed F_2 mutant recombinants. Preferably, subtraction of crossing strain variants can be performed by sequencing the crossing strain to identify all the variants it contains. Alternatively, crossing-strain variants can be identified by combining the lists of variants identified in several other strains that were all outcrossed to the same crossing strain, which is the approach we chose here (Figure 11C).

It is also possible, in a similar manner, to exclude background variants (pale blue diamonds) from the analysis if desired (Figure 11, B and C), although a larger number of noncrossing strain variants are likely to make the method more accurate. It is important to note that in the case of

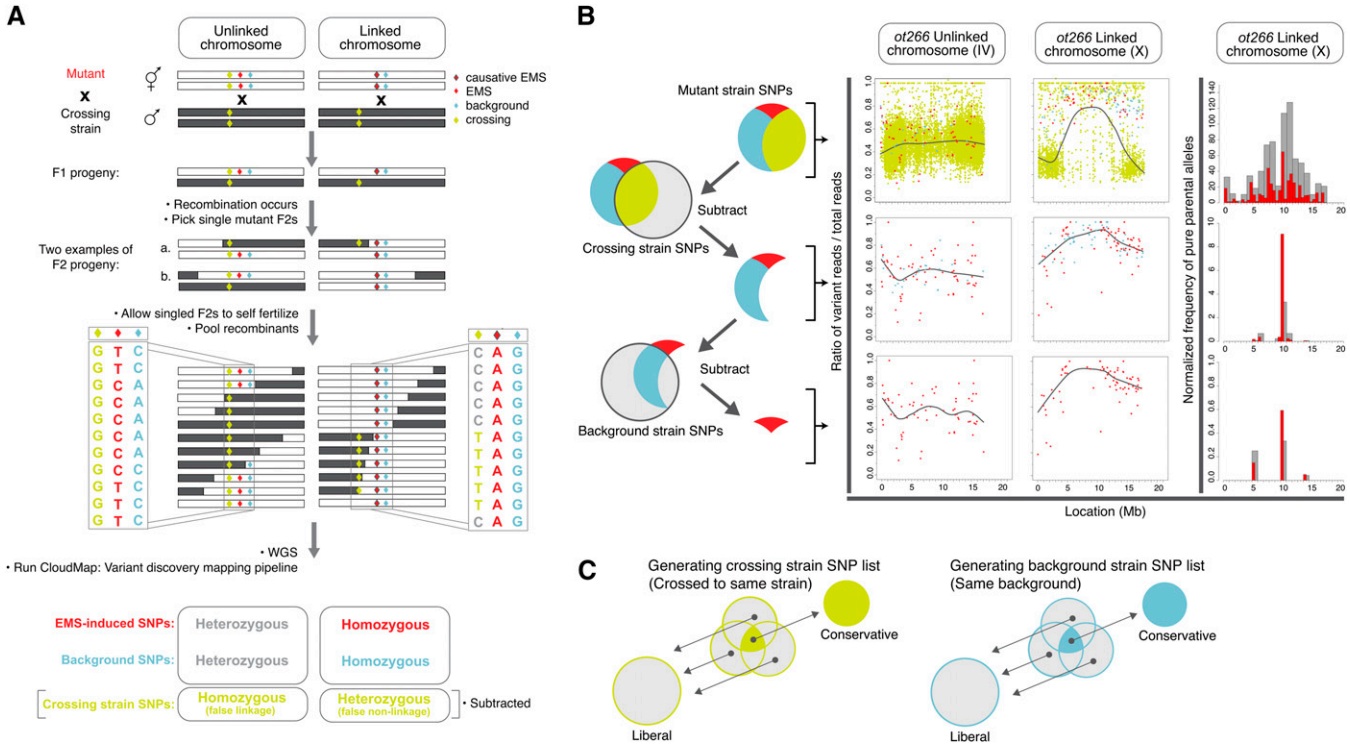


Figure 11 Variant Discovery Mapping. (A) Schematic representation of two extreme examples of the segregation of crossing strain variants (lime green diamonds), mutagen-induced variants (red diamonds), and background strain variants (pale blue diamonds) following an outcross of a mutant strain (white chromosome) to a nonparental (gray chromosome) strain. (B) Schematic representation of variant subtraction strategy for allele frequency plots. Allele frequency scatter plots display the ratio of variant reads/total reads at heterozygous and homozygous variant positions in the sequenced sample of pooled F_2 mutant progeny. Scatter plots are shown both prior to and after the successive subtraction of crossing-strain and background-strain variants. Color scheme is the same as in A. CloudMap Variant Discovery Mapping plots of normalized pure parental allele frequency for ot266. Note: *y*-axis scales are not consistent from panel to panel due to normalization. (C) Schematic representation of combining variant lists from other mutants to generate crossing-strain– or background-strain–specific variant lists for subtraction during Variant Discovery Mapping. Color scheme is the same as in A and B.

a suppressor screen, where the crossing strain is the parental premutagenesis strain, background variants will remain homozygous throughout and will be removed when subtracting crossing-strain variants. Thus in this particular case, only mutagen-induced variants can be used in the analysis.

