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

Genetics, December 2012

Speakers: Gregory Minevich (Hobert lab, Columbia University)
Richard Poole (Welcome Trust Fellow UCL)

Structure of talk

- 45 minute talk

- 15 minutes for questions

- Greg & Rich available for detailed discussion the remainder of the meeting

Overview

1. What problems does CloudMap address?

2. Choosing the right cross for your experiment

3. Navigating within Galaxy

4. Support (website & video user guides)

5. Galaxy hosting options

Sequencing costs are dropping faster than you think

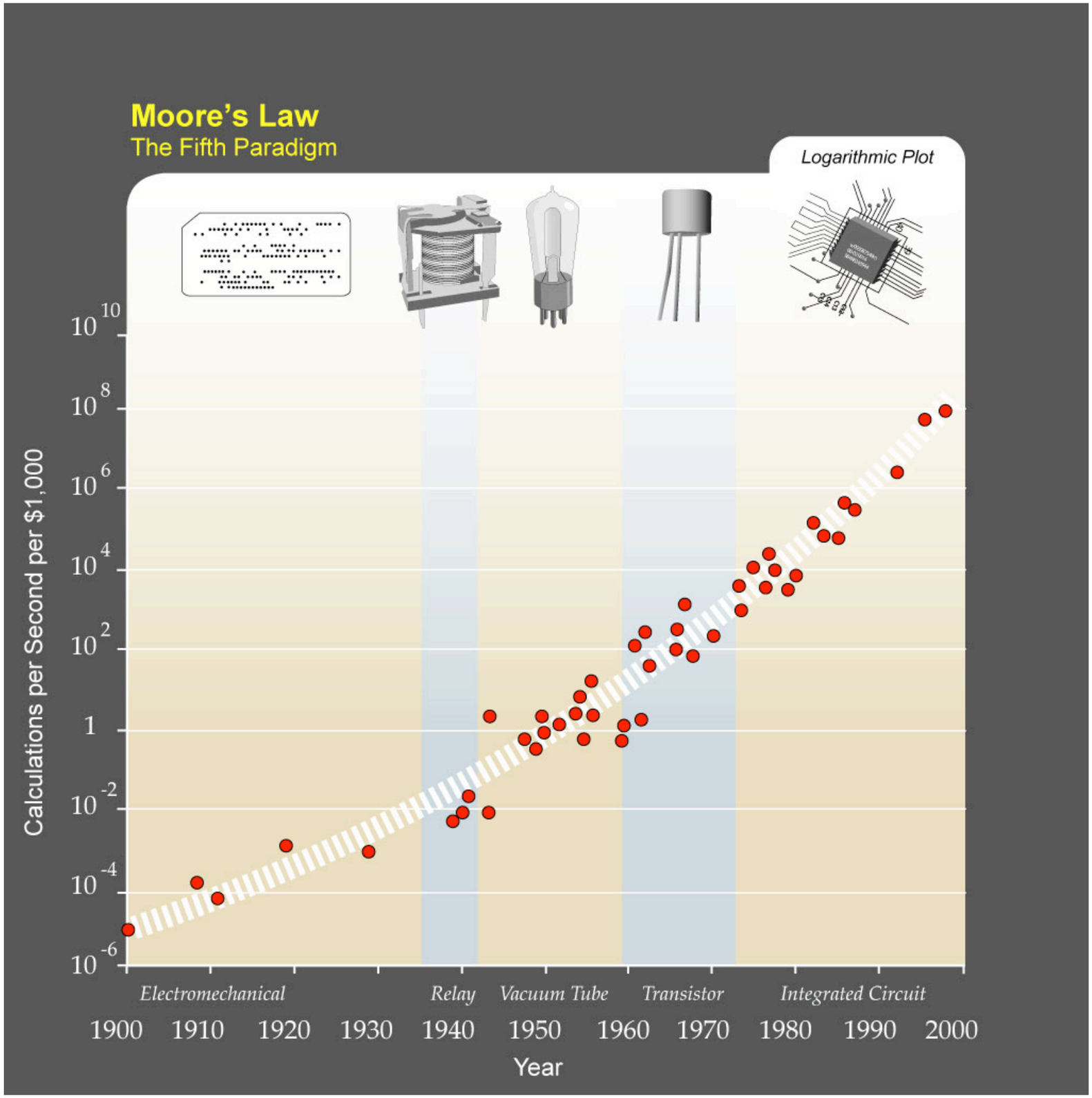


Image: Wikipedia

Sequencing costs are dropping faster than you think

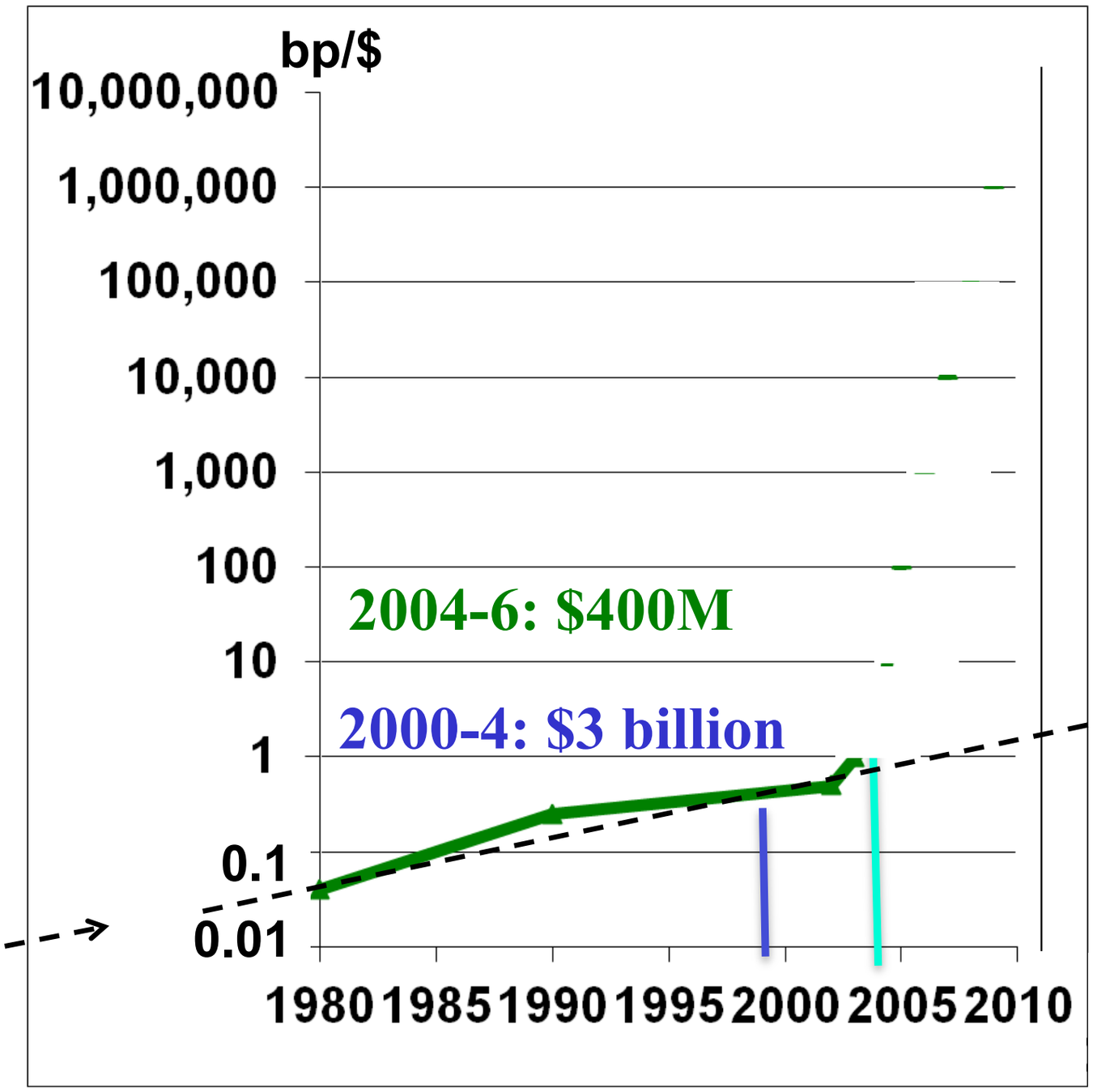
$1000 Genome

When?

2040

- - - - - -

Moore's law 1.5x/yr for electronics



Based on 1970-2004 exponential

Adapted from George Church 2012

(sequencing bp)/($) for whole genome sequencing is advancing faster than Moore's Law

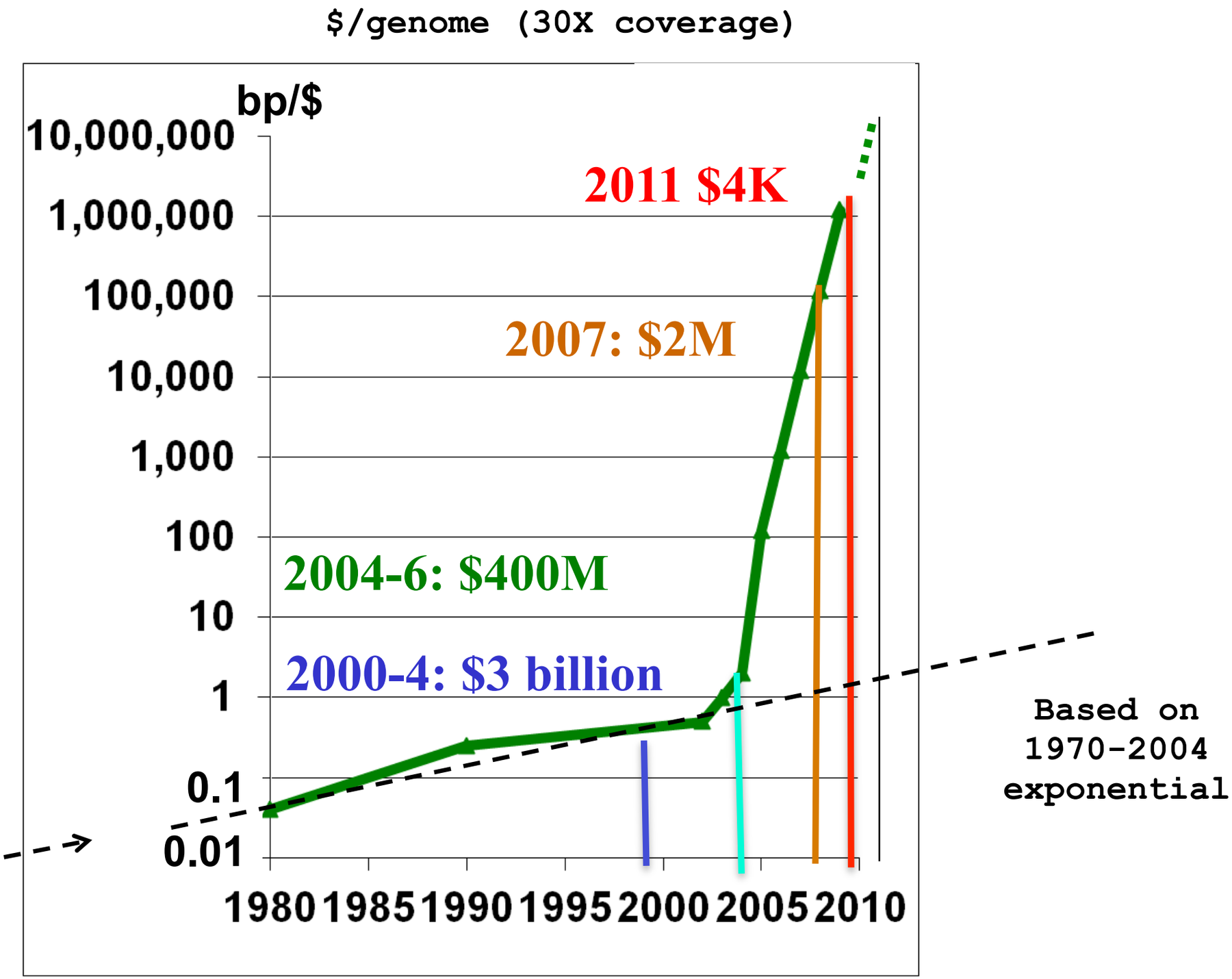
Factors of 10/yr

$1000 When?

2013

- - - - - -

Moore's law 1.5x/yr for electronics



Adapted from George Church 2012

whole genome sequencing is now affordable

The bottleneck is now data interpretation

- Flood of data requires expensive hardware
- Knowledgeable/expensive bioinformaticians
- Hardware or software may be quickly outdated.

CloudMap solves the data interpretation problem

- Cloud-based (Galaxy platform hosted at Penn State)
  - (can also install locally or use Amazon AMIs)
- Modular – latest aligners and variant callers
- Automated workflows
- Works with any organism
- Extensive pdf and video user guides
- **Free & open source**

CloudMap tools for finding the causal variant

- Clone mutants from mapping crosses (3 methods)

CloudMap tools for finding the causal variant

- Clone mutants from mapping crosses (3 methods)
- in silico complementation testing

CloudMap tools for finding the causal variant

- Clone mutants from mapping crosses (3 methods)

- in silico complementation testing

- Subtraction of common background variants and deletion analysis

CloudMap tools for finding the causal variant

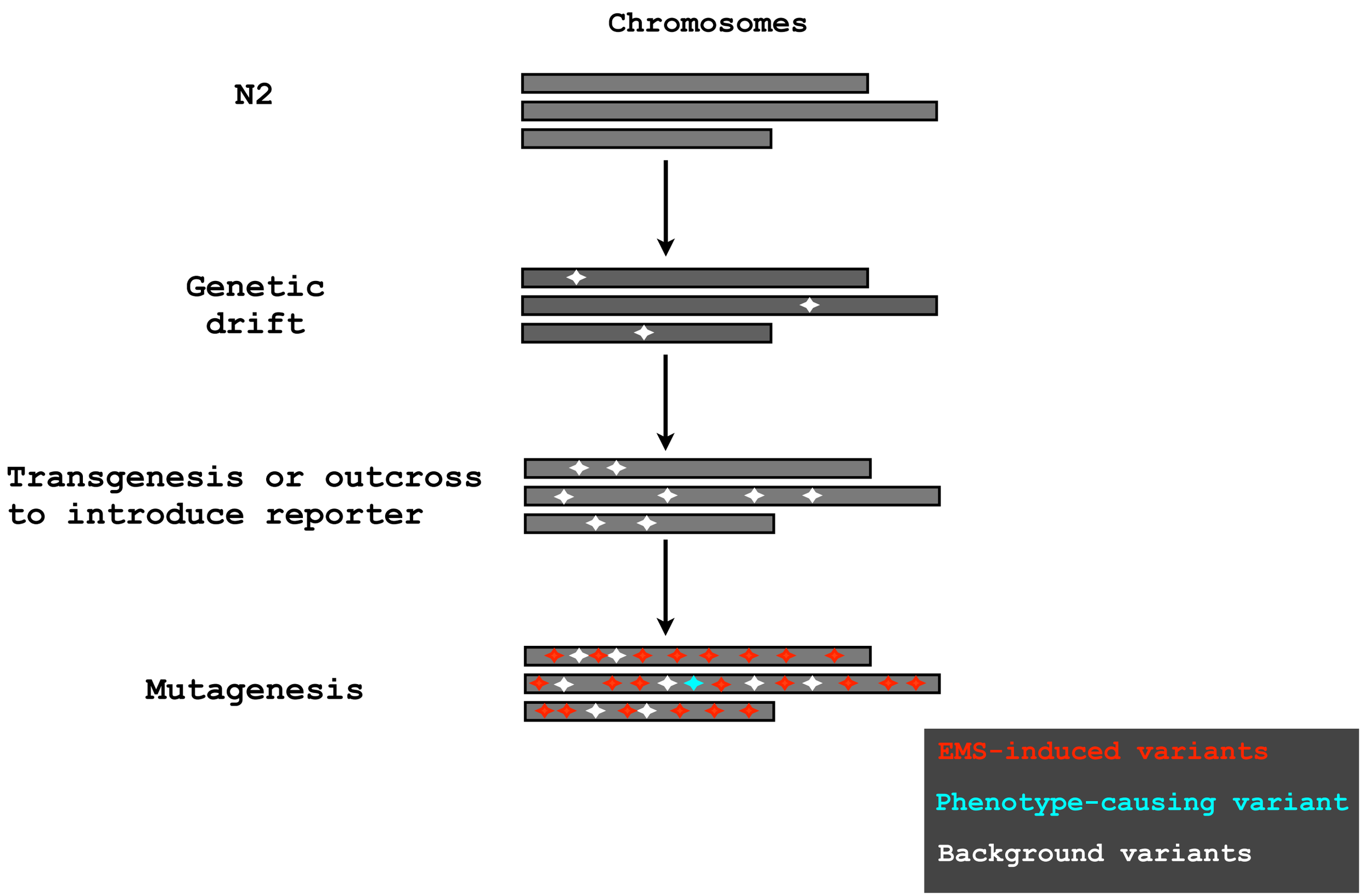
- Clone mutants from mapping crosses (3 methods)

