UAB DNA-Seq Analysis Workshop

John Osborne
Research Associate
Centers for Clinical and Translational Science
ozborn@uab.,edu

Thanks in advance

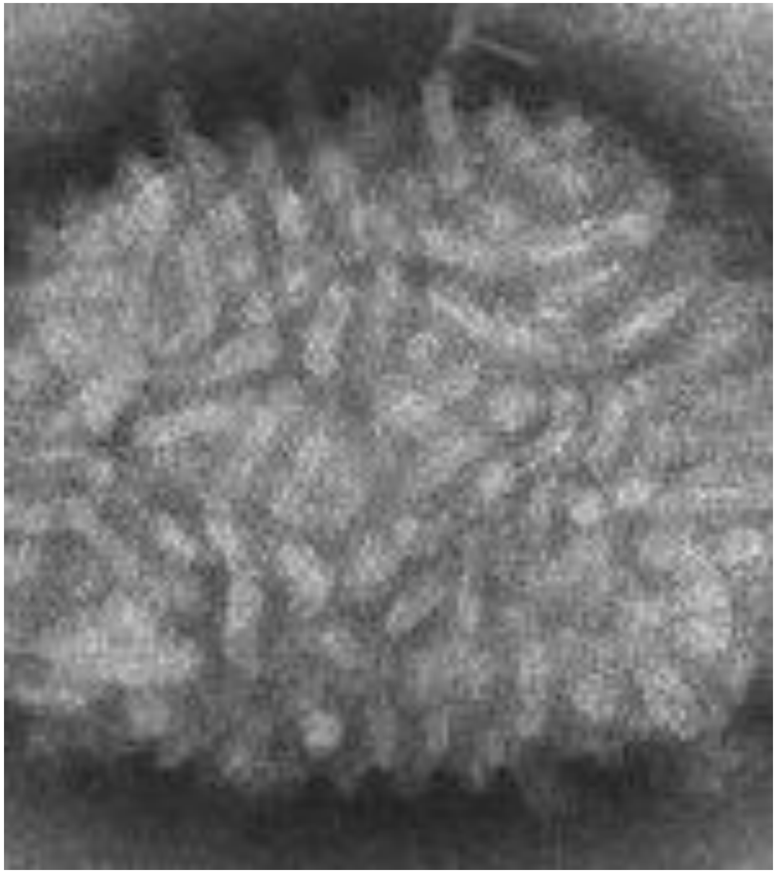
- You are the Guinea pigs for this workshop!
- At this point I hope you have:
  - Brought your laptop (Galaxy is pretty awkward with iPad)
  - Logged into Galaxy
  - Are eagerly awaiting some DNA-Seq data with a blank history pane
- Live demo
- Untimed

Format - Workshop

- Two important URLs
  - http://galaxy.uabgrid.uab.edu
  - http://docs.uabgrid.uab.edu/wiki/Galaxy_DNA-Seq_Tutorial
- Presentation more or less follows the Workshop Tutorial, but has extra details and less explicit instructions
- There is no pre-allocated time to helping people run the analysis so please interrupt me if something is unclear, wrong or just doesn't work
- Survey – how many people will be trying the DNA-Seq analysis?

Vaccinia Western Reserve (WR)

- Vaccinia strain Western Reserve is probably the most commonly used poxvirus strain for research
  - Vaccinia famous as the smallpox vaccine but has other uses as well
- Double stranded DNA virus with a AT rich genome approximately 195kb in length with inverted tandem repeats at the end of the genome



The Problem: Computational Biology Perspective

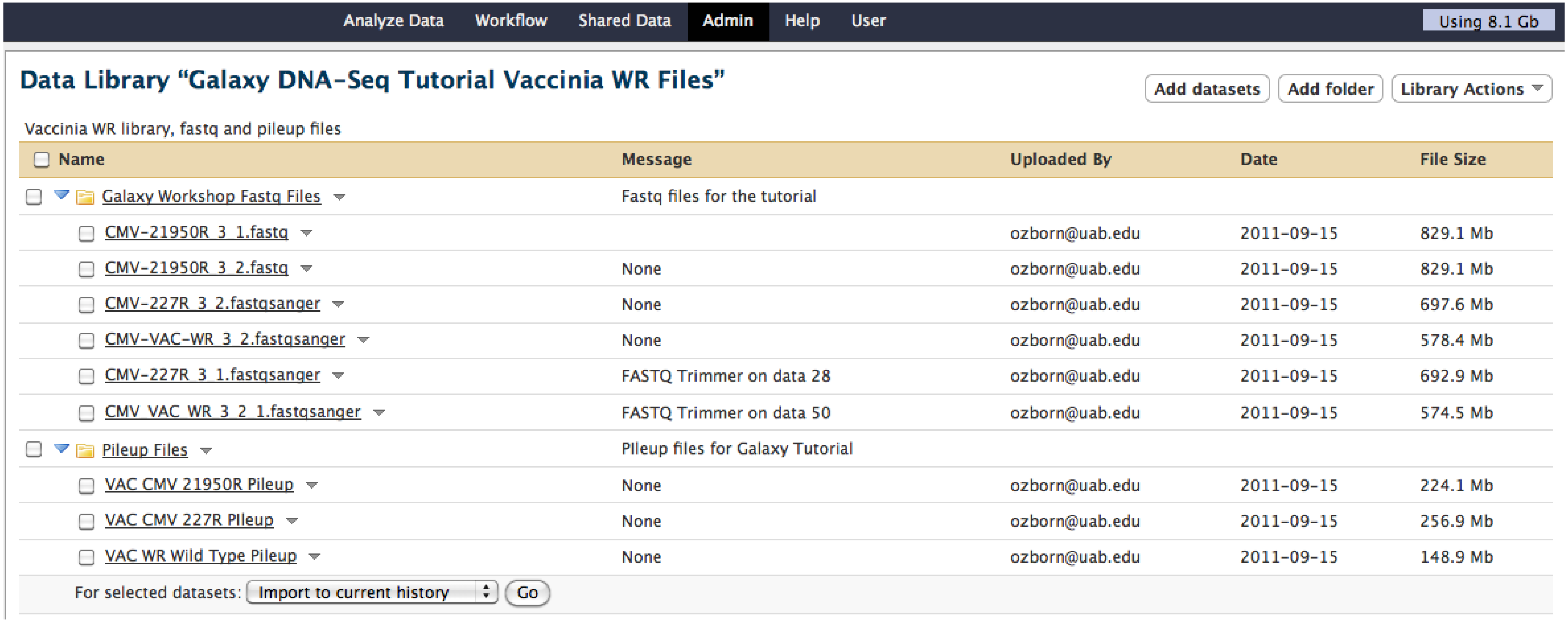
- 3 strains of Vaccinia were sequenced
- Parental strain VAC WR with wild type phenotype
- 2 mutant strains with different phenotypes
  - CMV 227R
  - CMV 21950R
- Illumunia paired end reads
- 6 Fastq files

- What are the genetic differences between these strains?
- Assumption is there is a genetic basis for the difference in phenotype
- Goal – List of potentially biologically relevant differences than can be followed up in the lab

Disclaimer

- Vaccinia selected for DNA-Seq because it is small enough to get results in a workshop and won't fill up the /lustre disk
- SNP analysis is different with mammals (particularly humans) where there are existing databases of SNP info (dbSNP being the most famous)
- GATK is probably more state of the art than Samtools for variant analysis and we use it here in our group but:
  - No available on Galaxy yet
  - Doesn't always make a difference
- Vaccinia is well designed for reference based sequencing, since there is a single source strain which has already sequenced
  - Faster and easier than de-novo sequencing which is not covered here
- There may be better ways to do this and I'm happy to hear about them
- Let's start!