The degree of mapping accuracy is likely to be strongly affected by the number of recombinants pooled, as well as both the total number and quality of variants analyzed. Our analysis and others have suggested that, in the case of *C. elegans*, both the accumulation of naturally occurring mutations and the addition of mutagen-induced mutations contribute to a high mutational load of close to 1000 SNPs across the genome of mutagenized strains (Flibotte *et al.* 2010; Sarin *et al.* 2010; Zuryn *et al.* 2010). This is likely to be enough to delineate a mapping region small enough to contain only a few candidate mutant variants.

Proof-of principle example for Variant Discovery Mapping

To determine the best parameters and to validate this strategy, we applied it to analyze the recessive *vab-3 (ot266)* allele that we have previously mapped and cloned using the polymorphic mapping strategy both with older WGS analysis tools (Doitsidou *et al.* 2010) and here with our new CloudMap pipeline. Utilizing the same dataset allows a direct comparison between polymorphic mapping and Variant Discovery Mapping. When the *vab-3(ot266)* data are processed using the Hawaiian polymorphic mapping strategy with the CloudMap Hawaiian Variant Mapping With WGS Data workflow, the mapping plots indicate a linked region on LG X between 10 and 11 Mb (Figure 6B). We find that the Variant Discovery Mapping strategy defines a very similar mapping region, despite using a total of only 575 background and EMS-induced SNPs and indels (Figure 11B, middle right plot). We describe below in more detail the individual steps required to perform Variant Discovery Mapping, all of which are automated when running the CloudMap Variant Discovery Mapping workflow that we have incorporated into our CloudMap pipeline (http://usegalaxy.org/cloudmap).

The first step is to determine which variants will be used for the analysis. One output of the CloudMap Variant Discovery Mapping workflow is a list of all high-quality heterozygous and homozygous variants present in the outcrossed pool of F_2 mutant recombinants as called by the GATK Unified Genotyper (DePristo *et al.* 2011). This genotyper uses a Bayesian genotype likelihood model to estimate genotypes and allele frequencies, detecting both SNPs and

indels (when used in combination with the BWA aligner). For Variant Discovery Mapping, we further apply a quality filter to this set of heterozygous and homozygous variants so that a single PHRED-scaled quality score reflects the probability of the variant allele (see *Materials and Methods*) (Danecek *et al.* 2011). Because we sequence pooled populations of F_2 mutant progeny, the definition of heterozygous and homozygous as defined by the GATK Unified Genotyper is not the same as it would be for a single strain. The ratio of variant reads/total reads at a given position approximately reflects the variant allele frequency in the pooled population and we therefore use the GATK defined quality score as a simple proxy for variants that are at least "heterozygous" in the pooled population of genomes to a large degree of confidence (see *Materials and Methods*). The goal is to plot variant read/total read ratios for all variants that are heterozygous or "homozygous" in the pooled population of recombinants. As GATK is an aggressive variant caller, many false positive heterozygous variant calls are those with low allele frequencies. Filtering for high-quality variants removes most of these false positive variants from the scatter plots. Rather than decreasing the prior likelihood value when running the genotyper, arbitrarily setting a read ratio cutoff point (*e.g.*, defining heterozygous variants as having a ratio of reference reads/variant reads >0.3), or invoking a complicated series of post-calling filters, all of which are valid methods to identify high-quality variants, we chose to filter using the single PHRED-scaled quality score assigned by GATK. The CloudMap Variant Discovery Mapping workflow produces output files filtered at Q100, Q200, and Q300 (see *Materials and Methods*). We empirically determined that filtering for a QScore ≥ 200 worked the best across several samples and several organisms. This score represents a 1 in $10^{\wedge}20$ chance of error in variant calling and we use this set of filtered variants for all downstream steps.