- in silico complementation testing

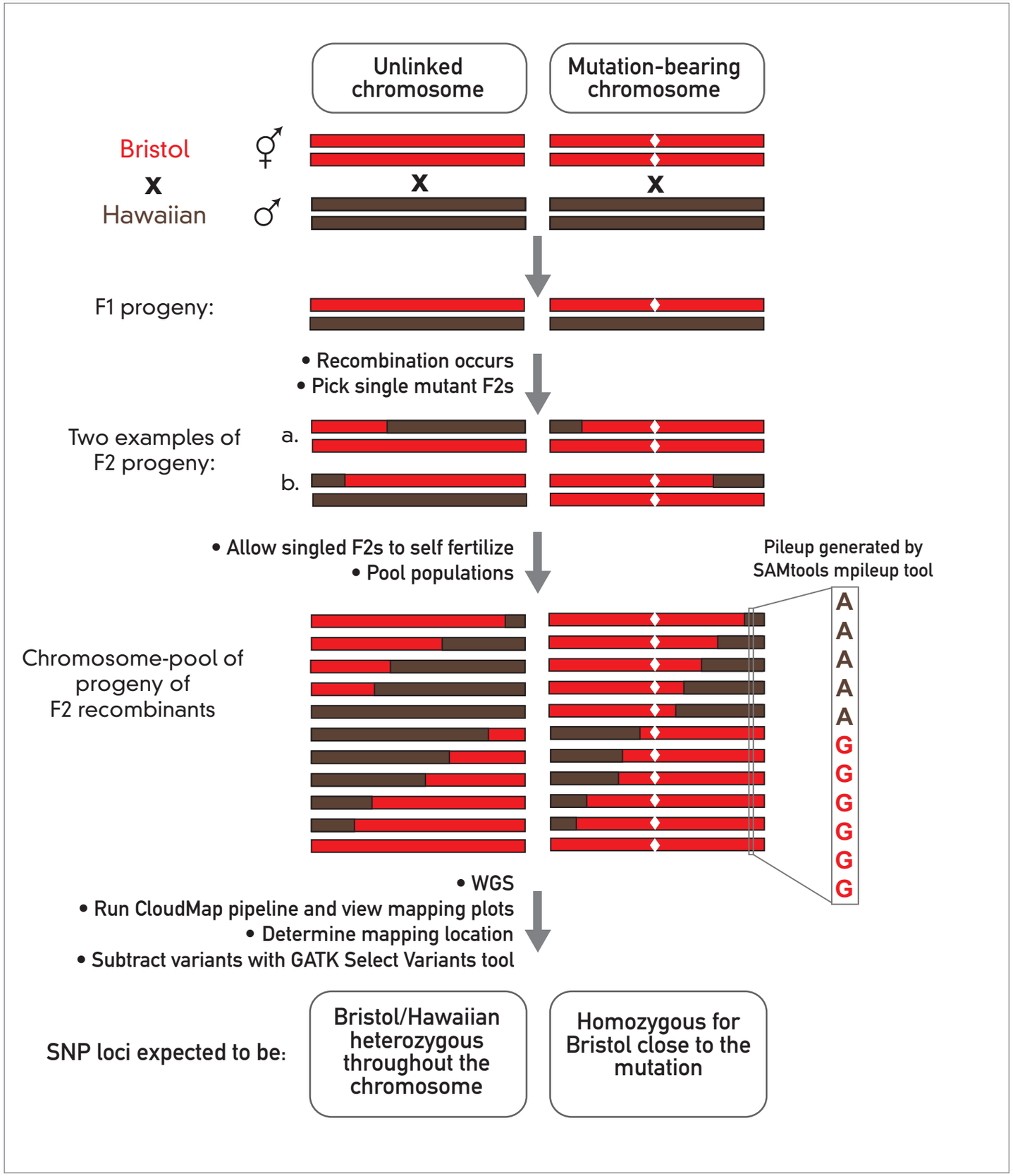
- Subtraction of common background variants and deletion analysis

- Query candidate gene lists

2. Choosing the right cross



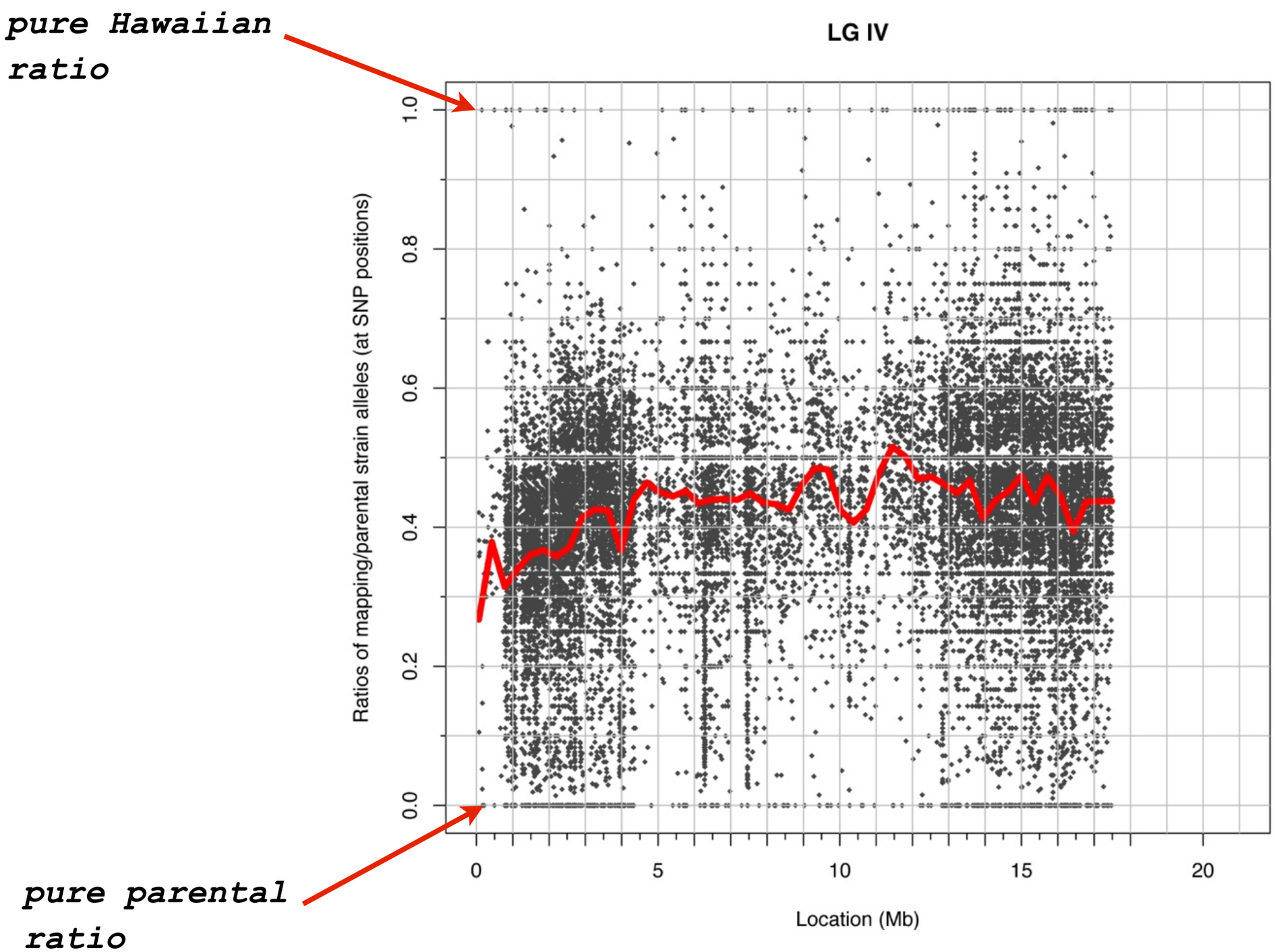
Hawaiian Variant Mapping concept



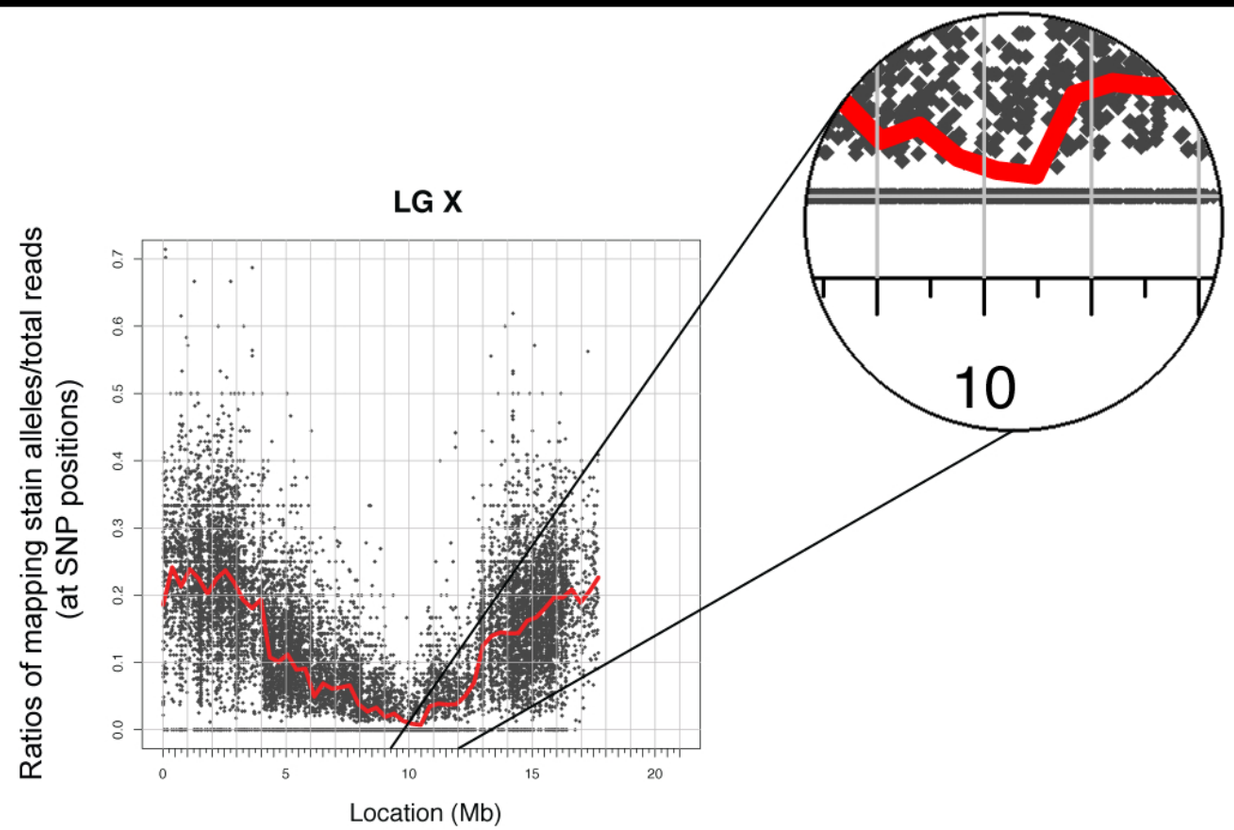
Schneeberger et al. 2009

Doitsidou et al. 2010

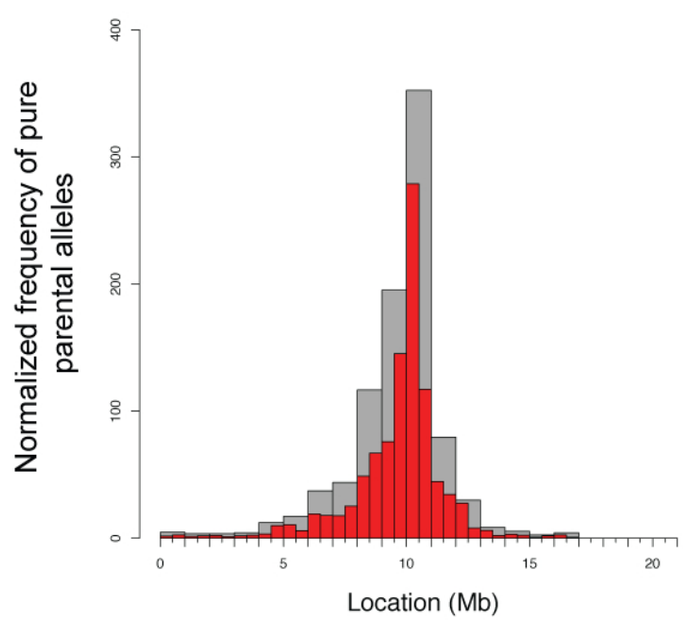
An unlinked chromosome



CloudMap's Hawaiian Variant Mapping plots



vab-3:LG X 10,503,393-
10,519,348



LG X

Hawaiian frequency plots normalized to 3 criteria

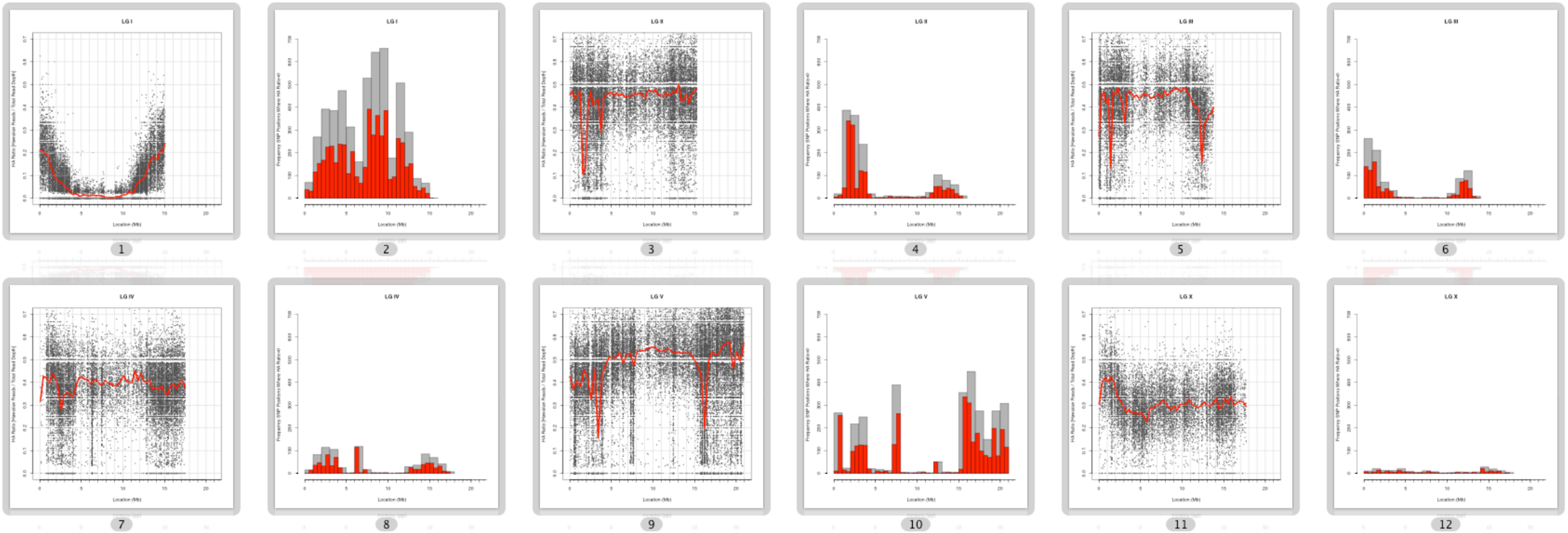
1. Amount of possible Hawaiian SNPs in a given Mb bin