Get the data



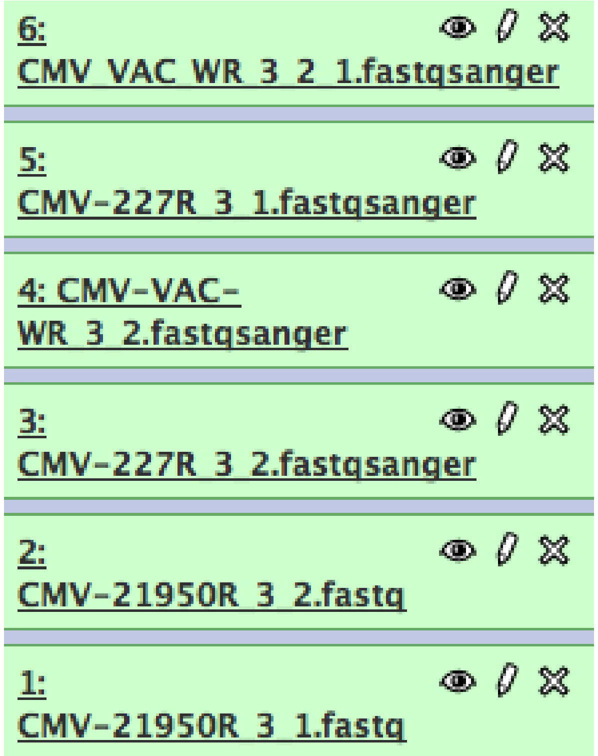
Data Library "Galaxy DNA-Seq Tutorial Vaccinia WR Files"

Vaccinia WR library, fastq and pileup files

□ Name	Message	Uploaded By	Date	File Size
□ Galaxy Workshop Fastq Files	Fastq files for the tutorial			
□ CMV-21950R 3 1.fastq		ozborn@uab.edu	2011-09-15	829.1 Mb
□ CMV-21950R 3 2.fastq	None	ozborn@uab.edu	2011-09-15	829.1 Mb
□ CMV-227R 3 2.fastqsanger	None	ozborn@uab.edu	2011-09-15	697.6 Mb
□ CMV-VAC-WR 3 2.fastqsanger	None	ozborn@uab.edu	2011-09-15	578.4 Mb
□ CMV-227R 3 1.fastqsanger	FASTQ Trimmer on data 28	ozborn@uab.edu	2011-09-15	692.9 Mb
□ CMV VAC WR 3 2 1.fastqsanger	FASTQ Trimmer on data 50	ozborn@uab.edu	2011-09-15	574.5 Mb
□ Pileup Files	PIleup files for Galaxy Tutorial			
□ VAC CMV 21950R Pileup	None	ozborn@uab.edu	2011-09-15	224.1 Mb
□ VAC CMV 227R PIleup	None	ozborn@uab.edu	2011-09-15	256.9 Mb
□ VAC WR Wild Type Pileup	None	ozborn@uab.edu	2011-09-15	148.9 Mb