The second step is to subtract crossing strain variants (as described in the previous section). In the case of *vab-3(ot266)*, it was crossed to the highly polymorphic Hawaiian strain, which will introduce a large number of crossing strain variants. To remove these variants, rather than specifically sequence our Hawaiian strain, we combined the heterozygous and homozygous variants detected in six other strains that were also crossed to Hawaiian as illustrated in Figure 11C in addition to the unfiltered list of known Hawaiian variants. We took the liberal approach and subtracted variants present in at least one of these strains. The effects of this subtraction can be seen in Figure 11B (middle row of plots).

In the final step, we used a custom Python script to plot the ratio of variant reads/total reads for each variant position and a LOESS regression line (as discussed in the CloudMap Hawaiian Variant Mapping With WGS Data tool section) is used to reveal the trend. With relatively few data points, we empirically determined that a span of 0.4 works the best for all tested samples and species, although this parameter can be adjusted as desired when running the Variant Discovery Mapping tool (see *Materials and Methods*). The tool additionally plots the normalized frequency of pure parental allele positions (variant reads/total reads is equal to 1) across each chromosome. As can be seen in Figure 11B (middle plots for linked chromosome X), both Variant Discovery Mapping plots define a mapping region on LG X between 10 and 11 Mb. Despite containing ~200-fold fewer variants, this is an identical mapping region to that defined by Hawaiian Variant Mapping and examination of homozygous variants in this region (using the less stringent list of homozygous variants output as part of the CloudMap Variant Discovery Mapping workflow) reveals only two protein coding variants, one of which is the premature stop in the *vab-3* locus.

It is also possible to distinguish background variants from EMS-induced variants by subtracting those variants present in other WGS strains from the same parental strain (in this case *ot260* and *ot263*) (Figure 11B, bottom row of plots) or to remove indels from the analysis. In the case of *vab-3 (ot266)*, removing background variants reduced the accuracy of the mapping (Figure 11B, bottom row of plots for linked chromosome X) but removing indels had relatively little effect (data not shown). We conclude that it is important to retain as many noncrossing strain variants as possible to achieve the best mapping accuracy.

Validation of the CloudMap Variant Discovery Mapping method in other organisms

In principle, Variant Discovery Mapping should work in any organism provided the numbers of homozygous background and mutagen-induced variants present in the starting strain are high enough. To test this, we analyzed the slow growth *Arabidopsis* mutant described in Schneeberger *et al.* (2009). This recessive mutant was isolated in the Columbia background, crossed to the polymorphic Landsberg strain, and 500 F_2 mutant recombinants were pooled and whole genome sequenced. We ran this publicly available WGS dataset through the CloudMap Variant Discovery Mapping workflow and compared these plots to standard Hawaiian Variant Mapping plots (generated using the CloudMap Hawaiian Variant Mapping With WGS Data workflow described above). We identified crossing-strain variants using the list of variants generated by analysis of a second mutant dataset that had also been crossed to the Landsberg polymorphic strain and subtracted them (Galvao *et al.* 2012). We find that Variant Discovery Mapping localizes the linked region to the far right end of chromosome IV between 16 and 18 Mb and is almost as accurate as polymorphic mapping (Figure S5). This indicates that our Variant Discovery Mapping pipeline is applicable to other species. The CloudMap Variant Discovery Mapping tool natively supports *Arabidopsis* and is easily configured for other species (see CloudMap user guides and videos at http://usegalaxy.org/cloudmap).

Conclusion

We have established a cloud-based pipeline that greatly simplifies the analysis of mutant genome sequences. Available

on the Galaxy platform, CloudMap requires no software installation when run on the cloud (as opposed to locally or via Amazon's EC2 service) and is modular and thus able to accommodate the latest software tools as they become available. CloudMap uses a series of predefined workflows that allow users to arrive at a mapping region and a list of variants with a few simple clicks. We encourage users to check for updates of the CloudMap pipeline at http://usegalaxy.org/cloudmap.

Regarding the preferred mapping strategy, our current best practice recommendation is to use Hawaiian Variant Mapping if there are no obvious issues with the polymorphic Hawaiian strain. In this approach, no additional WGS information from other strains is required. If a laboratory has extensive WGS information on other strains (the "crossing strain" in the Variant Discovery Mapping method) or if the Hawaiian strain causes problems, we recommend using the Variant Discovery Mapping approach.