2. Amount of pure parental alleles in that given Mb bin

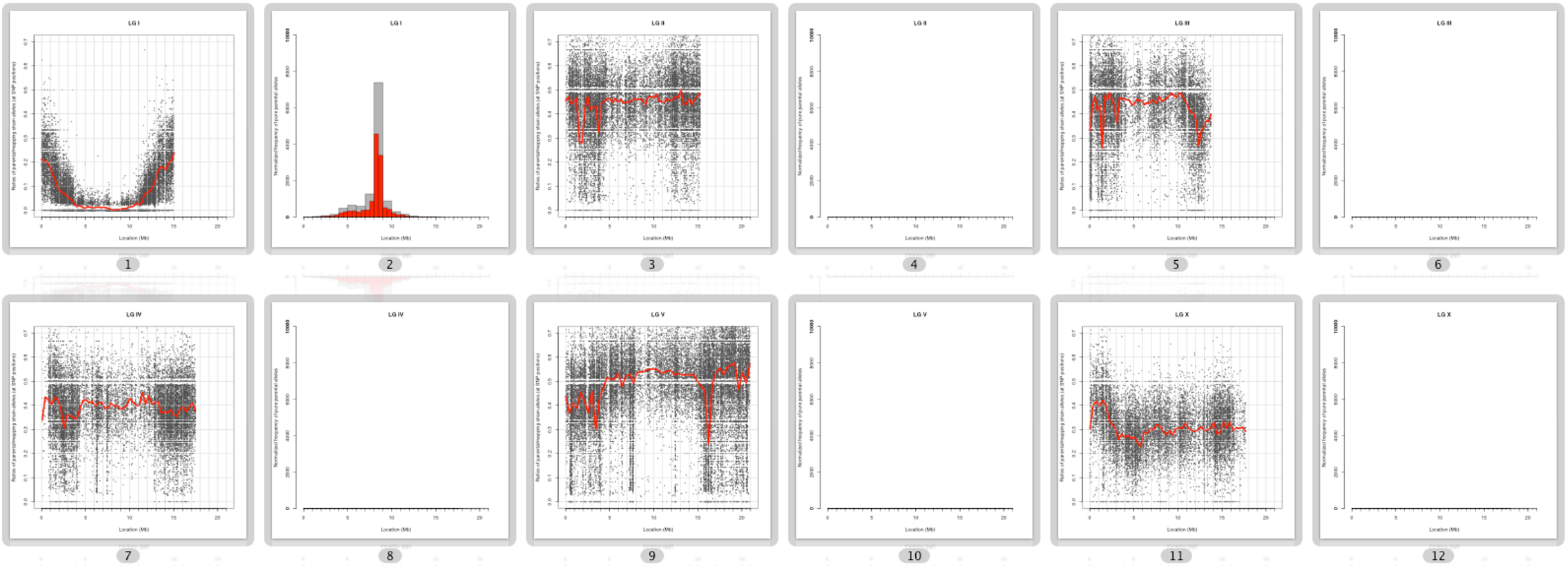
3. Average amount of Hawaiian variants in a given Mb bin per chromosome

HA mapping plots before/after normalization

Pre-Normalization:



Post-Normalization:



Variant Discovery Mapping (VDM)

- VDM allows you to map by crossing to N2 and **use the unique variants present in the EMS'd mutant for mapping**

Variant Discovery Mapping (VDM)

- VDM allows you to map by crossing to N2 and **use the unique variants present in the EMS'd mutant for mapping**

- What if you don't want to introduce ~110k Hawaiian mutations into your mutant strain?

  - e.g. behavioral screens

Variant Discovery Mapping (VDM)

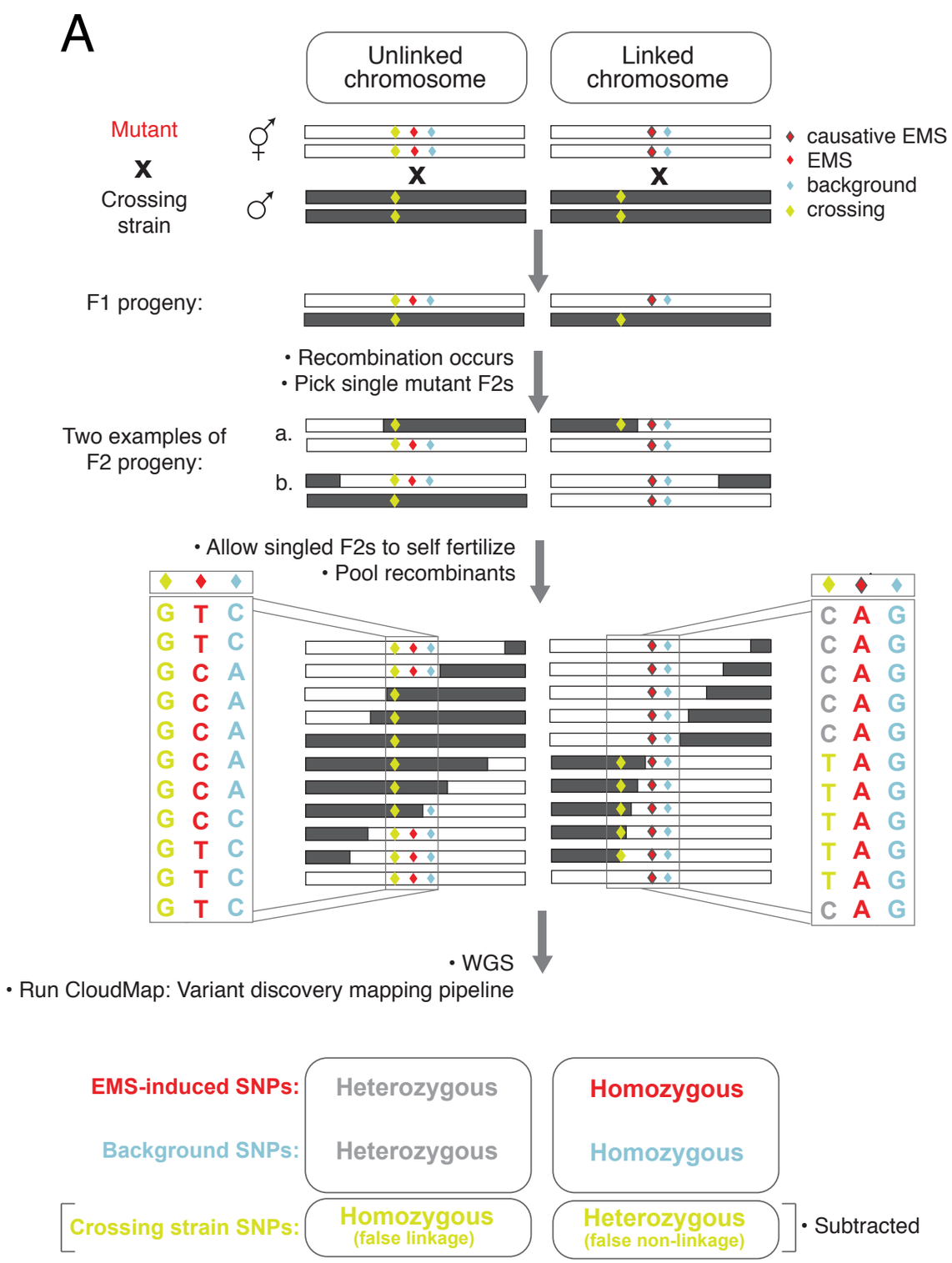
- VDM allows you to map by crossing to N2 and **use the unique variants present in the EMS'd mutant for mapping**

- What if you don't want to introduce ~110k Hawaiian mutations into your mutant strain?
  - e.g. behavioral screens

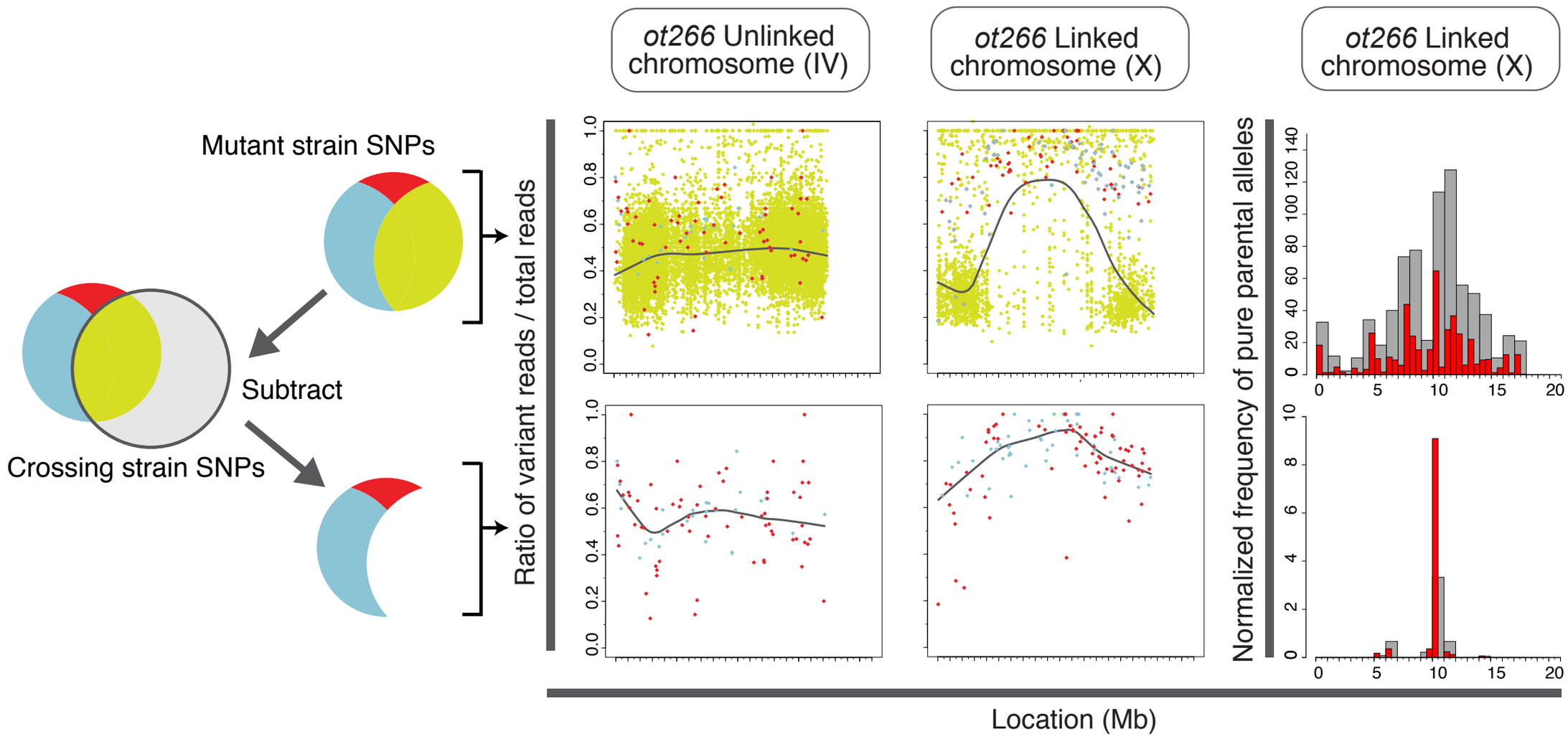
- What if you need to recover >1 mutation in your F2 progeny after a cross?
  - e.g. suppressor screens (**CAVEAT: not yet tried in worms, works in Arabidopsis**)

Variant Discovery Mapping

Fig.11



Variant Discovery Mapping



VDM frequency plots normalized to 3 criteria

1. Amount of potential parent strain SNPs in a given Mb bin

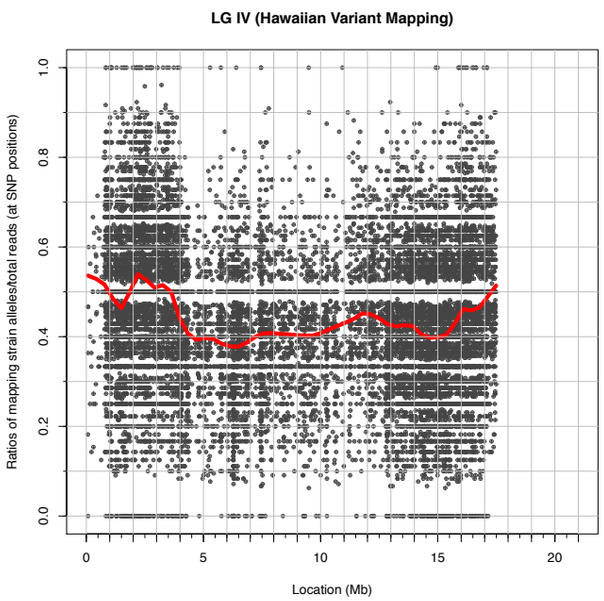
2. Amount of pure parental alleles in that given Mb bin

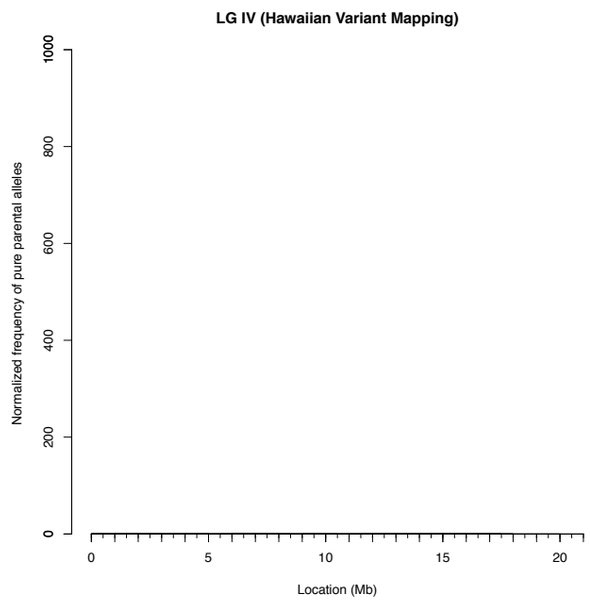
3. Average amount of pure parental alleles in a given Mb bin per chromosome

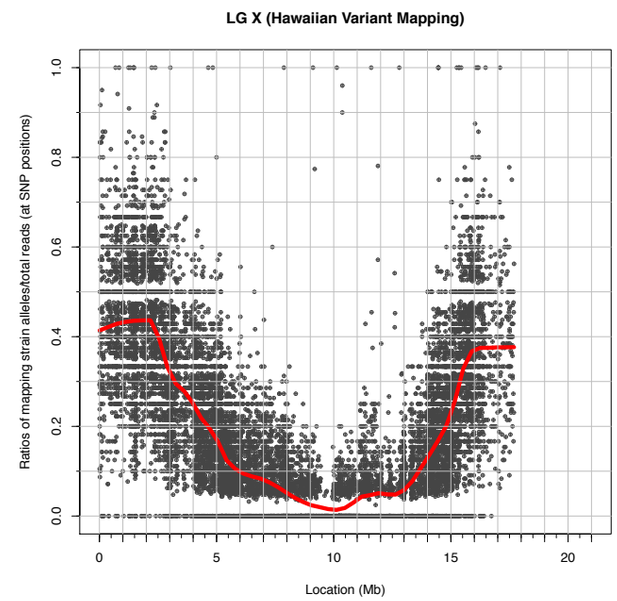
Hawaiian mapped mutants are also plotted with VDM automatically