For selected datasets: Import to current history Go

Put it in your history and make sure the details (Reference, format)are correct



6: CMV_VAC_WR_3_2_1.fastqsanger

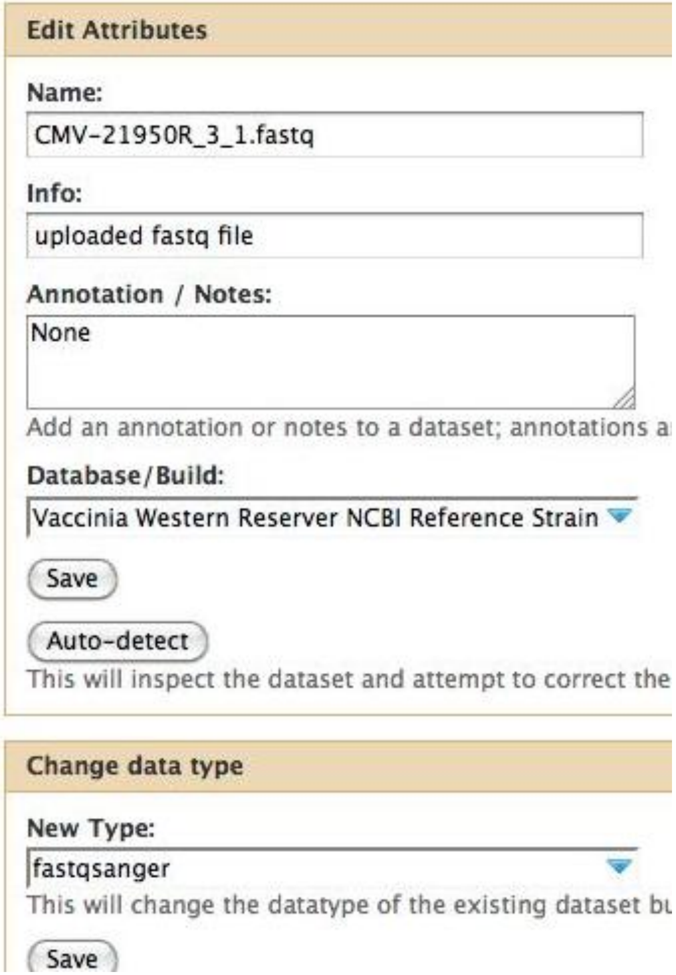
5: CMV-227R_3_1.fastqsanger

4: CMV-VAC-WR_3_2.fastqsanger

3: CMV-227R_3_2.fastqsanger

2: CMV-21950R_3_2.fastq

1: CMV-21950R_3_1.fastq



Edit Attributes

Name: CMV-21950R_3_1.fastq

Info: uploaded fastq file

Annotation / Notes: None

Add an annotation or notes to a dataset; annotations a

Database/Build: Vaccinia Western Reserver NCBI Reference Strain

This will inspect the dataset and attempt to correct the

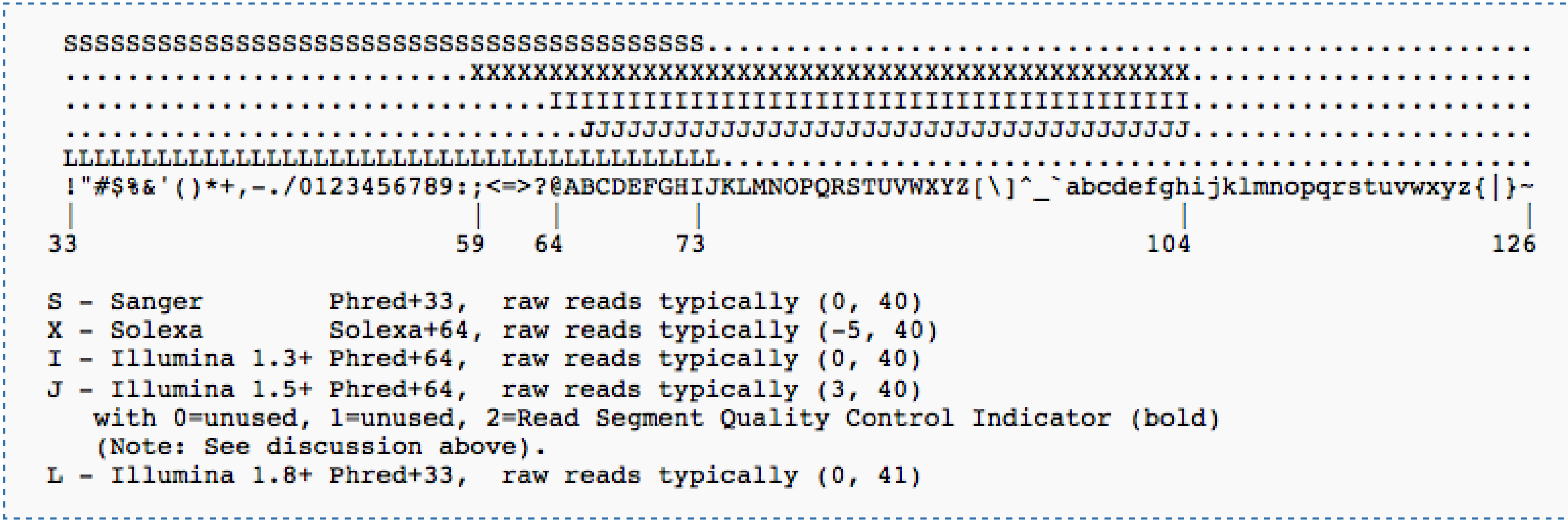
Change data type

New Type: fastqsanger

This will change the datatype of the existing dataset bu

Over 4GB of data for this this virus already!