Acknowledgments

We thank members of the Hobert lab, many colleagues in the *Caenorhabditis elegans* community for reading and commenting on the manuscript, and Alexander Boyanov for his expert handling of our WGS operations. This work was funded in part by the National Institutes of Health (NIH) (R01NS039996-05 and R01NS050266-03) and the Howard Hughes Medical Institute. G.M. was supported by a F31 predoctoral fellowship (1F31NS074841-01). We thank the Galaxy Team for their continuing help on this project. Galaxy is supported by US National Institutes of Health grants HG005133, HG004909 and HG006620 and US National Science Foundation grant DBI 0850103. Additional funding is provided, in part, by the Huck Institutes for the Life Sciences at Penn State, the Institute for Cyberscience at Penn State and a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

Literature Cited

Abe, A., S. Kosugi, K. Yoshida, S. Natsume, H. Takagi *et al.*, 2012 Genome sequencing reveals agronomically important loci in rice using MutMap. Nat. Biotechnol. 30: 174–178.

Afgan, E., D. Baker, N. Coraor, H. Goto, I. M. Paul *et al.*, 2011 Harnessing cloud computing with Galaxy Cloud. Nat. Biotechnol. 29: 972–974.

Bigelow, H., M. Doitsidou, S. Sarin, and O. Hobert, 2009 MAQGene: software to facilitate C. elegans mutant genome sequence analysis. Nat. Methods 6: 549.

Blankenberg, D., A. Gordon, G. Von Kuster, N. Coraor, J. Taylor *et al.*, 2010 Manipulation of FASTQ data with Galaxy. Bioinformatics 26: 1783–1785.

Cingolani, P., and A. Platts, L. Wang le, M. Coon, T. Nguyen *et al.*, 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6: 80–92.

Cleveland, W. S., E. Grosse, and W. M. Shyu, 1992 Local regression models, *Statistical Models in S*, edited by J. M. Chambers, and T. J. Hastie. CRC Press, Boca Raton, FL.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498.

Doitsidou, M., N. Flames, A. C. Lee, A. Boyanov, and O. Hobert, 2008 Automated screening for mutants affecting dopaminergic-neuron specification in C. elegans. Nat. Methods 5: 869–872.

Doitsidou, M., R. J. Poole, S. Sarin, H. Bigelow, and O. Hobert, 2010 C. elegans mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. PLoS ONE 5: e15435.

Ewing, B., and P. Green, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8: 186–194.

Flibotte, S., M. L. Edgley, I. Chaudhry, J. Taylor, S. E. Neil *et al.*, 2010 Whole-genome profiling of mutagenesis in *Caenorhabditis elegans*. Genetics 185: 431–441.

Flowers, E. B., R. J. Poole, B. Tursun, E. Bashllari, I. Pe'er *et al.*, 2010 The Groucho ortholog UNC-37interacts with the short Groucho-like protein LSY-22 to control developmental decisions in C. elegans. Development 137: 1799–1805.

Galvao, V. C., K. J. Nordstrom, C. Lanz, P. Sulz, J. Mathieu *et al.*, 2012 Synteny-based mapping-by-sequencing enabled by targeted enrichment. Plant J. 71: 517–526.

Harris, T. W., N. Chen, F. Cunningham, M. Tello-Ruiz, I. Antoshechkin *et al.*, 2004 WormBase: a multi-species resource for nematode biology and genomics. Nucleic Acids Res. 32(Database issue): D411–D417.

Hillier, L. W., G. T. Marth, A. R. Quinlan, D. Dooling, G. Fewell *et al.*, 2008 Whole-genome sequencing and variant discovery in C. elegans. Nat. Methods 5: 183–188.

Hobert, O., 2010 The impact of whole genome sequencing on model system genetics: get ready for the ride. Genetics 184: 317–319.

Hobert, O., 2012 The neuronal genome of Caenorhabditis elegans. WormBook (in press).

Janssen, T., M. Lindemans, E. Meelkop, L. Temmerman, and L. Schoofs, 2010 Coevolution of neuropeptidergic signaling systems: from worm to man. Ann. N. Y. Acad. Sci. 1200: 1–14.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle *et al.*, 2002 The human genome browser at UCSC. Genome Res. 12: 996–1006.

Kim, J. K., H. W. Gabel, R. S. Kamath, M. Tewari, A. Pasquinelli *et al.*, 2005 Functional genomic analysis of RNA interference in C. elegans. Science 308: 1164–1167.

Kim, S., J. A. Govindan, Z. J. Tu, and D. Greenstein, 2012 The SACY-1 DEAD-box helicase links the somatic control of oocyte meiotic maturation to the sperm-to-oocyte switch and gamete maintenance in *Caenorhabditis elegans*. Genetics (in press).