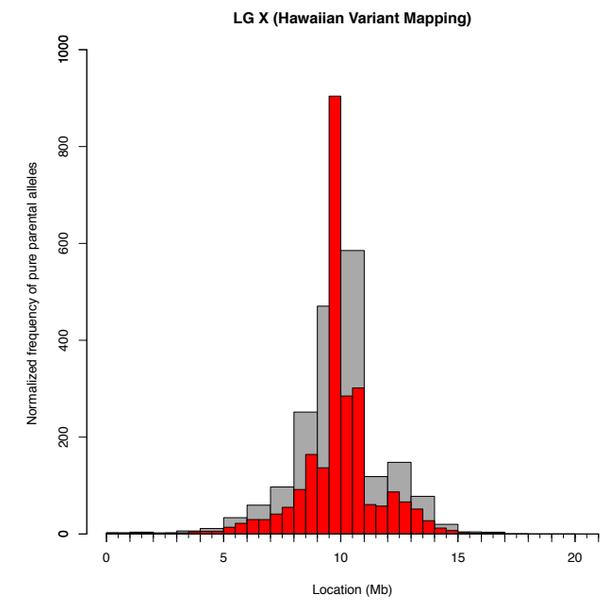
Unlinked LG

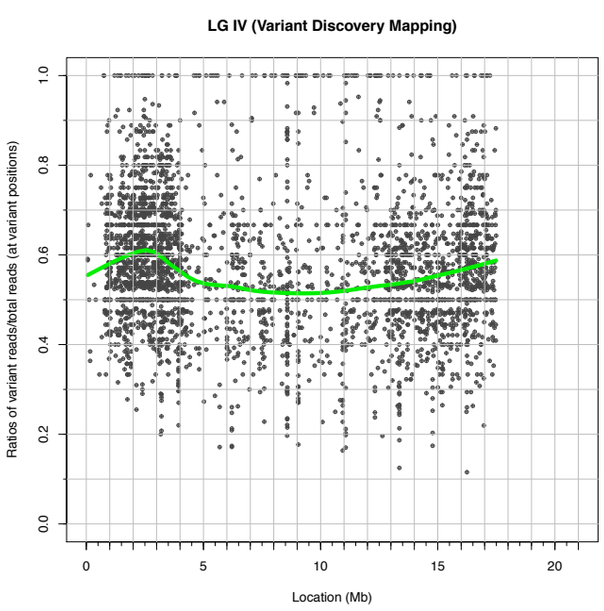
Linked LG

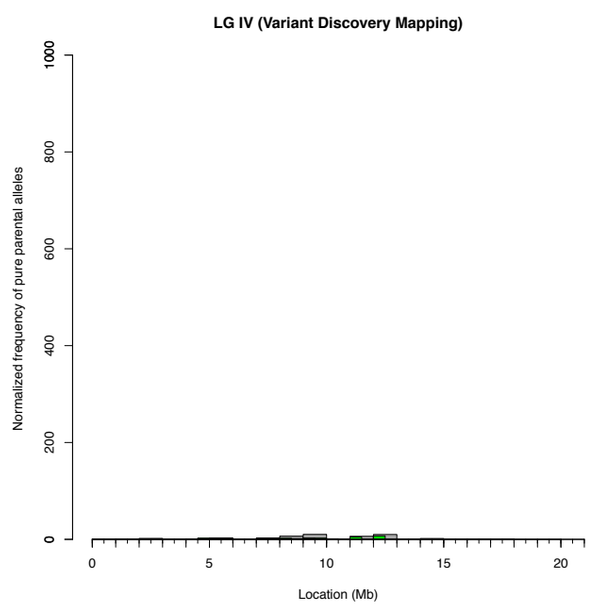


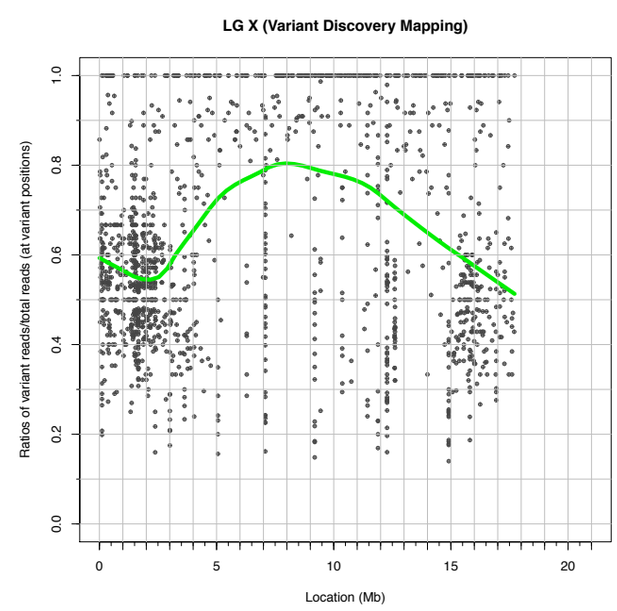


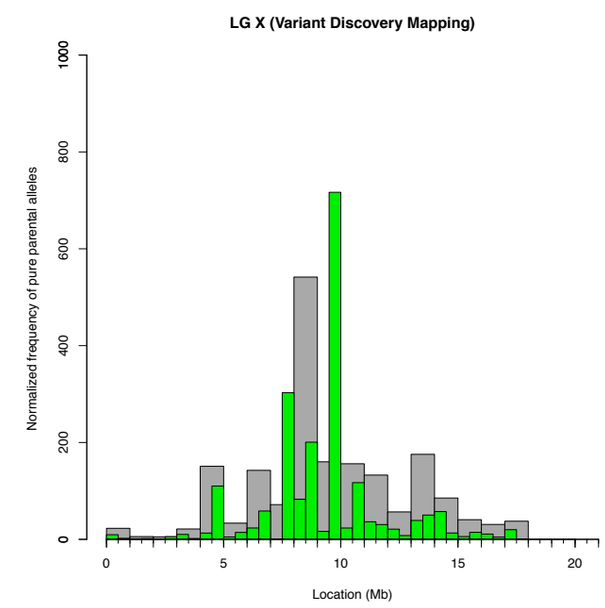










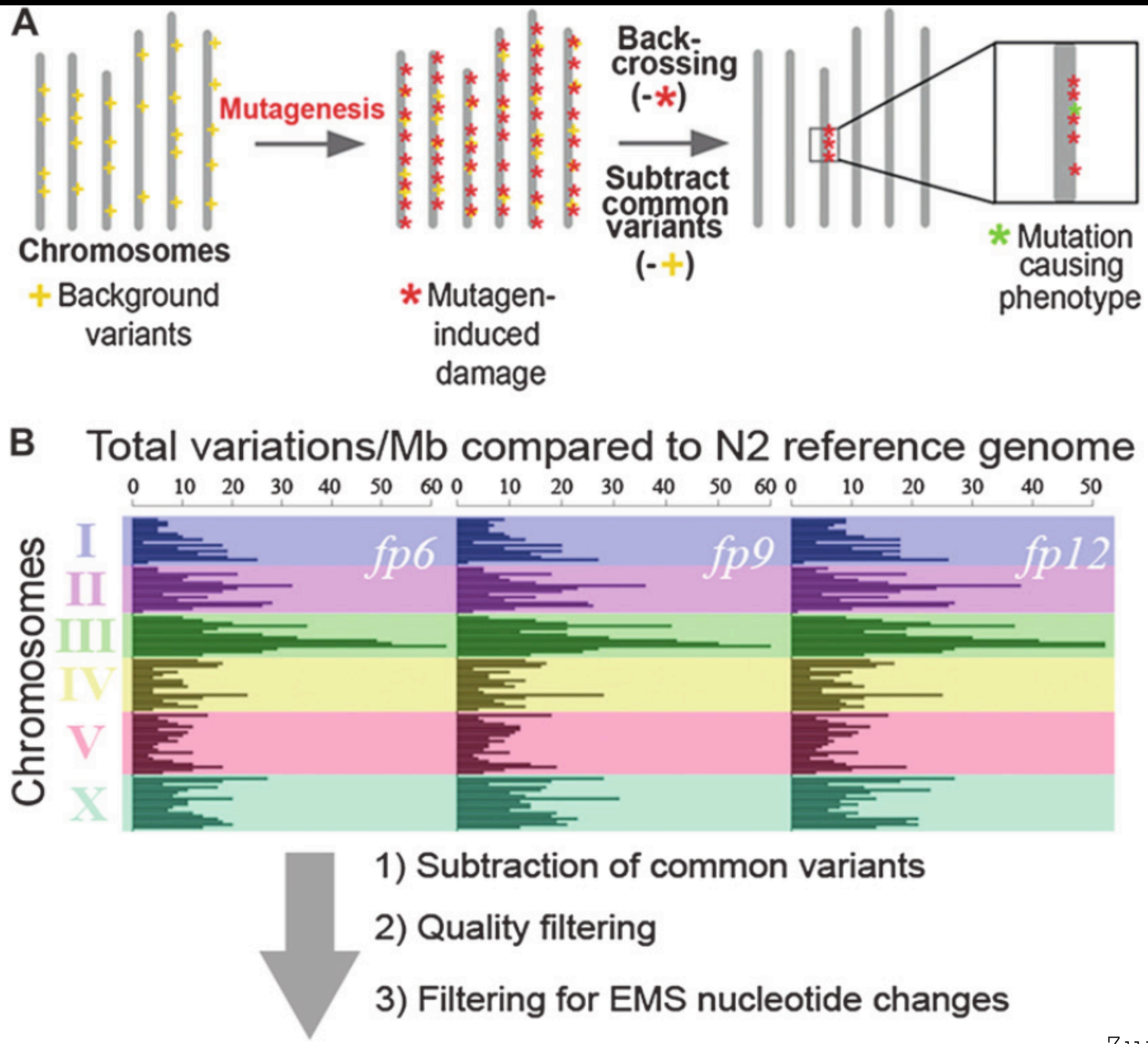


VDM is not as accurate as Hawaiian Mapping

Hawaiian mapping –> **~100,000** variants used for mapping

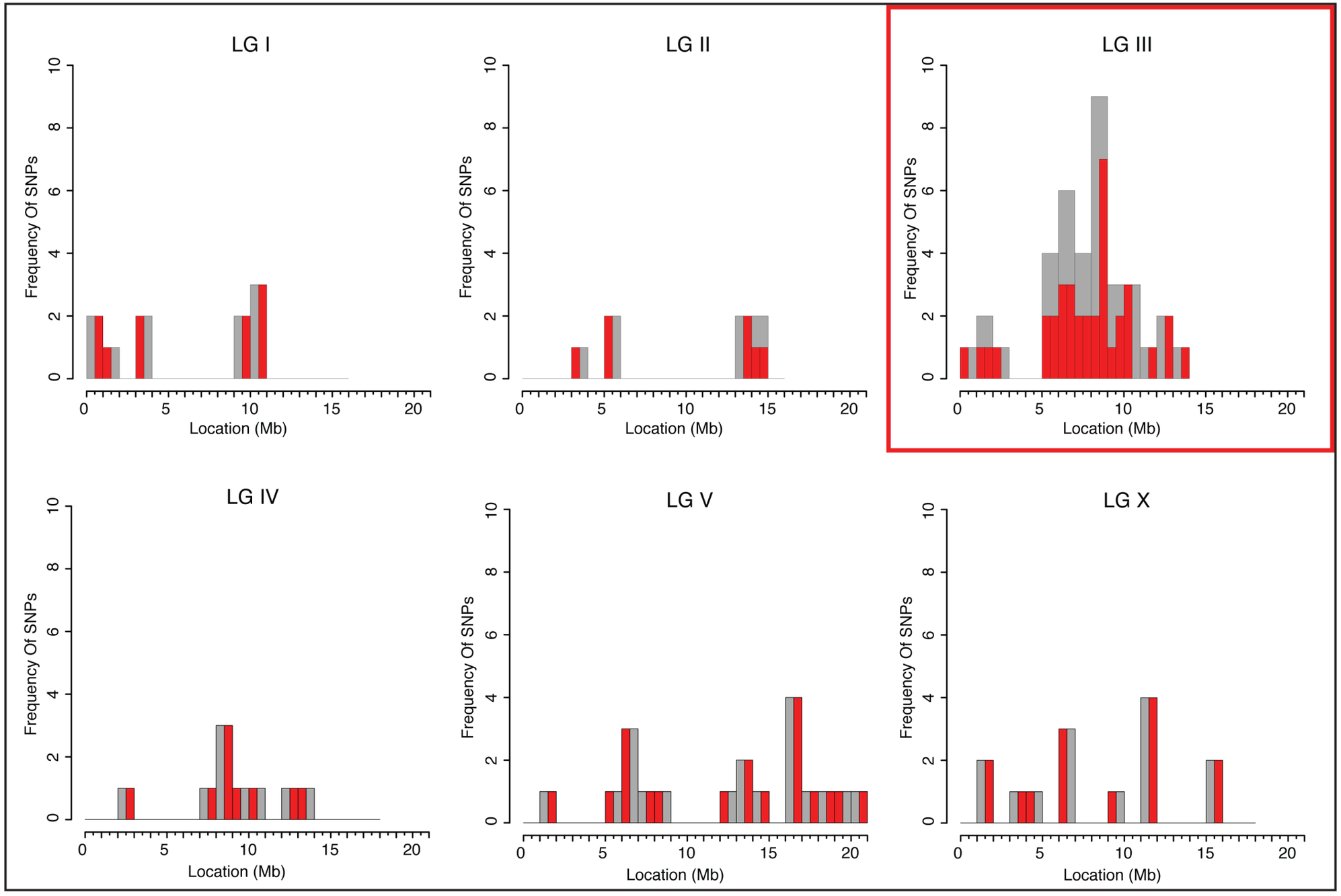
VDM –> **~1,000** variants used for mapping

EMS Density Mapping



Zuryn et al., 2010

CloudMap: EMS Variant Density Mapping Tool



Advantages of VDM over EMS Density Mapping

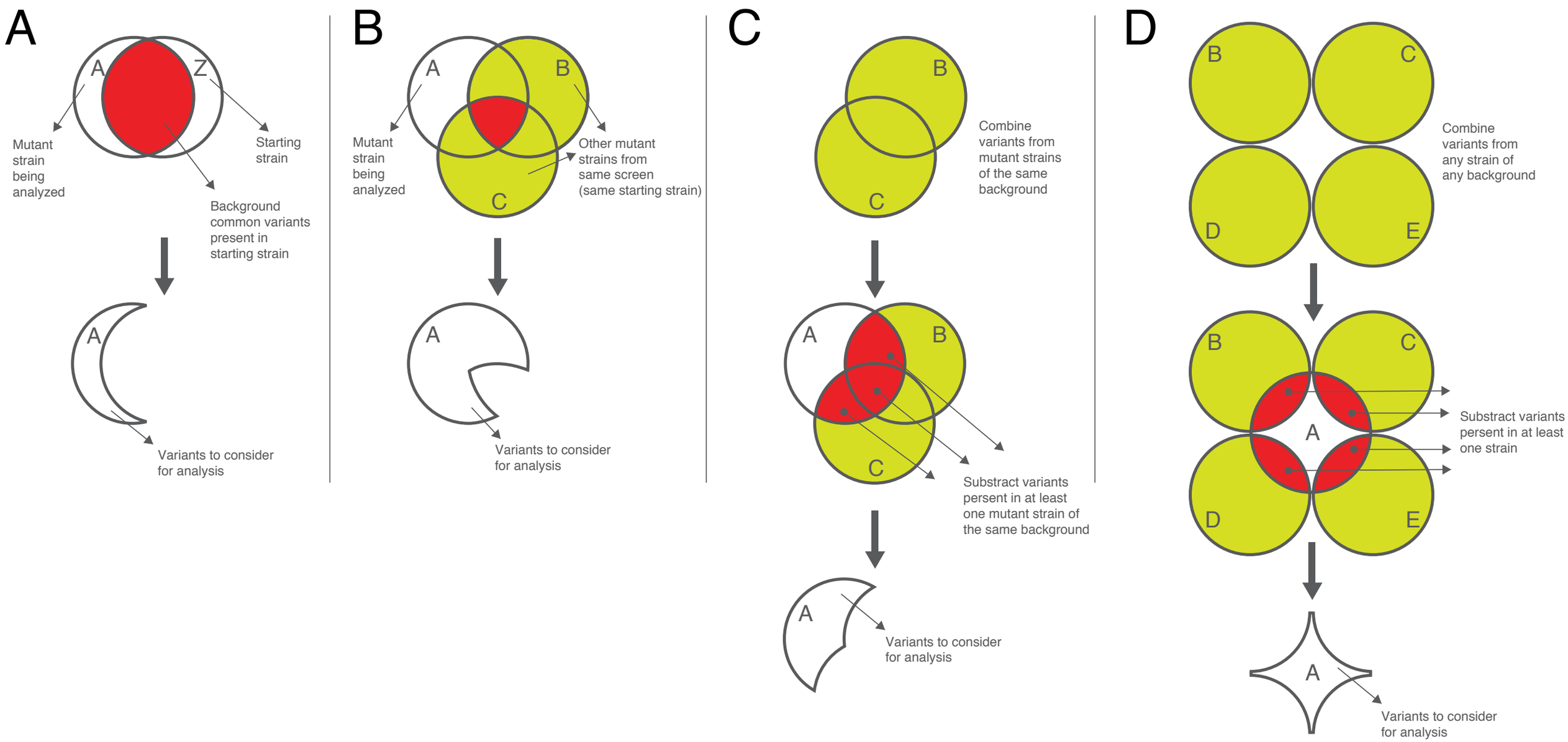
1. Bulk segregant approach gives **finer mapping**

   A. Pooling 50 F2's is conceptually equal to 50 serial backcrosses (same # crossover events that can be used for mapping)

2. **Faster** to pool 50 F2s derived from same cross (~6 days to pick F2s), than to serially backcross 50 times (50 x ~3 = ~150 days)

Variant subtraction and filtration (GATK)

Conservative —> Liberal



How can we analyze uncovered regions (putative deletions)?