FastQ Format



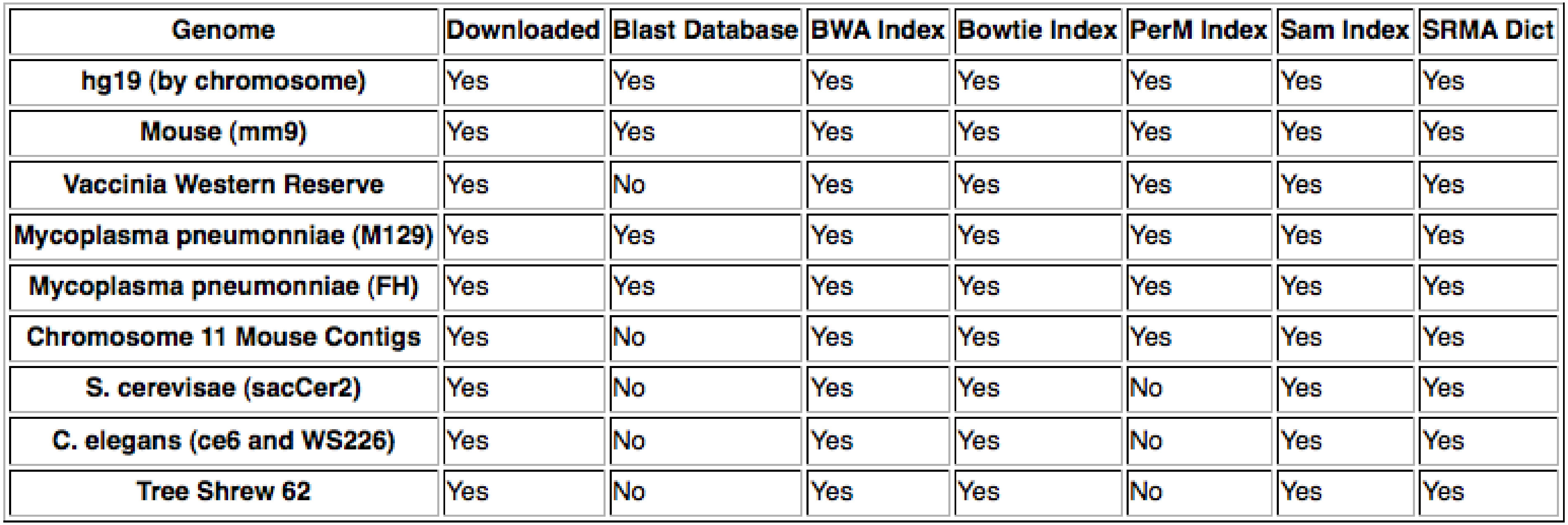
```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

- Highly recommend FastQ Groomer when dealing with new data
- Costly (time and space)

Available Reference Genomes

http://docs.uabgrid.uab.edu/wiki/Galaxy#Available_datasets

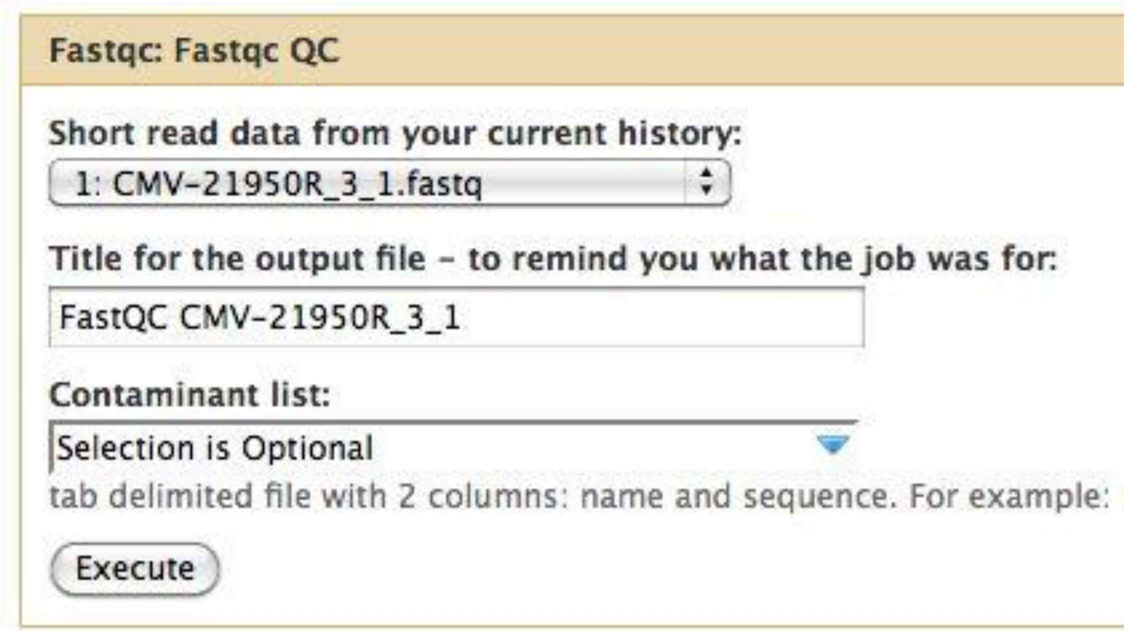
If you need a genome added or indexed, contact me. Set up for bwa, bowtie (tophat, cuffdiff), samtools, picard tools well scripted. Others less so



Genome	Downloaded	Blast Database	BWA Index	Bowtie Index	PerM Index	Sam Index	SRMA Dict
hg19 (by chromosome)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mouse (mm9)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Vaccinia Western Reserve	Yes	No	Yes	Yes	Yes	Yes	Yes
Mycoplasma pneumonniae (M129)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mycoplasma pneumonniae (FH)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Chromosome 11 Mouse Contigs	Yes	No	Yes	Yes	Yes	Yes	Yes
S. cerevisae (sacCer2)	Yes	No	Yes	Yes	No	Yes	Yes
C. elegans (ce6 and WS226)	Yes	No	Yes	Yes	No	Yes	Yes
Tree Shrew 62	Yes	No	Yes	Yes	No	Yes	Yes

FastQC

- Make sure reference genomes are correct including CMV-21950R_3_1.fastq
- Run FastQC
  - Select NGS: QC and manipulation -> FastQC