Labed, S. A., S. Omi, M. Gut, J. J. Ewbank, and N. Pujol, 2012 The pseudokinase NIPI-4 is a novel regulator of antimicrobial peptide gene expression. PLoS ONE 7: e33887.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25.

Li, H., and R. Durbin, 2010 Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589–595.

Li, H., and N. Homer, 2010 A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. 11: 473–483.

Li, H., J. Ruan, and R. Durbin, 2008 Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18: 1851–1858.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Nicol, J. W., G. A. Helt, S. G. Blanchard Jr., A. Raja, and A. E. Loraine, 2009 The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics 25: 2730–2731.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12: 443–451.

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Reece-Hoyes, J. S., A. Diallo, B. Lajoie, A. Kent, S. Shrestha *et al.*, 2011 Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. Nat. Methods 8: 1059–1064.

Sarin, S., V. Bertrand, H. Bigelow, A. Boyanov, M. Doitsidou *et al.*, 2010 Analysis of multiple ethyl methanesulfonate-mutagenized caenorhabditis elegans strains by whole-genome sequencing. Genetics 185: 417–430.

Schneeberger, K., S. Ossowski, C. Lanz, T. Juul, A. H. Petersen *et al.*, 2009 SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat. Methods 6: 550–551.

Seidel, H. S., M. V. Rockman, and L. Kruglyak, 2008 Widespread genetic incompatibility in C. elegans maintained by balancing selection. Science 319: 589–594.

Shaye, D. D., and I. Greenwald, 2011 OrthoList: a compendium of C. elegans genes with human orthologs. PLoS ONE 6: e20085.

Tursun, B., T. Patel, P. Kratsios, and O. Hobert, 2011 Direct conversion of C. elegans germ cells into specific neuron types. Science 331: 304–308.

Vastenhouw, N. L., K. Brunschwig, K. L. Okihara, F. Muller, M. Tijsterman *et al.*, 2006 Gene expression: long-term gene silencing by RNAi. Nature 442: 882.

Wang, D., S. Kennedy, D. Conte Jr., J. K. Kim, H. W. Gabel *et al.*, 2005 Somatic misexpression of germline P granules and enhanced RNA interference in retinoblastoma pathway mutants. Nature 436: 593–597.

Zhang, F., M. M. O'Meara, and O. Hobert, 2011 A left/right asymmetric neuronal differentiation program is controlled by the *Caenorhabditis elegans* lsy-27 zinc-finger transcription factor. Genetics 188: 753–759.

Zuryn, S., S. Le Gras, K. Jamet, and S. Jarriault, 2010 A strategy for direct mapping and identification of mutations by whole-genome sequencing. Genetics 186: 427–430.

Communicating editor: D. Greenstein

GENETICS

Supporting Information
http://www.genetics.org/content/suppl/2012/10/08/genetics.112.144204.DC1

CloudMap: A Cloud-Based Pipeline for Analysis of Mutant Genome Sequences

Gregory Minevich, Danny S. Park, Daniel Blankenberg, Richard J. Poole, and Oliver Hobert

DOI: 10.1534/genetics.112.144204

A

	A	B
1	C35E7.2	2
2	Y8A9A.2	2
3	Y16B4A.2	2

B

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Sample	# Chromo	Position	Reference	Change	Change_type	Homozygous	Quality	Coverage	Warnings	Gene_ID	Gene_name	Bio_type	Trancript_ID	Exon_ID	Exon_Rank	Effect	old_AA/new_AA	Old_codon/New_codon	Codon_Num	Codon_Dege	CDS_size
2	mutA	I	10841384	G	C	SNP	Hom	43.12	2		C35E7.2	C35E7.2	protein_coding	C35E7.2a	exon_I_10841103_10841965	1	NON_SYNONYMOUS_CODING	R/T	aGa/aCa	92	0	2124
3	mutB	I	10841434	A	C	SNP	Hom	80.72	3		C35E7.2	C35E7.2	protein_coding	C35E7.2a	exon_I_10841103_10841965	1	NON_SYNONYMOUS_CODING	I/L	Att/Ctt	109	0	2124
4	mutB	II	3796684	C	A	SNP	Hom	349.22	21		Y8A9A.2	Y8A9A.2	protein_coding	Y8A9A.2	exon_II_3796348_3797638	5	NON_SYNONYMOUS_CODING	P/Q	cCa/cAa	286	0	4083
5	mutC	II	3796759	A	T	SNP	Hom	1208.35	56		Y8A9A.2	Y8A9A.2	protein_coding	Y8A9A.2	exon_II_3796348_3797638	5	NON_SYNONYMOUS_CODING	N/I	aAt/aTt	311	0	4083
6	mutA	X	14766637	G	A	SNP	Hom	44.89	4		Y16B4A.2	Y16B4A.2	protein_coding	Y16B4A.2	exon_X_14766327_14766971	18	NON_SYNONYMOUS_CODING	S/F	tCc/tTc	1073	0	6504
7	mutB	X	14766625	*	—G	DEL	Hom	506.78	21		Y16B4A.2	Y16B4A.2	protein_coding	Y16B4A.2	exon_X_14766327_14766971	18	FRAME_SHIFT: Y16B4A.2					6504