- **known knowns** –> variants in your mapping region

How can we analyze uncovered regions (putative deletions)?

- **known knowns** —> variants in your mapping region

- **known unknowns** —> uncovered regions in mapping region (putative deletions)

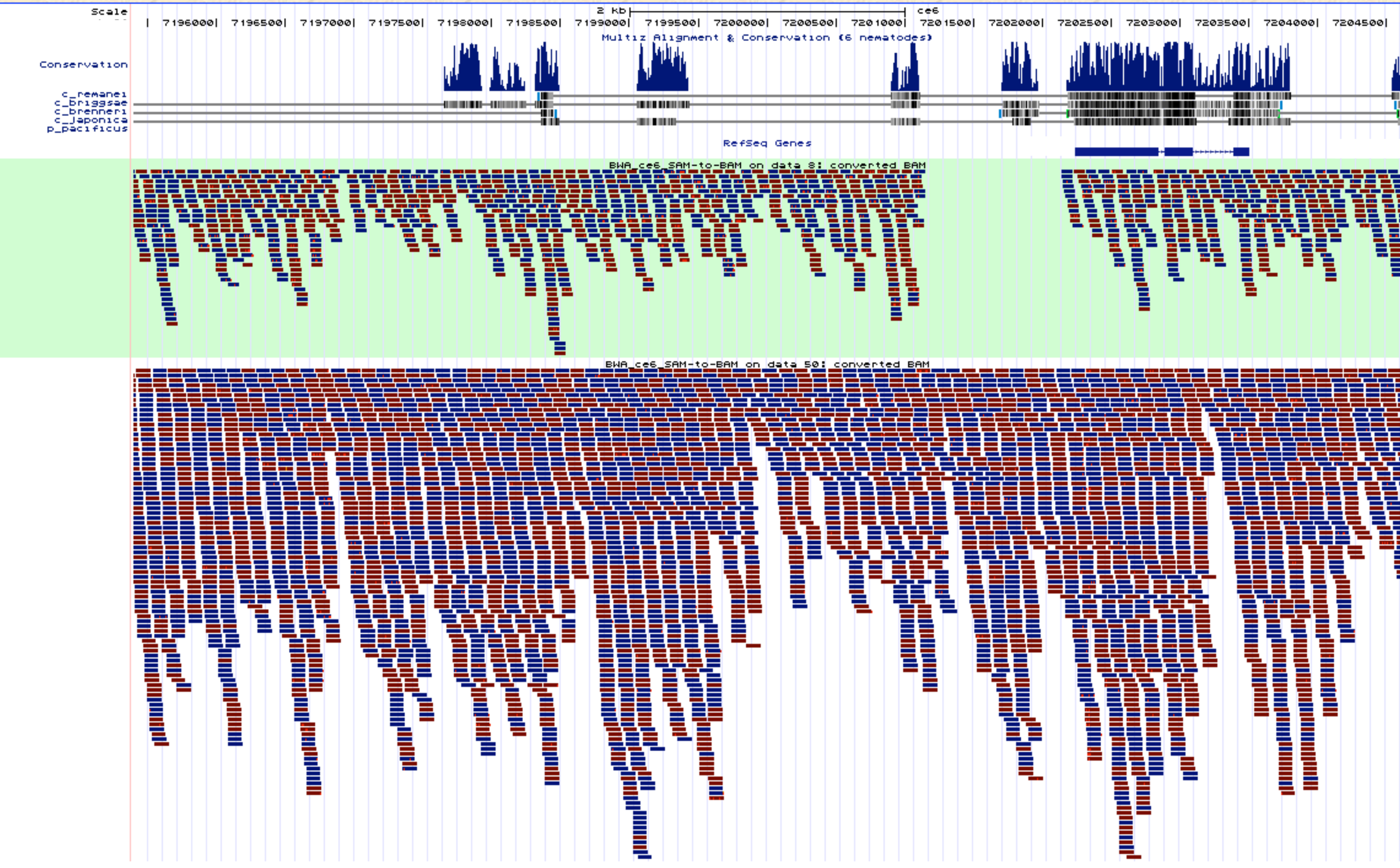
How can we analyze uncovered regions (putative deletions)?

- **known knowns** —> variants in your mapping region

- **known unknowns** —> uncovered regions in mapping region (putative deletions)

- **unknown unknowns** —> you're on your own

Find unique uncovered regions in your sample (putative deletions)



in silico complementation

- large genetic screens (especially suppressor screens) yield multiple alleles

in silico complementation

- large genetic screens (especially suppressor screens) yield multiple alleles
- complementation tests are time consuming & not always definitive
    - allelic complementation, non-allelic non-complementation, dominant alleles

in silico complementation

- large genetic screens (especially suppressor screens) yield multiple alleles
- complementation tests are time consuming & not always definitive
    - allelic complementation, non-allelic non-complementation, dominant alleles
- Just sequence many mutants from the same screen and see if you have multiple alleles of the same gene (if cost isn’t an issue)

in silico complementation

- large genetic screens (especially suppressor screens) yield multiple alleles

- complementation tests are time consuming & not always definitive

  - allelic complementation, non-allelic non-complementation, dominant alleles

- Just sequence many mutants from the same screen and see if you have multiple alleles of the same gene (if cost isn't an issue)

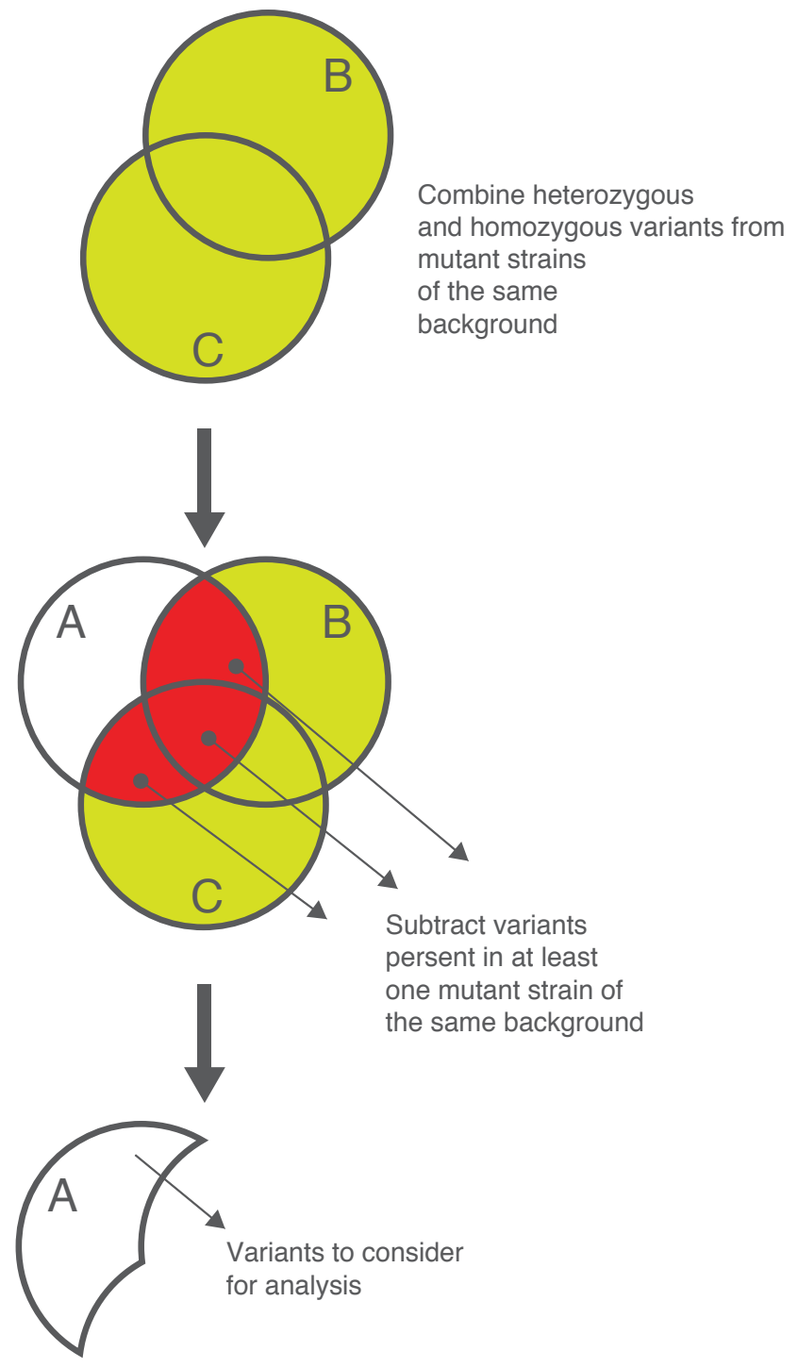
- Or, easily identify allelic variants in members of the same known complementation group

in silico complementation tool w/ variant subtration

1st pass – subtract background variants using a liberal subtraction strategy

- Tool returns results where allelic genetic loci contain non-identical hits in more than one sample

- Can control the upstream/downstream definition of a locus in bp



in silico complementation tool w/ variant subtration

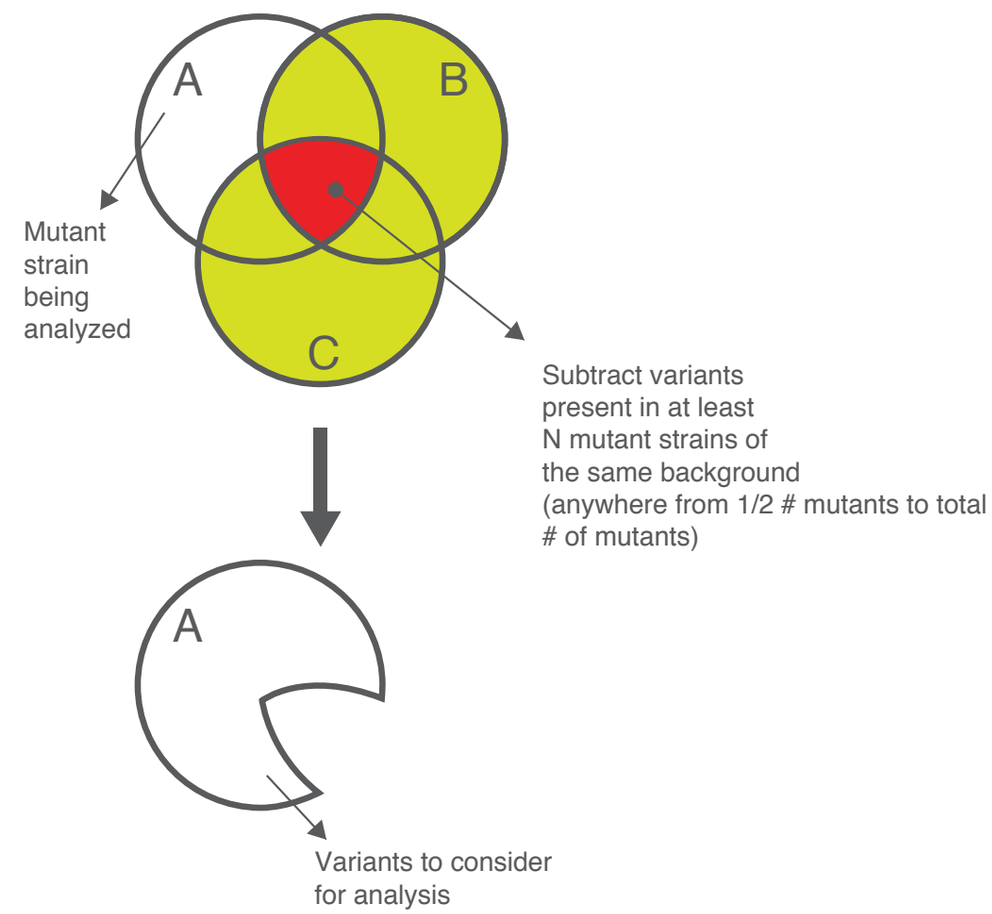
Caveat:

- Possible that 2 independent alleles of the same locus have the exact same variant – in which case the liberal subtraction would have subtracted that variant.

- Solution – subtract background variants using a conservative subtraction strategy

- Tool returns results where allelic genetic loci contain identical hits in more than one sample

- Downside – many non-phenotype causing variants will remain in the sample (background variants)



in silico complementation

Subtract background variants before running the tool

Tool returns instances where multiple alleles are present

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
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/Ne
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
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
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
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
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
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		

Variant calling & annotation (GATK & snpEff)

A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S	T	U
# Chromo	Position	Reference	Change	Change_t	Homozyg	Quality	Coverage	Gene_ID	Gene_nar	Bio_type	Trancript_	Exon_ID	Exon_Ran	Effect	old_AA/n	Old_codo	Codon_N	Codon_D	CDS_size
I	9879698	C	T	SNP	Het	27	62	T23H4.2	nhr-69	protein_codi	T23H4.2.2			INTRON					1122
I	9880489	C	T	SNP	Het	192	109	T23H4.2	nhr-69	protein_codi	T23H4.2.1			INTRON					1122
I	9880489	C	T	SNP	Het	192	109	T23H4.2	nhr-69	protein_codi	T23H4.2.2			INTRON					1122
I	9909039	G	A	SNP	Het	143	95	F52F12.6	ztf-11	protein_codi	F52F12.6			INTRON					1620
I	9909895	G	T	SNP	Het	162	68	F52F12.6	ztf-11	protein_codi	F52F12.6			INTRON					1620
I	9910274	*	#NAME?	INS	Het	4.42	102	F52F12.6	ztf-11	protein_codi	F52F12.6			INTRON					1620
I	9910270	*	#NAME?	INS	Het	4.42	103	F52F12.6	ztf-11	protein_codi	F52F12.6			INTRON					1620
I	9910274	*	#NAME?	INS	Het	217	103	F52F12.6	ztf-11	protein_codi	F52F12.6			INTRON					1620
I	9994773	A	G	SNP	Het	142	73	T23D8.8	cfi-1	protein_codi	T23D8.8	exon_I_9994	7	NON_SYNON	I/T	aTc/aCc	451	0	1404
I	10139575	C	G	SNP	Het	16.1	28							INTERGENIC					
I	10163982	A	G	SNP	Het	182	125	C25A1.2	fkh-10	protein_codi	C25A1.2.2	exon_I_1016	3	SYNONYMOU	H/H	caT/caC	112	1	585
I	10163982	A	G	SNP	Het	182	125	C25A1.2	fkh-10	protein_codi	C25A1.2.1	exon_I_1016	3	SYNONYMOU	H/H	caT/caC	112	1	585
I	10194812	G	A	SNP	Het	133	79	C25A1.11	aha-1	protein_codi	C25A1.11b			INTRON					1356
I	10194812	G	A	SNP	Het	133	79	C25A1.11	aha-1	protein_codi	C25A1.11a			INTRON					1362
I	10195403	A	T	SNP	Het	222	147	C25A1.11	aha-1	protein_codi	C25A1.11b	exon_I_1019	5	SYNONYMOU	R/R	cgT/cgA	212	3	1356
I	10195403	A	T	SNP	Het	222	147	C25A1.11	aha-1	protein_codi	C25A1.11a	exon_I_1019	5	SYNONYMOU	R/R	cgT/cgA	212	3	1362
I	10207646	A	C	SNP	Het	49	480							INTERGENIC					
I	10209783	A	G	SNP	Hom	27	45							INTERGENIC					
I	10209806	G	A	SNP	Het	3.55	16							INTERGENIC					
I	10248569	G	T	SNP	Het	141	100	ZC247.3	lin-11	protein_codi	ZC247.3			INTRON					1218
I	10248578	C	T	SNP	Het	39	92	ZC247.3	lin-11	protein_codi	ZC247.3			INTRON					1218
I	10251364	T	C	SNP	Het	113	115	ZC247.3	lin-11	protein_codi	ZC247.3			INTRON					1218
I	10251848	C	T	SNP	Het	135	102	ZC247.3	lin-11	protein_codi	ZC247.3	exon_I_1025	5	SYNONYMOU	S/S	tcC/tcT	202	3	1218
I	10383040	G	A	SNP	Het	40	99	F45H11.4	mgl-2	protein_codi	F45H11.4.2			UTR_3_PRIME: 590 bases from CDS					
I	10411503	A	G	SNP	Het	218	77	F25D7.3	blmp-1	protein_codi	F25D7.3b			INTRON					2406
I	10411503	A	G	SNP	Het	218	77	F25D7.3	blmp-1	protein_codi	F25D7.3a			INTRON					2454
I	10477650	C	T	SNP	Het	120	83	F37D6.2	F37D6.2	protein_codi	F37D6.2a.1			INTRON					1749