- Queue all 6 up, we do this in parallel!



Fastqc: Fastqc QC

Short read data from your current history:

1: CMV-21950R_3_1.fastq

Title for the output file - to remind you what the job was for:

FastQC CMV-21950R_3_1

Contaminant list:

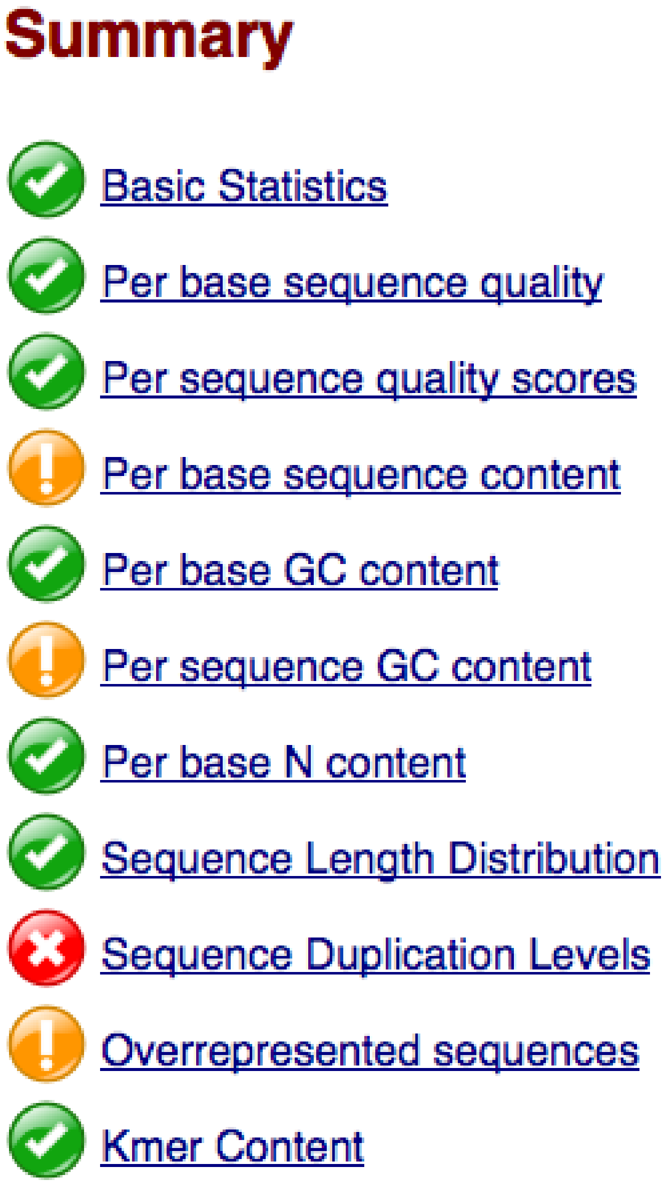
Selection is Optional

tab delimited file with 2 columns: name and sequence. For example:

FastQC Summary

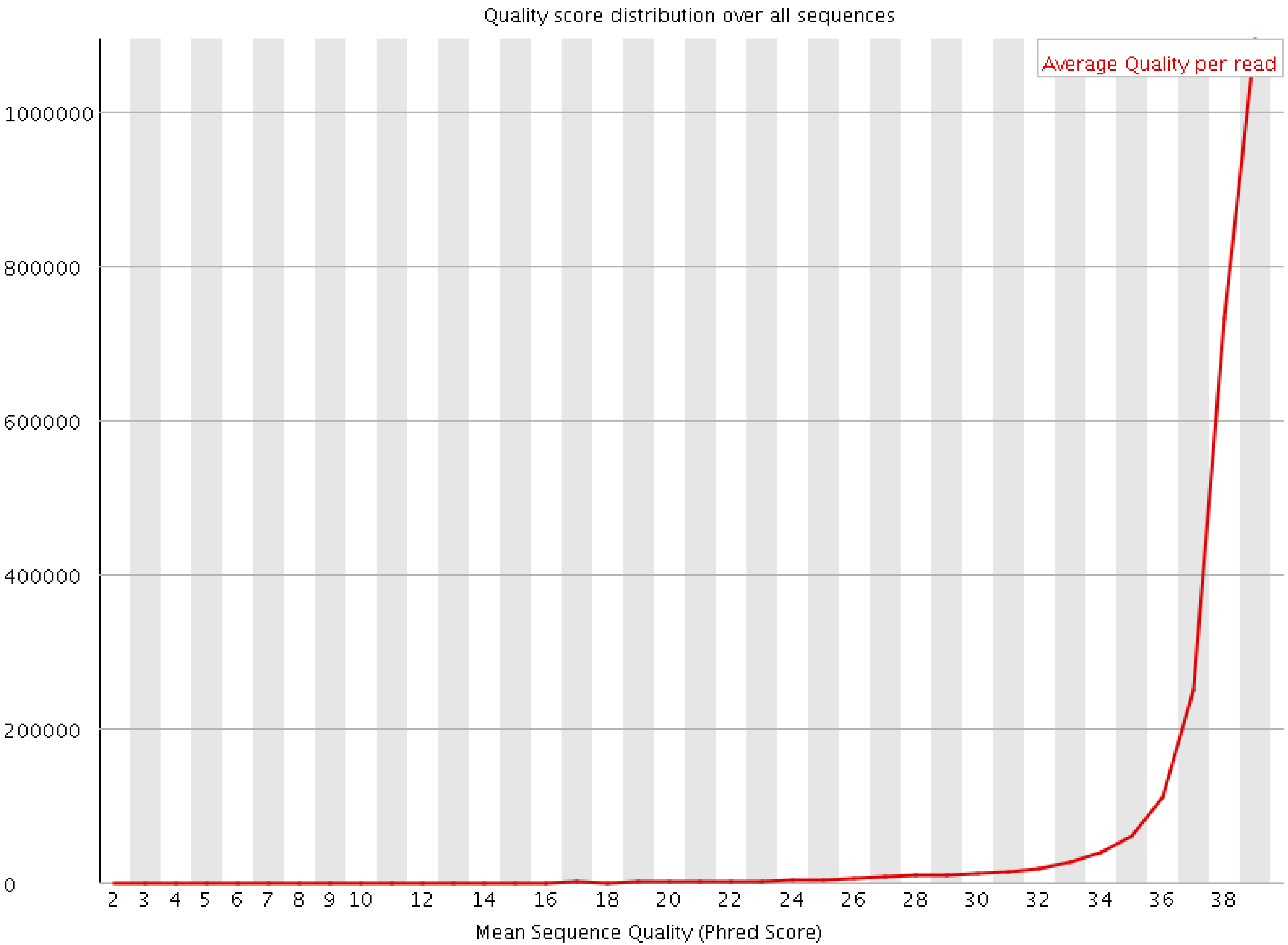
Should we be worried about sequence duplication levels?

What about sequence content and per base GC content?

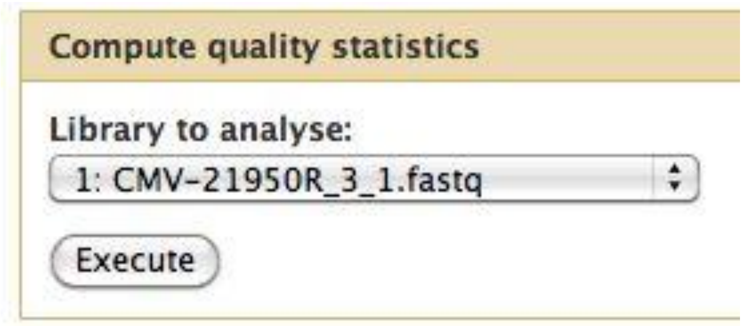


Summary

- ✓ 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


Quality score distribution over all sequences
Average Quality per read
1000000
800000
600000
400000
200000
0
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
Mean Sequence Quality (Phred Score)

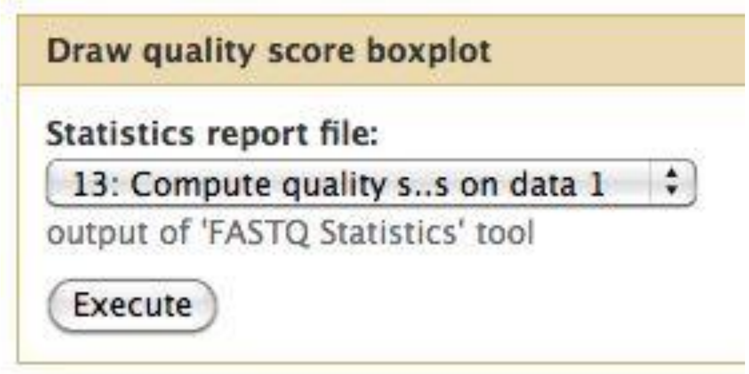
More detailed quality breakdown



Compute quality statistics

Library to analyse:

1: CMV-21950R_3_1.fastq



Draw quality score boxplot

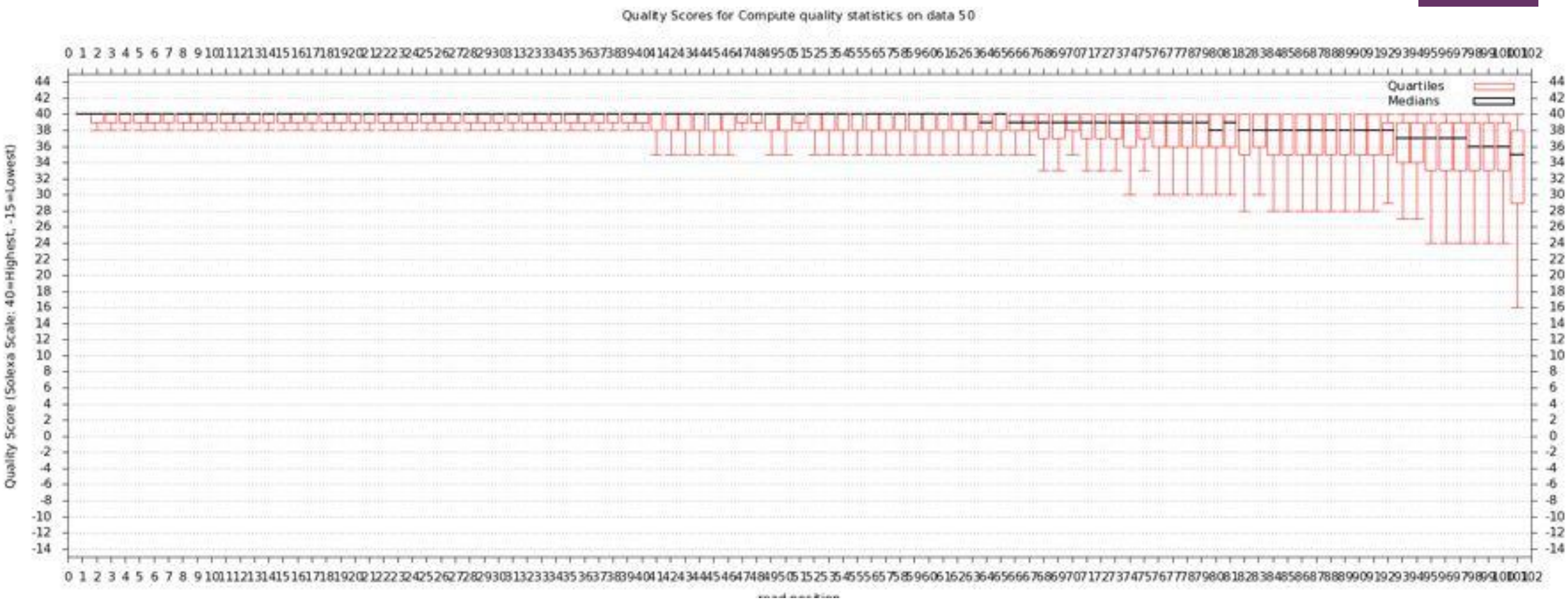
Statistics report file:

13: Compute quality s..s on data 1

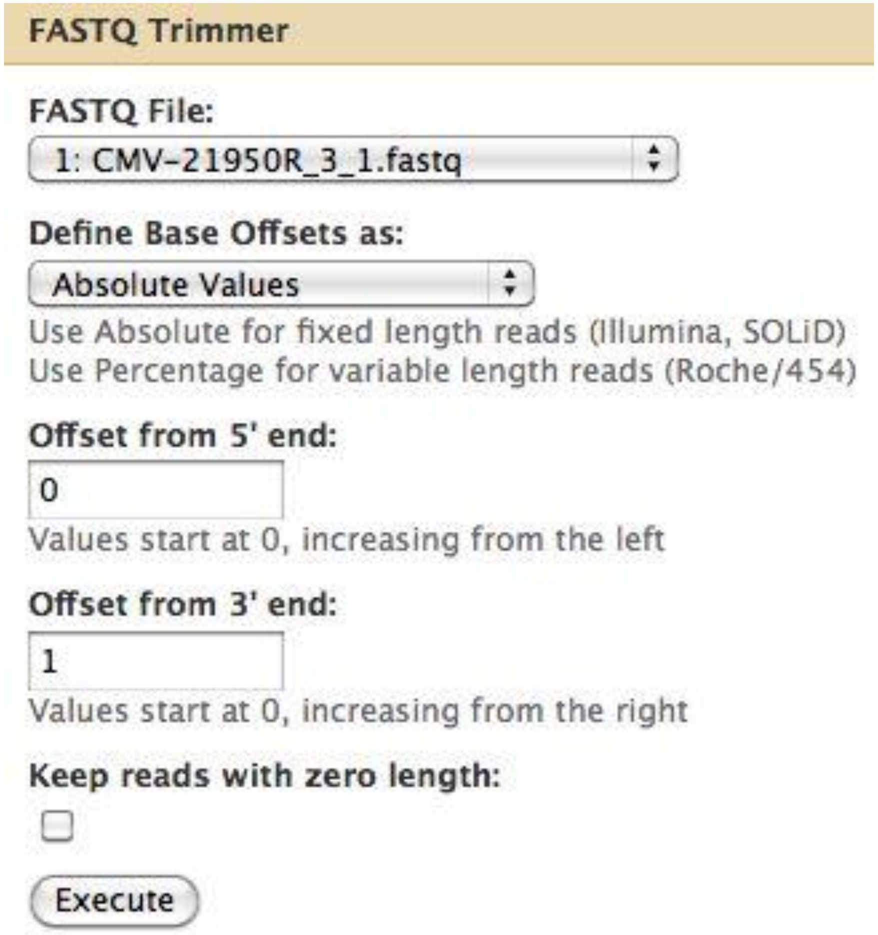
output of 'FASTQ Statistics' tool

- Compute quality statistics and draw a boxplot for all 6 Fastq files