Figure S1 CloudMap *in silico Complementation Test* tool. **A:** Summary output. CloudMap allows for large scale *in silico* comparison of annotated WGS variants (that have been filtered for quality and had common variants subtracted) between many samples. The summary output from this comparison shows the number of alleles of each gene sorted from most to fewest. **B:** Comprehensive output. For each in silico Complementation Test summary output file, CloudMap provides the corresponding detailed list of snpEff-annotated, allelic gene hits that is also sorted from most to fewest alleles.

Figure S2 Uncovered region confirmed to be a genomic deletion. CloudMap contains a workflow for annotating uncovered regions that may be genomic deletions. Users are encouraged to check if uncovered regions repeatedly appear in other strains and also to view these putative deletions in an alignment viewer. We find that true deletions tend to exhibit a cliff of high coverage followed by zero coverage on both uncovered boundary regions. Regions of high coverage flanking the putative deletion also often have SNPs or insertions present in many of the reads — indicating that distant genomic regions are now adjacent to one another. The deletion shown was confirmed to be a deletion via PCR and Sanger sequencing. The IGV viewer is used to display the alignment (Robinson et al., 2011, *Nature Biotechnology* 29, 24–26).

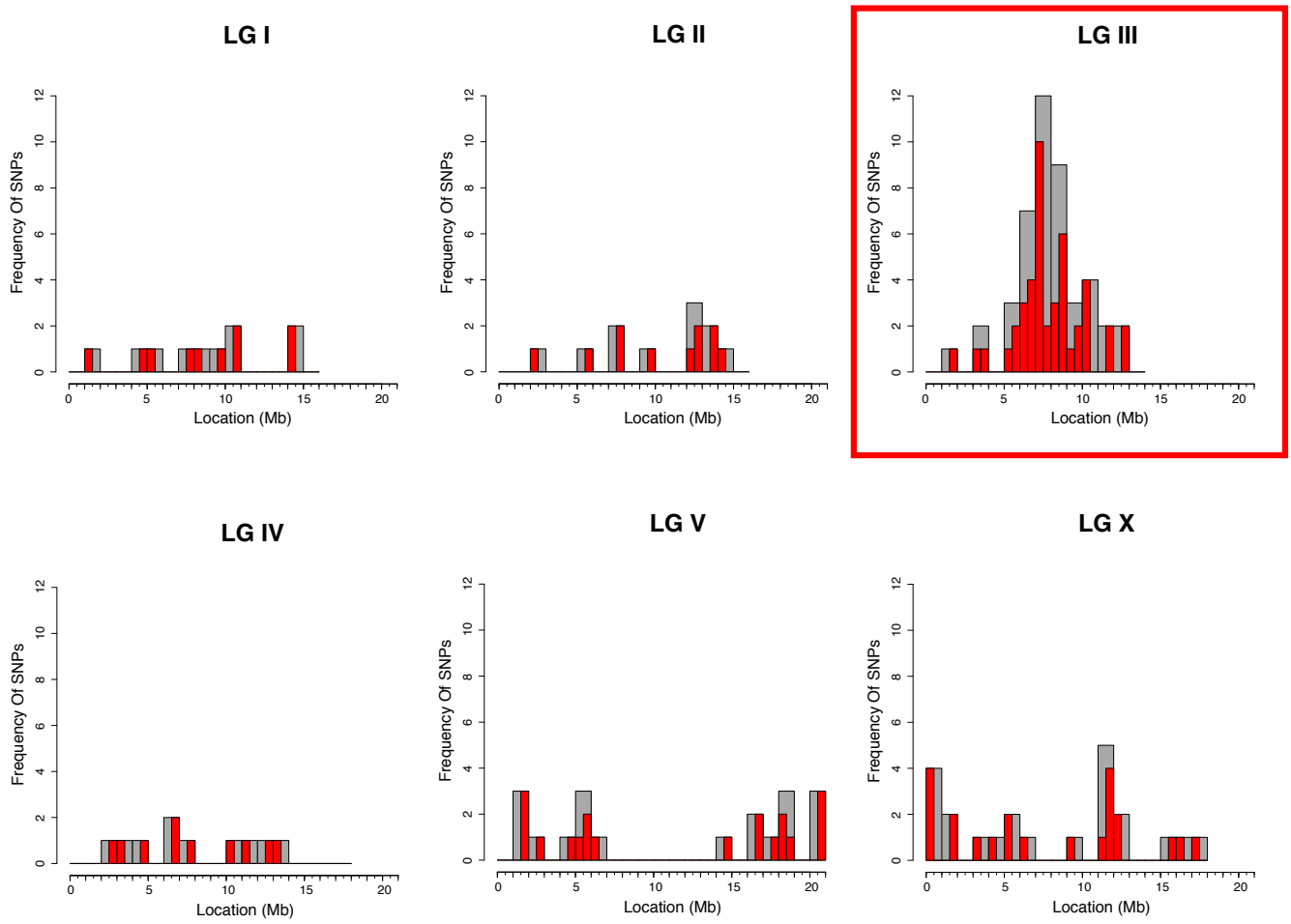


Figure S3 *EMS Variant Density Mapping* tool: CloudMap also supports the approach detailed in Zuryn et al., that involves plotting frequencies of variant density in a mutant *C. elegans* strain that has been backcrossed to its (pre-mutagenesis) starting strain (ZURYN *et al.* 2010).

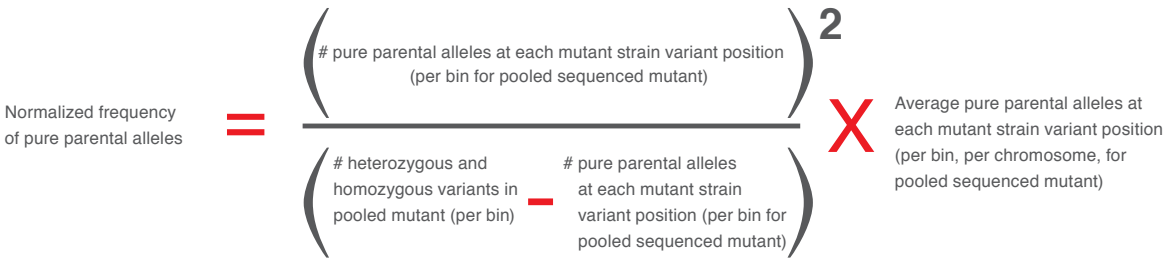


Figure S4 *Variant Discovery Mapping* normalization equation. Pure parental alleles are defined as those positions in the pooled sequenced mutant where variant reads/total reads = 1 (after the appropriate variant subtraction strategy has been applied). Normalization is applied by default although users have the option of turning it off.