List of variants + Annotation information –> List of variant effects

SNPs & indels <=5bp annotated

For pooled samples, make sure you check the BAM alignment to determine if a called variant is real before proceeding

CloudMap variant annotation candidate checker

A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S	T	U	Y	Z
# Chromo	Position	Reference	Change	Change_t	Homozyg	Quality	Coverage	Gene_ID	Gene_nar	Bio_type	Trancript_	Exon_ID	Exon_Ran	Effect	old_AA/n	Old_codo	Codon_N	Codon_D	CDS_size	TFs	
I	9879698	C	T	SNP	Het	27	62	T23H4.2	nhr-69	protein_codi	T23H4.2.2			INTRON					1122	ZF - NHR	
I	9880489	C	T	SNP	Het	192	109	T23H4.2	nhr-69	protein_codi	T23H4.2.1			INTRON					1122	ZF - NHR	
I	9880489	C	T	SNP	Het	192	109	T23H4.2	nhr-69	protein_codi	T23H4.2.2			INTRON					1122	ZF - NHR	
I	9909039	G	A	SNP	Het	143	95	F52F12.6	ztf-11	protein_codi	F52F12.6			INTRON					1620	ZF - C2HC 2 fingers	
I	9909895	G	T	SNP	Het	162	68	F52F12.6	ztf-11	protein_codi	F52F12.6			INTRON					1620	ZF - C2HC 2 fingers	
I	9910274	*	#NAME?	INS	Het	4.42	102	F52F12.6	ztf-11	protein_codi	F52F12.6			INTRON					1620	ZF - C2HC 2 fingers	
I	9910270	*	#NAME?	INS	Het	4.42	103	F52F12.6	ztf-11	protein_codi	F52F12.6			INTRON					1620	ZF - C2HC 2 fingers	
I	9910274	*	#NAME?	INS	Het	217	103	F52F12.6	ztf-11	protein_codi	F52F12.6			INTRON					1620	ZF - C2HC 2 fingers	
I	9994773	A	G	SNP	Het	142	73	T23D8.8	cfi-1	protein_codi	T23D8.8	exon_I_9994	7	NON_SYNON	I/T	aTc/aCc	451	0	1404	ARID/BRIGHT	
I	10139575	C	G	SNP	Het	16.1	28							INTERGENIC						WH - Fork Head, AT Hook	
I	10163982	A	G	SNP	Het	182	125	C25A1.2	fkh-10	protein_codi	C25A1.2.2	exon_I_1016	3	SYNONYMOU	H/H	caT/caC	112	1	585	WH - Fork Head	
I	10163982	A	G	SNP	Het	182	125	C25A1.2	fkh-10	protein_codi	C25A1.2.1	exon_I_1016	3	SYNONYMOU	H/H	caT/caC	112	1	585	WH - Fork Head	
I	10194812	G	A	SNP	Het	133	79	C25A1.11	aha-1	protein_codi	C25A1.11b			INTRON					1356	bHLH	
I	10194812	G	A	SNP	Het	133	79	C25A1.11	aha-1	protein_codi	C25A1.11a			INTRON					1362	bHLH	
I	10195403	A	T	SNP	Het	222	147	C25A1.11	aha-1	protein_codi	C25A1.11b	exon_I_1019	5	SYNONYMOU	R/R	cgT/cgA	212	3	1356	bHLH	
I	10195403	A	T	SNP	Het	222	147	C25A1.11	aha-1	protein_codi	C25A1.11a	exon_I_1019	5	SYNONYMOU	R/R	cgT/cgA	212	3	1362	bHLH	
I	10207646	A	C	SNP	Het	49	480							INTERGENIC						WH - Fork Head, AT Hook	
I	10209783	A	G	SNP	Hom	27	45							INTERGENIC						WH - Fork Head, AT Hook	
I	10209806	G	A	SNP	Het	3.55	16							INTERGENIC						WH - Fork Head, AT Hook	
I	10248569	G	T	SNP	Het	141	100	ZC247.3	lin-11	protein_codi	ZC247.3			INTRON					1218	HD - LIM	
I	10248578	C	T	SNP	Het	39	92	ZC247.3	lin-11	protein_codi	ZC247.3			INTRON					1218	HD - LIM	
I	10251364	T	C	SNP	Het	113	115	ZC247.3	lin-11	protein_codi	ZC247.3			INTRON					1218	HD - LIM	
I	10251848	C	T	SNP	Het	135	102	ZC247.3	lin-11	protein_codi	ZC247.3	exon_I_1025	5	SYNONYMOU	S/S	tcC/tcT	202	3	1218	HD - LIM	
I	10383040	G	A	SNP	Het	40	99	F45H11.4	mgl-2	protein_codi	F45H11.4.2			UTR_3_PRIME: 590 bases from CDS						bZIP	
I	10411503	A	G	SNP	Het	218	77	F25D7.3	blmp-1	protein_codi	F25D7.3b			INTRON					2406	ZF - C2H2 - 4 fingers	
I	10411503	A	G	SNP	Het	218	77	F25D7.3	blmp-1	protein_codi	F25D7.3a			INTRON					2454	ZF - C2H2 - 4 fingers	
I	10477650	C	T	SNP	Het	120	83	F37D6.2	F37D6.2	protein_codi	F37D6.2a.1			INTRON					1749	ZF - C2H2 - 5 fingers	

1) Transcription factors

2) Transgene silencers

3) Genes expressed in the nervous system

4) Anything you want. . .

Summary of CloudMap strategies:

- Clone mutants from mapping crosses

1. Hawaiian Mapping
2. Variant Discovery Mapping
3. EMS Density Mapping

Summary of CloudMap strategies:

- Clone mutants from mapping crosses

  1. Hawaiian Mapping

  2. Variant Discovery Mapping

  3. EMS Density Mapping

- Find variants unique to your sample by subtracting variants in other strains

Summary of CloudMap strategies:

- Clone mutants from mapping crosses

  1. Hawaiian Mapping

  2. Variant Discovery Mapping

  3. EMS Density Mapping

- Find variants unique to your sample by subtracting variants in other strains

- Putative deletion analysis

Summary of CloudMap strategies:

- Clone mutants from mapping crosses

  1. Hawaiian Mapping

  2. Variant Discovery Mapping

  3. EMS Density Mapping

- Find variants unique to your sample by subtracting variants in other strains

- Putative deletion analysis

- in silico complementation testing

Summary of CloudMap strategies:

- Clone mutants from mapping crosses

  1. Hawaiian Mapping

  2. Variant Discovery Mapping

  3. EMS Density Mapping

- Find variants unique to your sample by subtracting variants in other strains

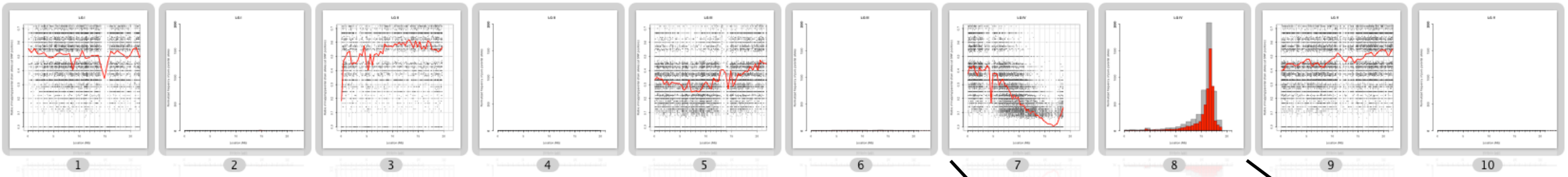
- Putative deletion analysis

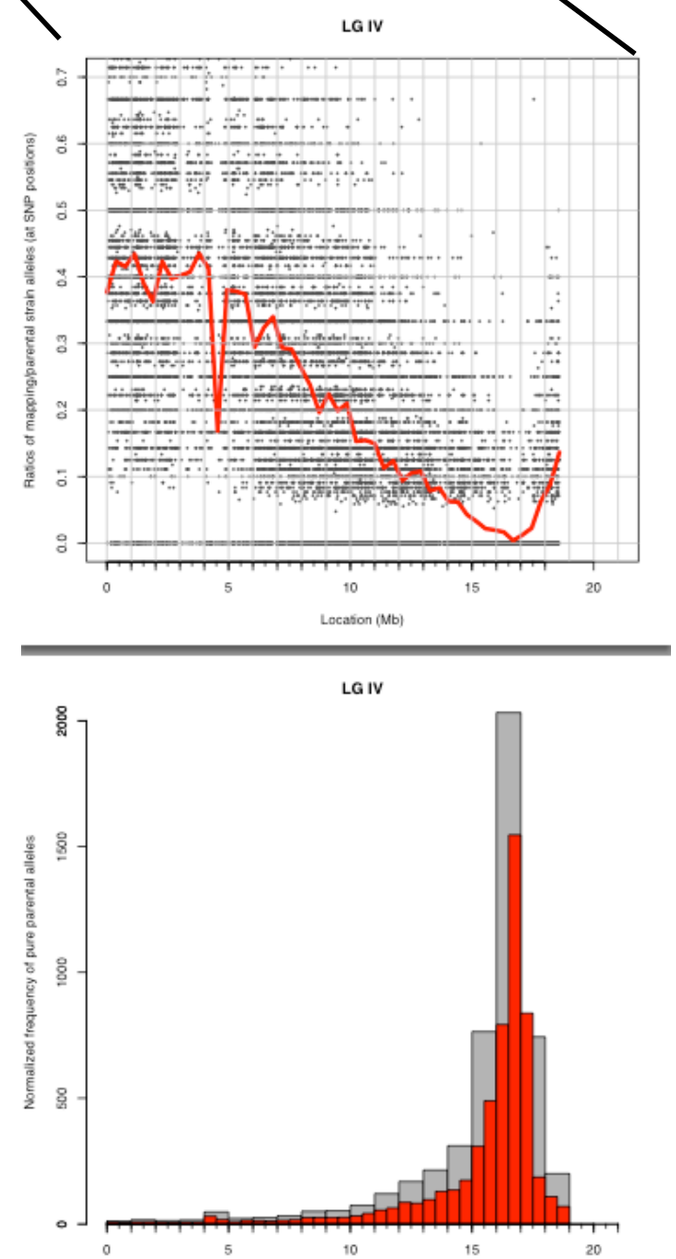
- in silico complementation testing

- Query candidate gene lists

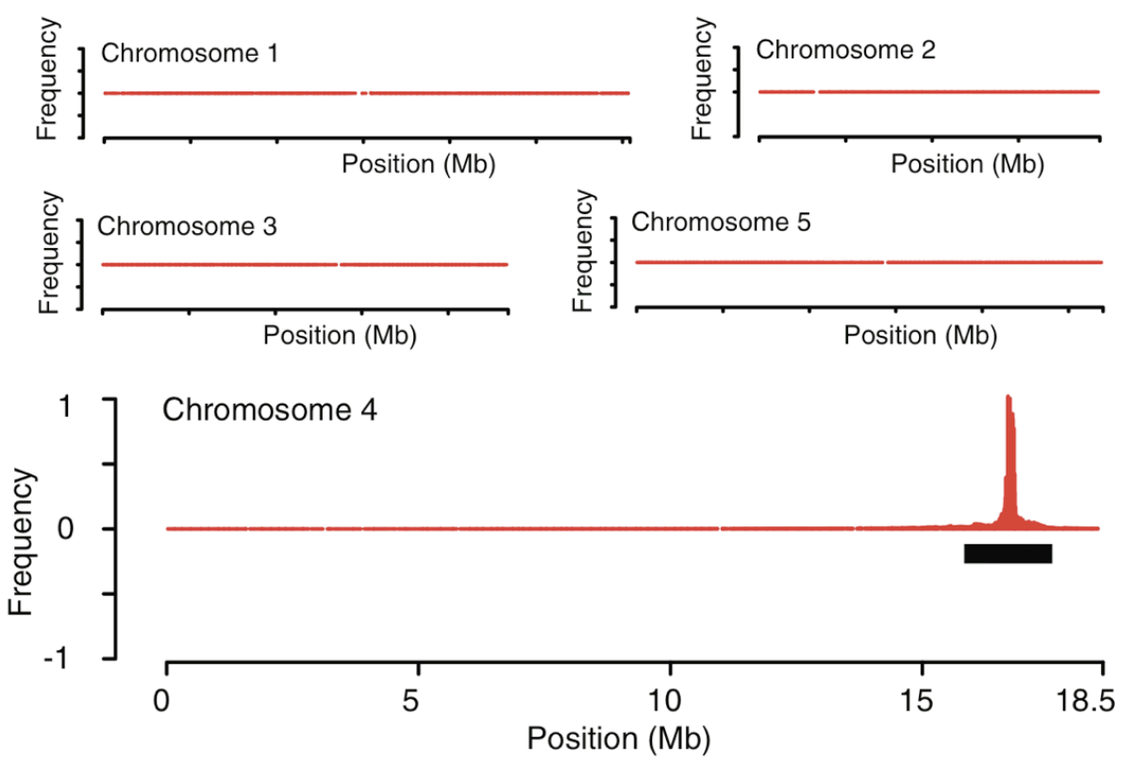
Tested with Arabidopsis, Brachypodium, zebrafish

Arabadopsis_AT4G35090gene_Galaxy39-[CloudMap_SNP_mapping_with_WGS].pdf





y-axis: relative parental allele frequency



3) Navigating within Galaxy

https://main.g2.bx.psu.edu/root

For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User | Using 86%

Tools

Phenotype Association
EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Variant Detection

- CloudMap: in silico complementation Perform in silico complementation analysis on multiple tabular snpEff output files
- CloudMap: Variant Discovery Mapping with WGS data Map a mutation using in silico bulk segregant linkage analysis using variants that are already present in the mutant strain of interest (rather than those introduced by a cross to a polymorphic strain).
- FreeBayes - Bayesian genetic variant detector
- CloudMap: Hawaiian Variant Mapping with WGS data Map a mutation by plotting recombination frequencies resulting from crossing to a highly polymorphic strain
- CloudMap: EMS Variant Density Mapping Map a mutation by linkage to regions of high mutation density using WGS data

NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
NGS: Picard (beta)
BEDTools
snpEff

RGENETICS

SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models

Workflows

- All workflows

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

Please select the species:
C. elegans

WGS Mutant VCF File:
43: Heterozygous and ..ubtraction)
WGS Mutant VCF file from pooled F2 mutants that have been crossed to a mapping strain. The VCF should contain data from only mapping strain (e.g. Hawaiian) SNP positions

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

Y-axis upper limit for scatter plot:
0.7

Y-axis upper limit for frequency plot:
500

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

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

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

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

Execute

What it does:

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

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

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

Sample CloudMap output for a linked chromosome:

History

CloudMap_ot266_Proof_of_Principle (with hidden data)
12.6 GB

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

Automated analyses w/ CloudMap workflows

Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User | Using 86%

Tools

search tools

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Genome Diversity
Phenotype Association
EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Variant Detection
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
NGS: Picard (beta)
BEDTools

Successfully ran workflow "imported: CloudMap Variant Discovery Mapping (Subtracts Crossing Strain from Input List of Variants run w/ GATK Unified Genotyper default settings)". The following datasets have been added to the queue:

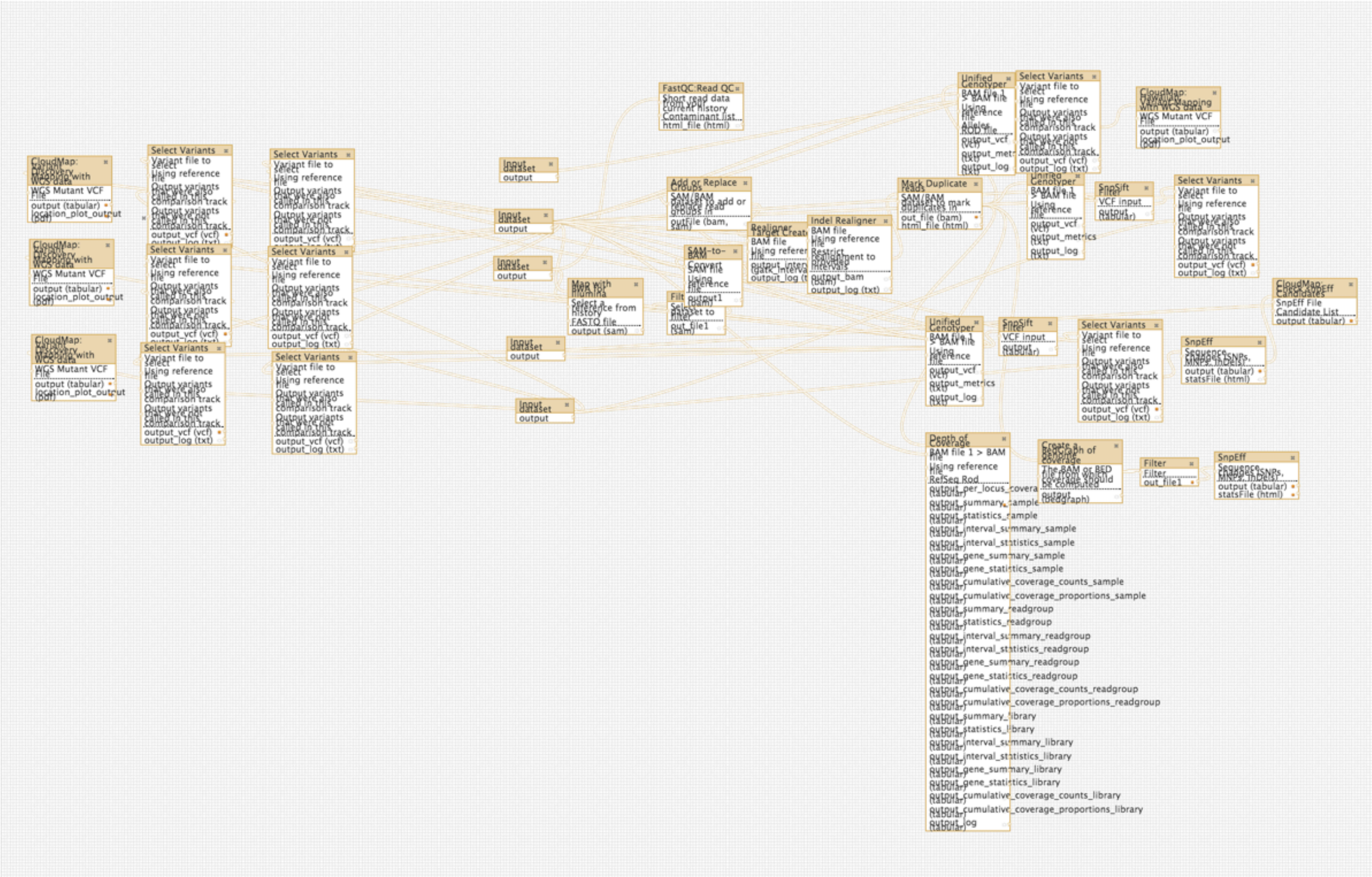
1: WS220.64_chr.fa
3: hu97HA_Heterozygous and Homozygous variants (lower quality, for cloning)
2: HA_SNPS_Unfiltered_112061Variants_WS220.64_chr.vcf
22: Mutant Strain SNPs QUAL100 VCF
23: Select Variants on data 1 and data 3 (log)
24: Mutant Strain SNPs QUAL200 VCF
25: Select Variants on data 1 and data 3 (log)
26: Mutant Strain SNPs QUAL300 VCF
27: Select Variants on data 1 and data 3 (log)
28: Mutant Strain SNPs QUAL100 minus crossing strain SNPs VCF
29: Select Variants on data 1, data 22, and data 2 (log)
30: Mutant Strain SNPs QUAL200 minus crossing strain SNPs VCF
31: Select Variants on data 1, data 24, and data 2 (log)
32: Mutant Strain SNPs QUAL300 minus crossing strain SNPs VCF
33: Select Variants on data 1, data 26, and data 2 (log)
34: Mutant Strain SNPs QUAL100 minus crossing+background strain VDM plot
35: CloudMap: Variant Discovery Mapping with WGS data on data 28
36: Mutant Strain SNPs QUAL200 minus crossing+background strain VDM plot
37: CloudMap: Variant Discovery Mapping with WGS data on data 30
38: Mutant Strain SNPs QUAL300 minus crossing+background strain VDM plot
39: CloudMap: Variant Discovery Mapping with WGS data on data 32

History

VDM_example
76.9 MB

39: CloudMap: Variant Discovery Mapping with WGS data on data 32
38: Mutant Strain SNPs QUAL300 minus crossing+background strain VDM plot
37: CloudMap: Variant Discovery Mapping with WGS data on data 30
36: Mutant Strain SNPs QUAL200 minus crossing+background strain VDM plot
35: CloudMap: Variant Discovery Mapping with WGS data on data 28
34: Mutant Strain SNPs QUAL100 minus crossing+background strain VDM plot
33: Select Variants on data 1, data 26, and data 2 (log)
32: Mutant Strain SNPs QUAL300 minus crossing strain SNPs VCF
31: Select Variants on data 1, data 24, and data 2 (log)
30: Mutant Strain SNPs QUAL200 minus crossing strain SNPs VCF
29: Select Variants on data 1, data 22, and data 2 (log)
28: Mutant Strain SNPs QUAL100 minus crossing strain SNPs VCF
27: Select Variants on data 1 and data 3 (log)
26: Mutant Strain SNPs QUAL300 VCF
25: Select Variants on data 1 and data 3 (log)
24: Mutant Strain SNPs QUAL200 VCF
23: Select Variants on data 1 and data 3 (log)
22: Mutant Strain SNPs QUAL100 VCF
3: hu97HA_Heterozygous and Homozygous variants (lower quality, for cloning)
2: HA_SNPS_Unfiltered_112061Variants_WS220.64_chr.vcf
1: WS220.64_chr.fa

CloudMap workflows can be edited



CloudMap data libraries contain a proof of principle dataset and config files

Data Library "CloudMap"

Name	Message
CloudMap Candidate Gene Lists	For CloudMap Check snpEff Candidates tool
CloudMap_C.elegansGenesWithHumanOrthologs.txt	
CloudMap_ChromatinFactors.txt	
CloudMap_TranscriptionFactors_wTF2.2.txt	
CloudMap EMS Variant Density Mapping	Use this dataset to try out the CloudMap EMS Variant Density Mapping tool
Zuryn_et_al_2010_mutA(subtracted_mutD).vcf	
CloudMap ot266 proof of principle dataset	Use these files to run the CloudMap ot266 proof of principle example
Hawaiian SNP reference files filtered (WS220.64)	Filtered set of Hawaiian SNP variants (used by CloudMap SNP Mapping with WGS tool)
HA_SNPs_Filtered_103346Variants_WS220.vcf	
Hawaiian SNP reference files unfiltered (WS220.64)	Unfiltered set of Hawaiian SNP variants (used by CloudMap SNP Mapping with WGS tool)
HA_SNPS_Unfiltered_112061Variants_WS220.64_chr.vcf	
ot266_ProofOfPrinciple_Small.fastqsanger	None
WS220.64_chr.fa	
CloudMap user guides	Detailed guides for using the CloudMap pipeline
CloudMap_Userguide_11-28-2012_large.pdf	
CloudMap_Userguide_11-28-2012_small.pdf	
Hawaiian Variant Mapping with WGS Data Other Species Configuration Files	Use these files to run Hawaiian Variant Mapping tools with species other than C. elegans or Arabidopsis
A.thaliana_Hawaiian_Variant_Mapping_config.txt	
C.elegans_Hawaiian_Variant_Mapping_config.txt	
D.rerio_Hawaiian_Variant_Mapping_config.txt	
ot260 and ot263 BEDs for uncovered subtraction	Use these BED files for the CloudMap ot266 proof of principle for uncovered region subtraction
ot260_Uncovered_regions.bed	
ot263_Uncovered_regions.bed	
ot266_Uncovered_regions.bed	
ot260 and ot263 VCFs for variant subtraction	Use these VCF files for the CloudMap ot266 proof of principle variant subtraction

For selected datasets: Import to current history Go

FASTQ statistics (FASTQC tool)

FastQC Report

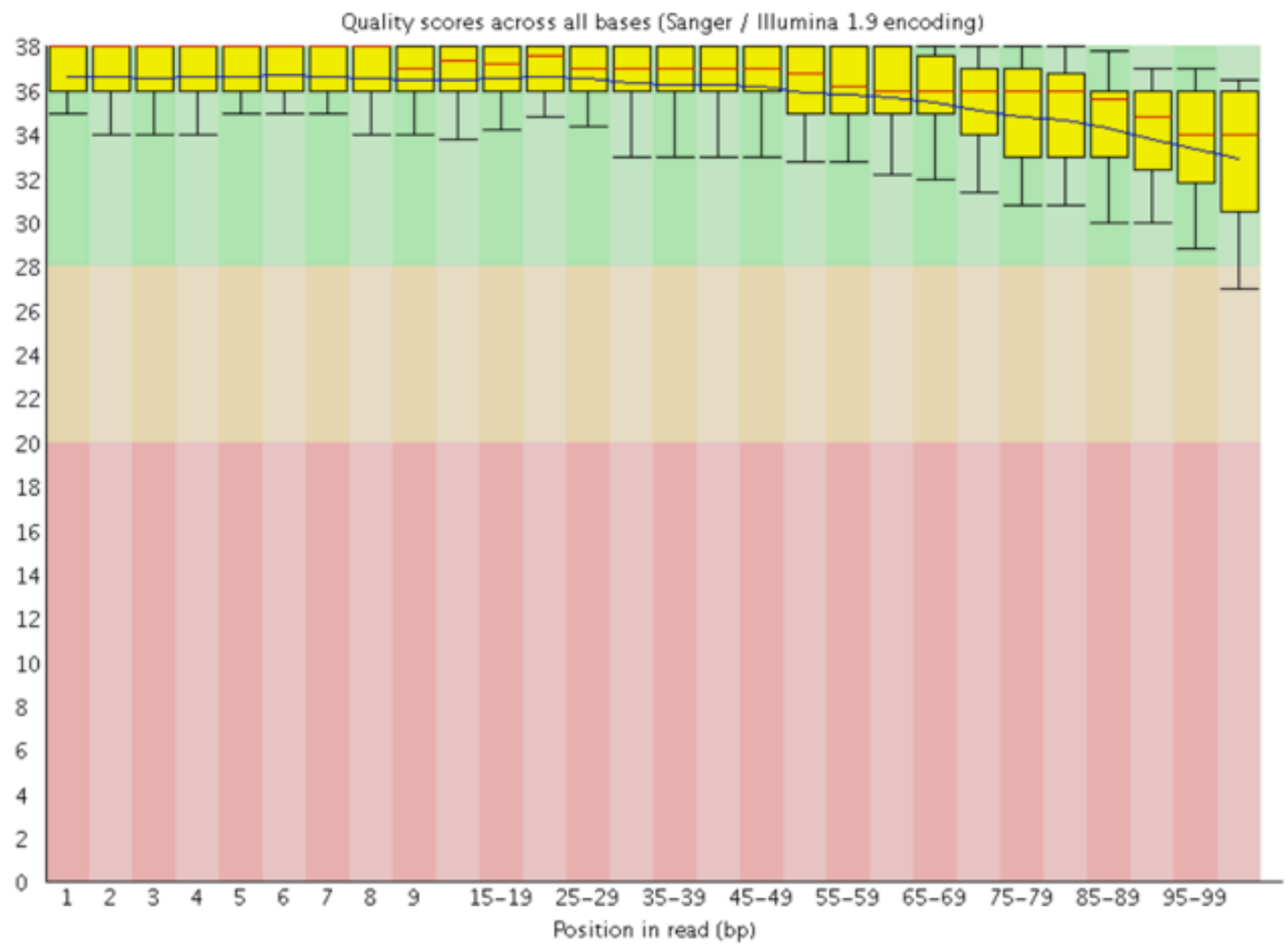
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

Measure	Value
Filename	ot266_ProofOfPrinciple_Small.fastqsanger
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10194621
Filtered Sequences	0
Sequence length	101
%GC	33

Per base sequence quality



Babraham Bioinformatics

FASTQ statistics (FASTQC tool)

FastQC Report

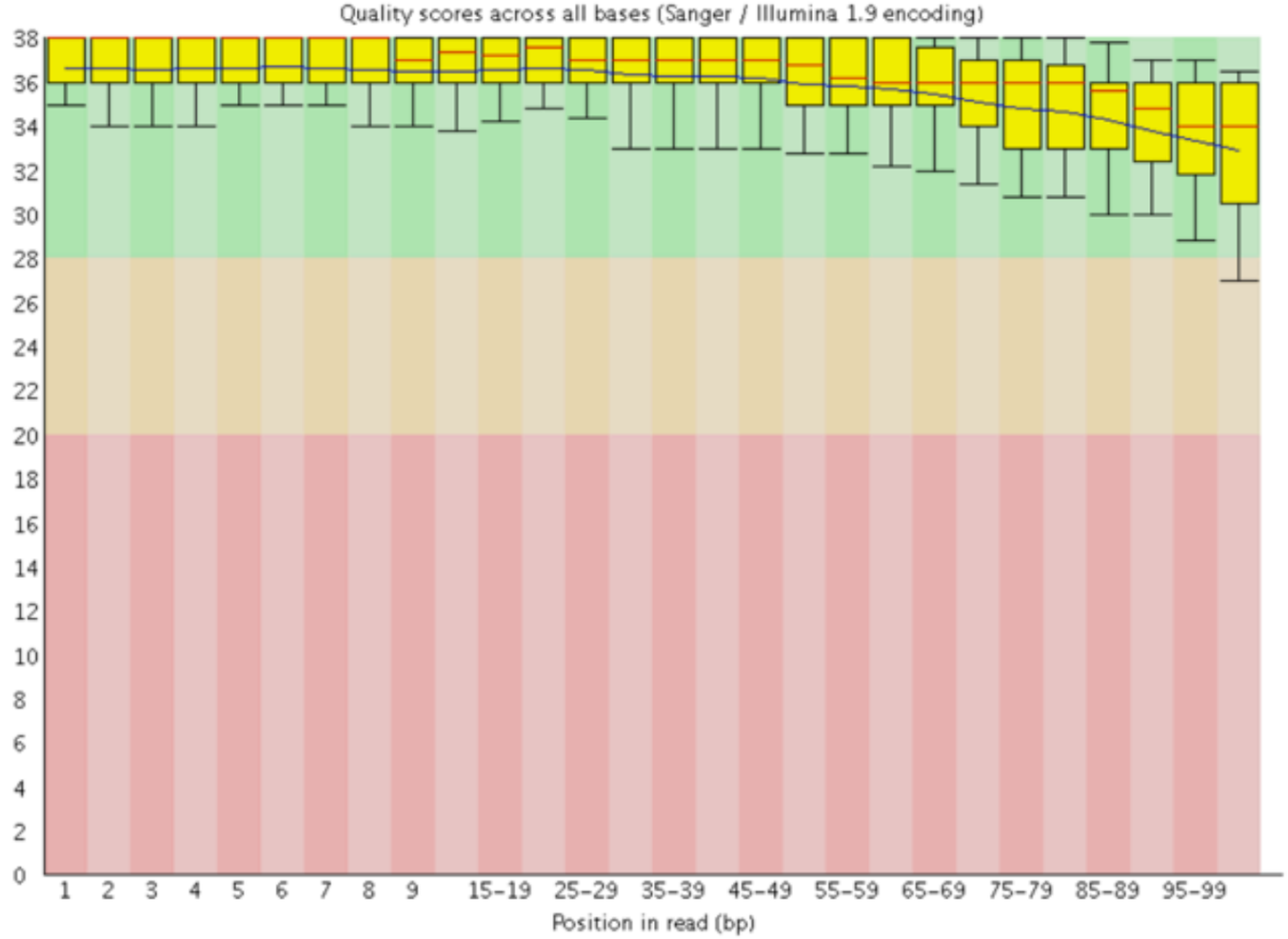
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