  - 12 jobs total
  - Queue them up, no need to wait for previous job to finish

The last base sequenced (102) is of poor quality



Trim CMV-21950R_3_1.fastq from the 3` end by one base pair



FASTQ Trimmer

FASTQ File:

1: CMV-21950R_3_1.fastq

Define Base Offsets as:

Absolute Values

Use Absolute for fixed length reads (Illumina, SOLiD)
Use Percentage for variable length reads (Roche/454)

Offset from 5' end:

0

Values start at 0, increasing from the left

Offset from 3' end:

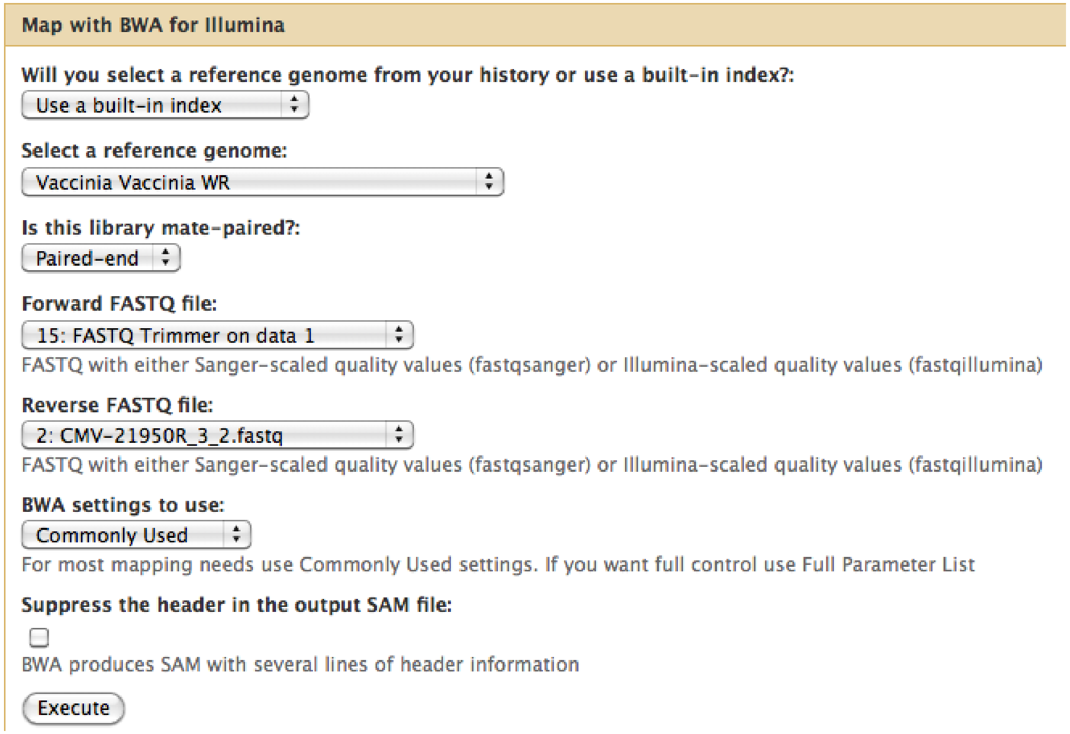
1

Values start at 0, increasing from the right

Keep reads with zero length:

□

Run our BWA alignments



Map with BWA for Illumina

Will you select a reference genome from your history or use a built-in index?: Use a built-in index

Select a reference genome: Vaccinia Vaccinia WR

Is this library mate-paired?: Paired-end

Forward FASTQ file: 15: FASTQ Trimmer on data 1

FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)

Reverse FASTQ file: 2: CMV-21950R_3_2.fastq

FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)

BWA settings to use: Commonly Used

For most mapping needs use Commonly Used settings. If you want full control use Full Parameter List

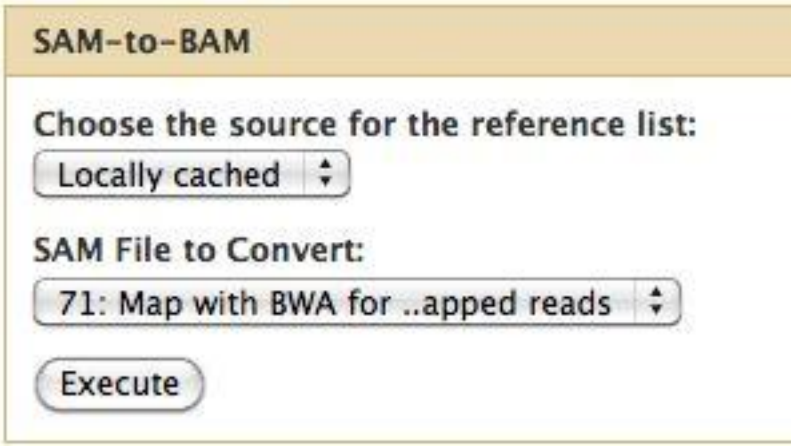
Suppress the header in the output SAM file: □