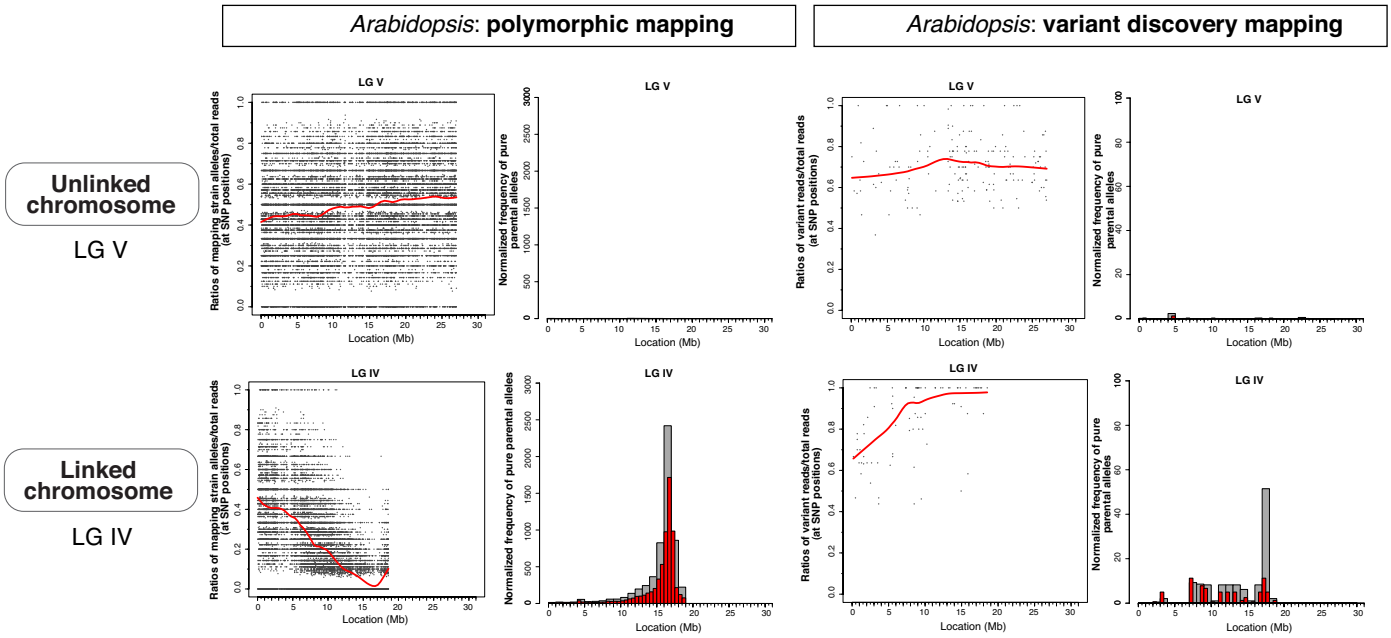


Figure S5 Multi-organism *Variant Discovery Mapping* support. CloudMap natively supports the variant discovery mapping method for *Arabidopsis* as shown here for data from Schneeberger *et al.* 2009. Users must provide a simple configuration file for organisms other than C. elegans and Arabidopsis. Configuration files and instructions for other organism support are provided at http://usegalaxy.org/cloudmap

Table S1 WS220 Hawaiian Variants filtered to assist in CloudMap *Hawaiian Variant Mapping with WGS Data* plot normalization. This table provides details on the numbers of Hawaiian SNPs subtracted from mapping analysis for purposes of SNP mapping plot normalization. Details provided in text and in **Fig.7**.

Mutant	Trans-gene	Location of trans-gene (LG)	Location of Mutation (LG)	LG I	LG II	LG III	LG IV	LG V	LG X	< .05 or > .95 ratio positions	< .05 or > .95 ratio positions after transgene & mutation LGs removed
ot219	*otIs114*	I	V		✓	✓	✓		✓	15,873	4,003
ot266	*vtIs1*	V	X	✓	✓	✓	✓		✓*	13,669	11,033
ot628	*oxIs12, vsIs33*	V, X	I, V		✓	✓	✓			22,268	3,367
ot641	*otIs138, vIs33*	V, X	X	✓	✓	✓	✓			9,616	4,314
ot642	*otIs138, vsI33*	V, X	X	✓	✓	✓	✓		✓**	13,589	5,021
ot704	*otIs341*	X	I		✓	✓	✓	✓		14,805	7,197
ot705	*otIs341*	X	III	✓	✓		✓	✓		14,686	7,969

✓ Considered SNPs on this chromosome for filtering

* Excluding 6-13 Mb

** Excluding 0-6 Mb

Total dataset used:

WS220_WormMart HA SNPs: 112,061 Variants

<.05 or >.95 ratio positions in at least 2 samples: 8,715 Variants

<.05 & >.95 ratio positions in all 7 samples:1,563 Variants

Final curated Hawaiian SNP list (WS220.64): 103,346 Variants

Table S2 CloudMap comparison with MAQGene. This table shows variants in the mapping region of *ot266* as determined by CloudMap vs. MAQGene. CloudMap was able to identify a smaller mapping region than MAQGene (1Mb vs. 2.13Mb).

	MAQGene	**CloudMap**
# of pooled recombinants	50	50
Defined mapping interval	8,841,415-10,975,250	10,000,000-11,000,000
	(aligned to WS201)	(aligned to WS220)
Defined mapping region in Mb	2.13	1
# Variants in the region	26	22
(pre-variant subtraction)		
# Variants in the region	not performed	10
(post- variant subtraction)		
# of protein coding variants in respective mapping regions	3	2
Premature stops	1	1

Additional Supporting Materials

Video user guides, automated workflows, and up to date CloudMap tools are available at: http://usegalaxy.org/cloudmap