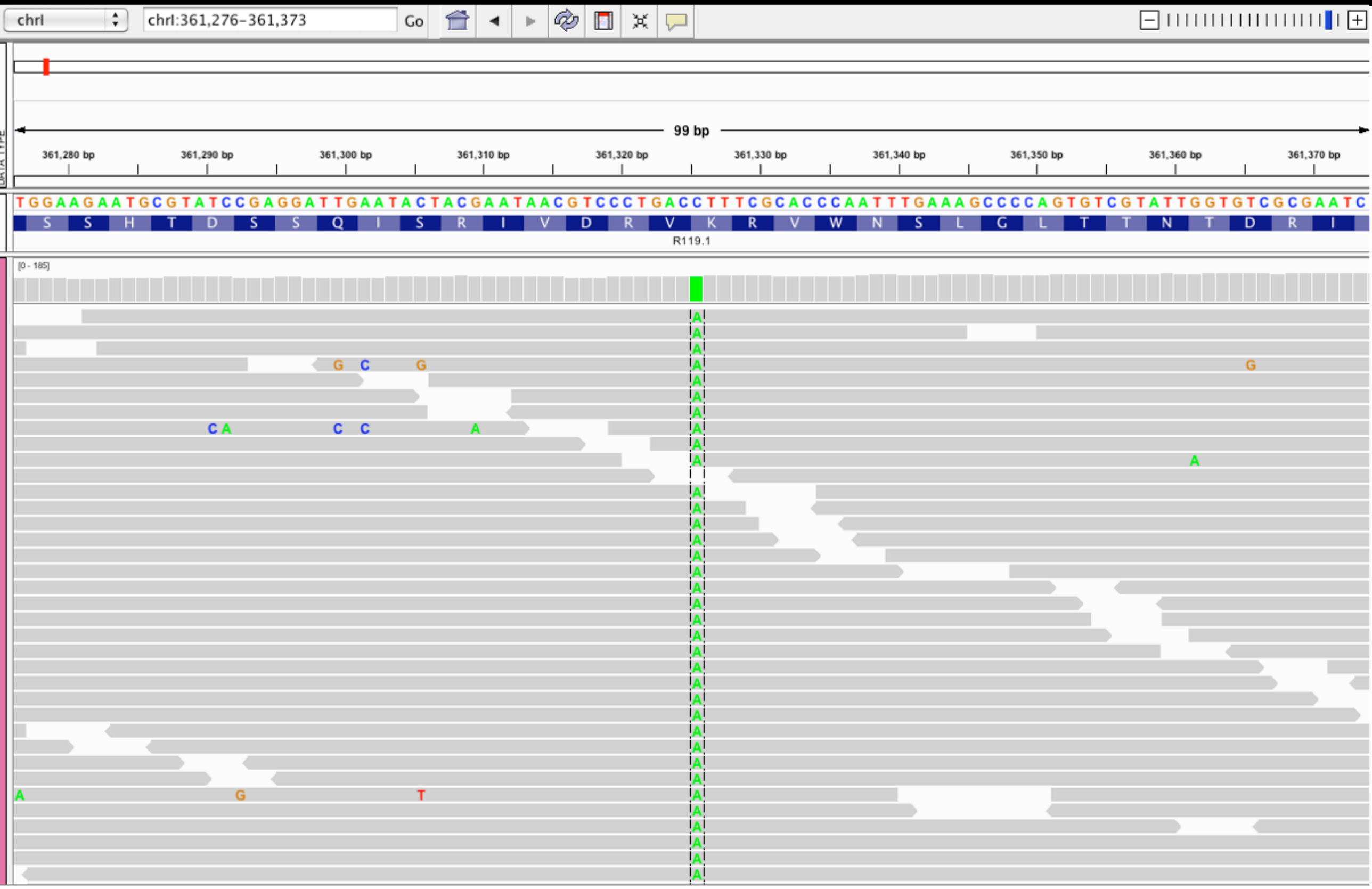
Measure	Value
Filename	ot266_ProofOfPrinciple_Small.fastqsanger
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10194621
Filtered Sequences	0
Sequence length	101
%GC	33

Per base sequence quality



Babraham Bioinformatics

Alignments (BWA, Bowtie, GATK, PICARD)



Integrated Genome Viewer

Variant annotation (snpEff)

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	245	0.774%
LOW	16,895	53.371%
MODERATE	501	1.583%
MODIFIER	14,015	44.273%

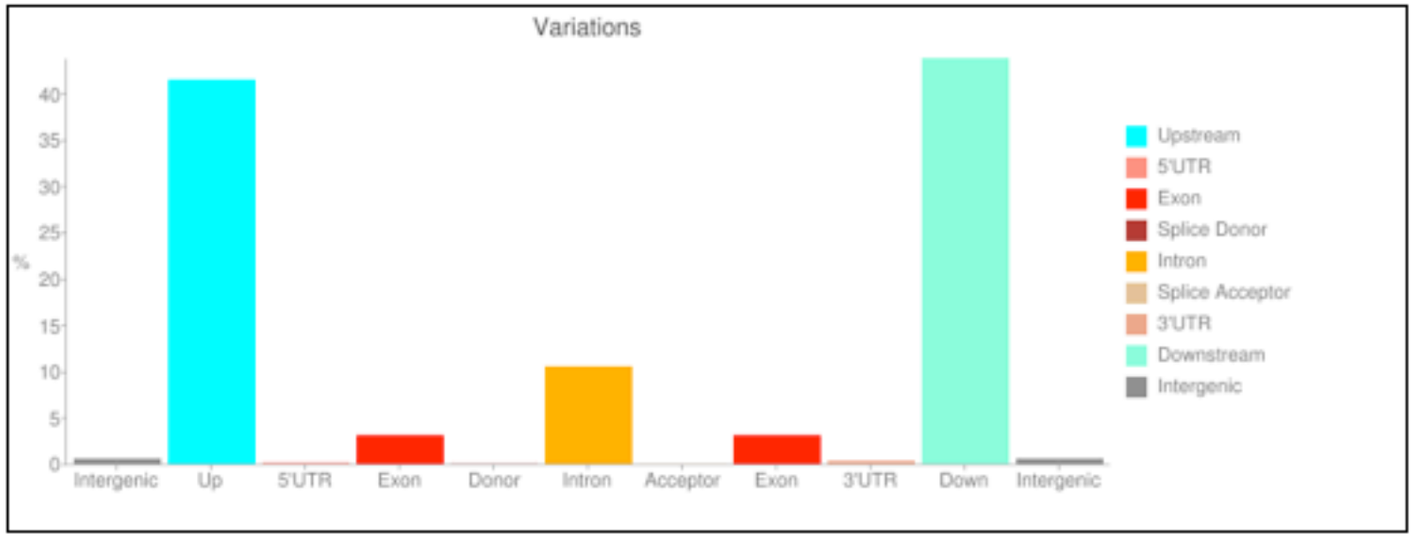
Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
CODON_CHANGE_PLUS_CODON_DELETION	1	0.003%
CODON_CHANGE_PLUS_CODON_INSERTION	1	0.003%
CODON_DELETION	2	0.006%
CODON_INSERTION	4	0.013%
DOWNSTREAM	13,868	43.808%
FRAME_SHIFT	211	0.667%
INTERGENIC	177	0.559%
INTRON	3,332	10.526%
NON_SYNONYMOUS_CODING	493	1.557%
SPLICE_SITE_ACCEPTOR	6	0.019%
SPLICE_SITE_DONOR	9	0.028%
STOP_GAINED	15	0.047%
STOP_LOST	4	0.013%
SYNONYMOUS_CODING	250	0.79%
SYNONYMOUS_START	1	0.003%
SYNONYMOUS_STOP	1	0.003%
UPSTREAM	13,134	41.49%
UTR_3_PRIME	104	0.329%
UTR_5_PRIME	43	0.136%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	13,868	43.808%
EXON	983	3.105%
INTERGENIC	177	0.559%
INTRON	3,332	10.526%
SPLICE_SITE_ACCEPTOR	6	0.019%
SPLICE_SITE_DONOR	9	0.028%
UPSTREAM	13,134	41.49%
UTR_3_PRIME	104	0.329%
UTR_5_PRIME	43	0.136%



Variant annotation (snpEff)

Base changes (SNPs)

	A	C	G	T
A	0	17	17	28
C	13	0	11	183
G	172	14	0	20
T	29	19	10	0

Ts/Tv (transitions / transversions)

Note: Only SNPs are used for this statistic.
Note: This Ts/Tv ratio is a 'raw' ratio. Some people prefer to use a ratio of rates, not observed events. In that case, you need to multiply by 2.0 (since there are twice as many possible transitions than transversions, E[Ts/Tv] ratio is twice the ratio of events).

Transitions	391
Transversions	142
Ts/Tv ratio	2.7535

All variants:

```
      Sample          : Total
   Transitions    : 391      391
   Transversions  : 142      142
Ts/Tv           : 2.754    2.754
```

Only known variants (i.e. the ones having a non-empty ID field):

```
No results available (empty input?)
```

Variant annotation (snpEff)

Amino acid changes

How to read this table:
- Rows are reference amino acids and columns are changed amino acids. E.g. Row 'A' column 'E' indicates how many 'A' amino acids have been replaced by 'E' amino acids.
- Red background colors indicate that more changes happened (heat-map).
- Diagonals are indicated using grey background color
- WARNING: This table may include different translation codon tables (e.g. mamalian DNA and mitochondrial DNA).

	*	-	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
*																						
-								1														
A			2												1			1	4	8		
C				2			1															2
D					1	1		2												3		
E		1				7					11											
F							3			1												
G					6	20		2									3			1		
H																						3
I										1												
K											5			1								
L							4					6										
M										6		1										
N														2								
P												9			2			13				
Q									1							12	2					
R				1						1	11						2					
S												2		1	1			2				
T										2					1				1			
V								1		1			2							2		
W	2																					
Y																						

4) Support (usegalaxy.org/cloudmap)

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User

Published Pages | gm2123 | CloudMap

CloudMap Materials:

User guides

Video user guide demonstrating the Hawaiian Variant Mapping workflow using the ot266 proof of principle dataset from the CloudMap paper:

https://vimeo.com/51082571

Note: In the interest of allowing users to quickly run a Hawaiian Variant Mapping example, the ot266 FASTQ sample dataset is a small subset of all the ot266 reads. For this reason, plots and variant lists generated by the example will not exactly match the ot266 figures in the CloudMap paper.

Video user guides demonstrating all workflows:

Coming soon. . .

PDF user guide:

Dataset 'CloudMap_Userguide_11-28-2012_large.pdf'

Dataset 'CloudMap_Userguide_11-28-2012_small.pdf'

Workflows

Hawaiian Variant Mapping workflow using the ot266 proof of principle dataset from the CloudMap paper (workflow can be used for any strain that has been crossed to a mapping strain e.g. Hawaiian):

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

Workflow 'CloudMap Hawaiian Variant Mapping with WGS and Variant Calling workflow (no candidate genes)'

Variant Discovery Mapping workflow:

Coming soon...

EMS Variant Density Mapping workflow (takes VCF of heterozygous and homozygous background variants to subtract):

Workflow 'CloudMap EMS Variant Density Mapping workflow (takes VCF of heterozygous and homozygous variants to subtract)'

Note: Unmapped mutant workflow can be used to generate the VCF of heterozygous and homozygous variants to subtract from the primary sample to be mapped using EMS variant density)

EMS Variant Density Mapping workflow (takes FASTQ reads from a second sample, creates VCF, and subtracts those variants from the primary sample):

Coming soon...

Unmapped mutant workflow (no variants from other strains to subtract):

Workflow 'CloudMap Unmapped Mutant workflow'

Unmapped mutant workflow (allows for subtraction of variants from other strains)

Workflow 'CloudMap Unmapped Mutant workflow (w/ subtraction of other strains)'

Uncovered Region Subtraction workflow (allows for subtraction of uncovered regions from other strains)

Workflow 'Cloudmap Uncovered Region Subtraction workflow'

Subtract variants workflow (1 set of candidates, 2 sets of variants to subtract)

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

Shared Histories

Shared history from the ot266 proof of principle dataset from the CloudMap paper (all the files generated from the workflow above):

History 'CloudMap_ot266_Proof_of_Principle (with hidden data)'

History 'CloudMap_ot266_Proof_of_Principle (with unhidden data)'

CloudMap Tools & Data

CloudMap tools can be downloaded from the Galaxy toolshed. They can be run in a local Galaxy install, or run as standalone Python scripts on a computer that has Python and R installed, or in Galaxy on Amazon's Elastic Compute (EC2) cloud service:

http://toolshed.g2.bx.psu.edu/

Shared data library (for use case examples from the paper and user guide and also contains key references files):

Go to http://usegalaxy.org/library and search for CloudMap

Alternative ways to run Galaxy & CloudMap

Galaxy and CloudMap on Amazon's Elastic Compute Cloud (EC2):

http://wiki.g2.bx.psu.edu/CloudMan

Running Galaxy locally:

4) Support (hobertlab.org/cloudmap)

Hobert Lab | People | Research | Publications | Methods & Protocols | Links | Location

[edit]

CloudMap

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

Minevich G, Park DS, Blankenberg D, Poole RJ, Hobert O.

Genetics. 2012 Dec;192(4):1249-69. doi: 10.1534/genetics.112.144204. Epub 2012 Oct 10.

Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical Institute, Columbia University Medical Center, New York, New York 10032.

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 ofCloudMap 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.

Video User Guides

Galaxy Install steps and CloudMap Dependencies

Frequently Asked Questions (FAQs):

CloudMap Questions:

How much coverage do I need for this to work?

Why does my annotated variant not correspond to the same position in Wormbase?

5) Hosting options

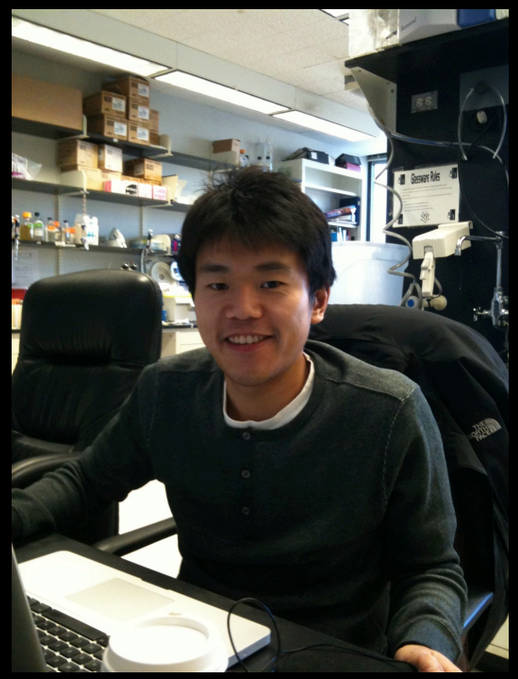
Install locally

Install on Amazon

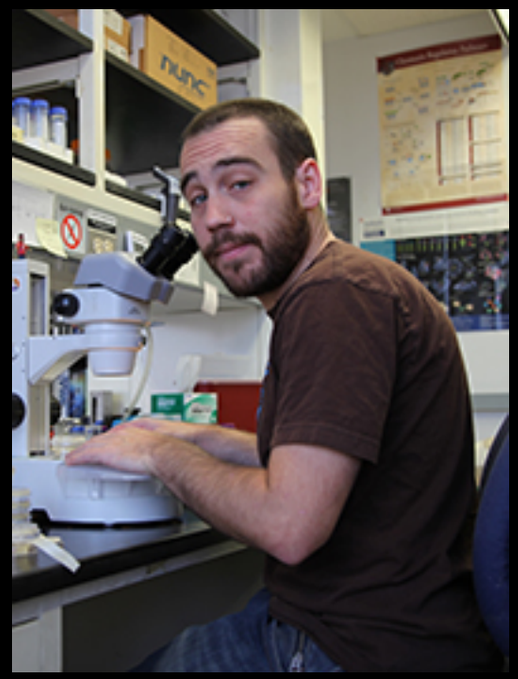
Instructions: hobertlab.org/cloudmap

Acknowledgements

Danny S. Park



Richard Poole



Daniel Blankenberg
(& Galaxy Team)

Oliver Hobert

http://usegalaxy.org/cloudmap

http://hobertlab.org/cloudmap

gm2123@columbia.edu – sign up to receive email updates