BWA produces SAM with several lines of header information

- BWA will align all the short reads to our reference genome
  - BWA is the algorithm of choice for DNA-Seq with Illumina data
  - CASAVA 1.8 may do as well for SNPs, BWA does indels better
- Select NGS Mapping -> Map with BWA for Illumina
- Set reference genome as vaccinia
- Set paired-end, be careful selecting pairs as the library is out of order
- Run for all 3 – reminder - don't wait for previous results

Samtools – SAM to BAM

- We have an alignment – how do we look at it?
- One common option is to convert the SAM file to a BAM file and examine in a viewer
- BAM files are required by a lot of downstream software and are smaller than the SAM files
- Could do filtering and sorting of the SAM file before doing this, but not needed in this case
- Convert all 3 SAM files to BAM
- This should work even if the SAM files are "empty"

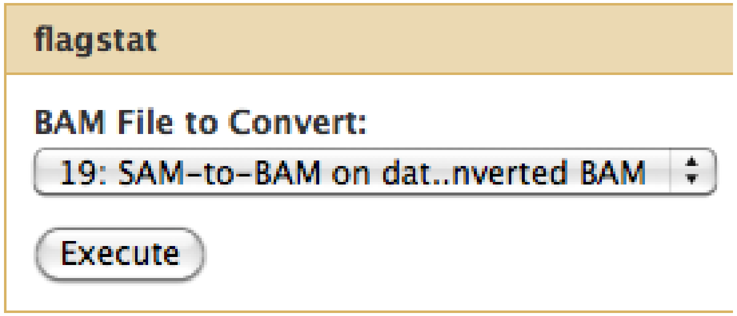


SAM-to-BAM

Choose the source for the reference list:
Locally cached

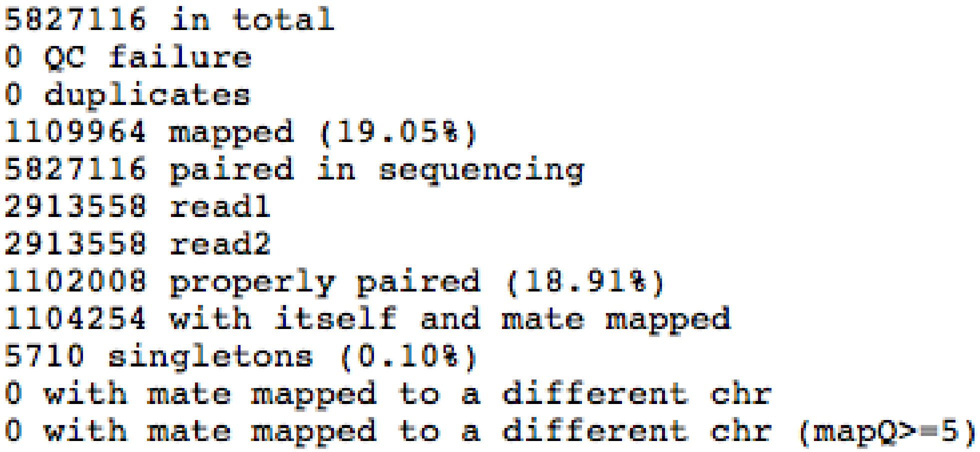
SAM File to Convert:
71: Map with BWA for ..apped reads

Samtool - Flagstats



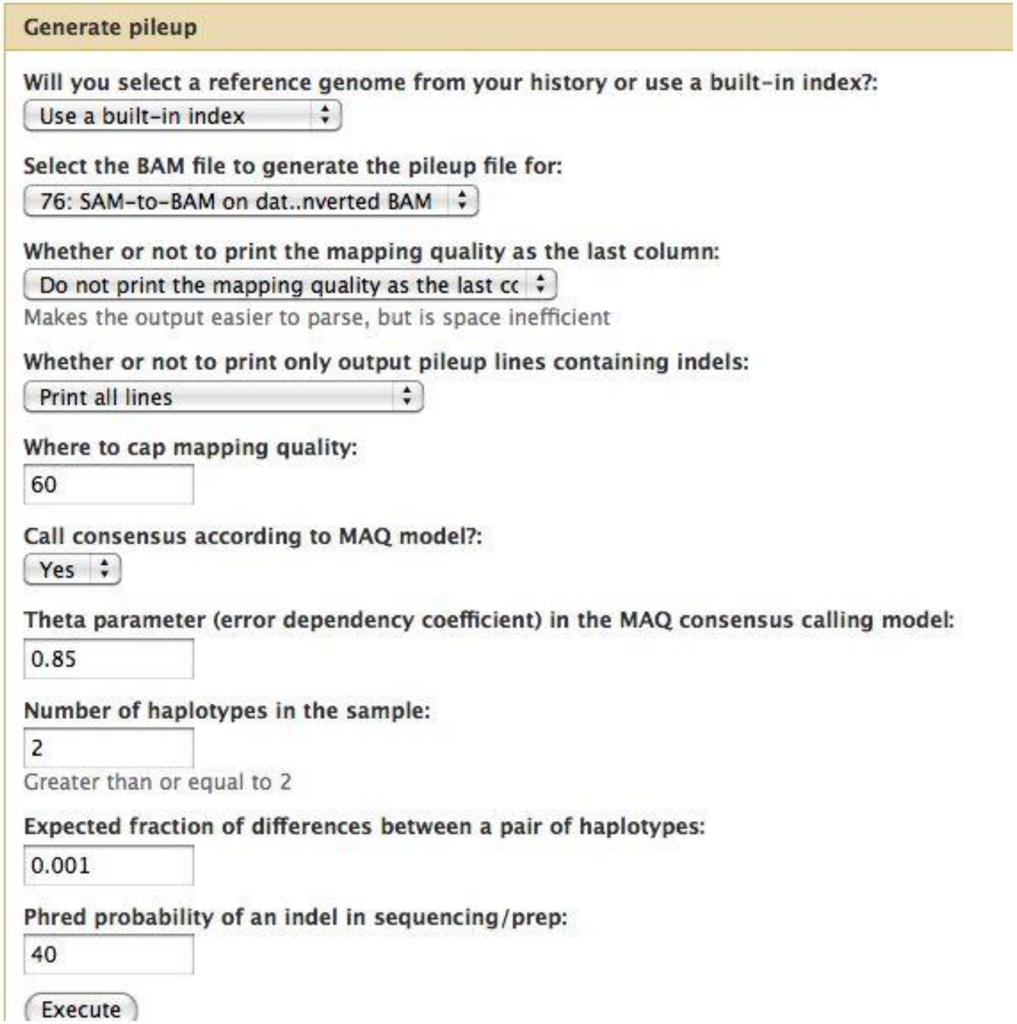
flagstat

BAM File to Convert: 19: SAM-to-BAM on dat..nverted BAM



```
5827116 in total
0 QC failure
0 duplicates
1109964 mapped (19.05%)
5827116 paired in sequencing
2913558 read1
2913558 read2
1102008 properly paired (18.91%)
1104254 with itself and mate mapped
5710 singletons (0.10%)
0 with mate mapped to a different chr
0 with mate mapped to a different chr (mapQ>=5)
```

- No QC failures or optical duplicates, so we didn't need to flag those
- Could have removed improperly paired reads but that would remove a lot of singletons mapping to different ITRs
- The big surprising result – low mapping of results!
- Contamination? I don't know yet.



Generate pileup

Will you select a reference genome from your history or use a built-in index?:
Use a built-in index

Select the BAM file to generate the pileup file for:
76: SAM-to-BAM on dat..nverted BAM

Whether or not to print the mapping quality as the last column:
Do not print the mapping quality as the last cc
Makes the output easier to parse, but is space inefficient

Whether or not to print only output pileup lines containing indels:
Print all lines

Where to cap mapping quality:
60

Call consensus according to MAQ model?:
Yes

Theta parameter (error dependency coefficient) in the MAQ consensus calling model:
0.85

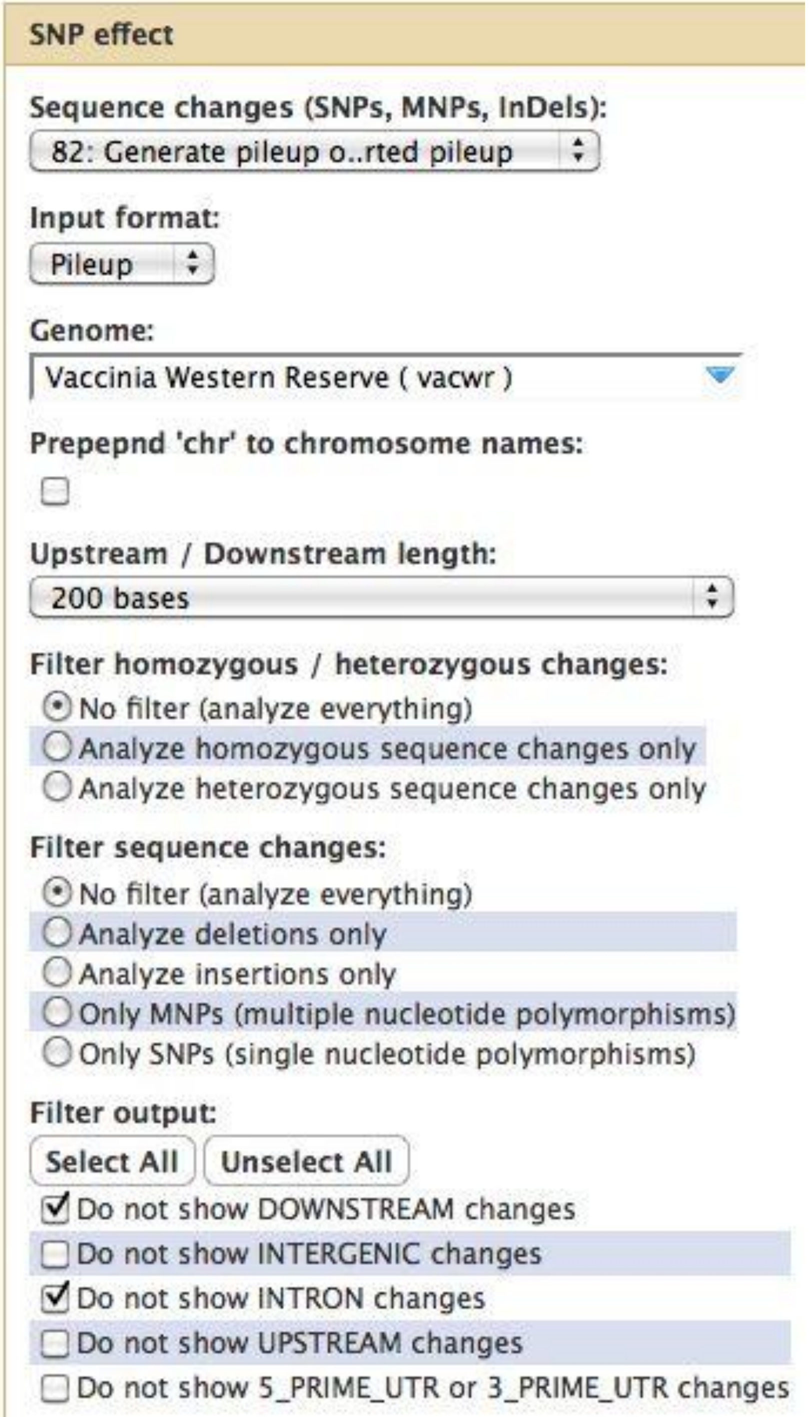
Number of haplotypes in the sample:
2
Greater than or equal to 2

Expected fraction of differences between a pair of haplotypes:
0.001

Phred probability of an indel in sequencing/prep:
40

Samtools – Generate Pileup

Make sure consensus is called according to the MAQ model



SNP effect

Sequence changes (SNPs, MNPs, InDels):

82: Generate pileup o..rted pileup

Input format:

Pileup

Genome:

Vaccinia Western Reserve (vacwr)

Prepepnd 'chr' to chromosome names:

□

Upstream / Downstream length:

200 bases

Filter homozygous / heterozygous changes:

● No filter (analyze everything)
○ Analyze homozygous sequence changes only
○ Analyze heterozygous sequence changes only

Filter sequence changes:

● No filter (analyze everything)
○ Analyze deletions only
○ Analyze insertions only
○ Only MNPs (multiple nucleotide polymorphisms)
○ Only SNPs (single nucleotide polymorphisms)

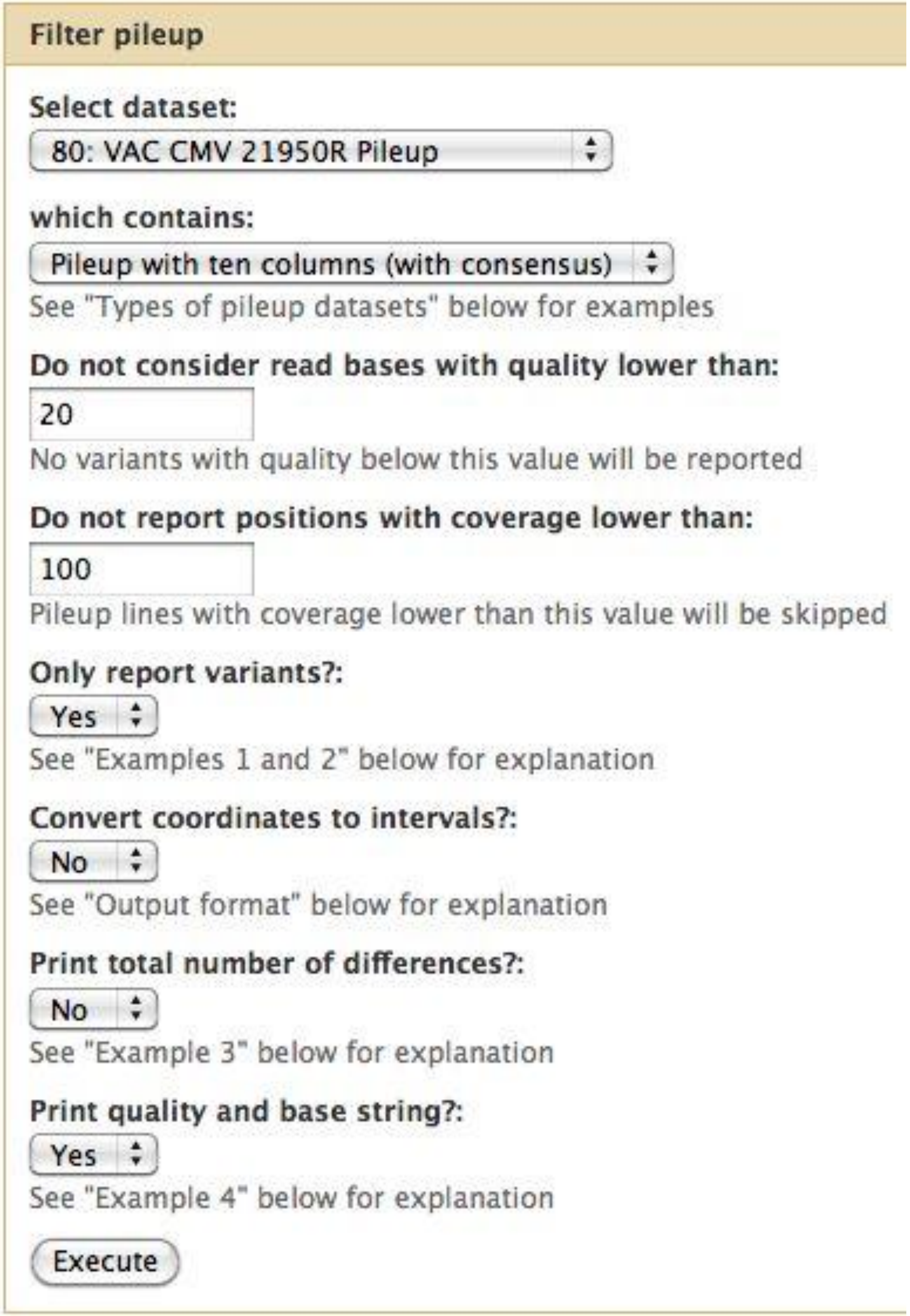
Filter output:

Select All | Unselect All

☑ Do not show DOWNSTREAM changes
□ Do not show INTERGENIC changes
☑ Do not show INTRON changes
□ Do not show UPSTREAM changes
□ Do not show 5_PRIME_UTR or 3_PRIME_UTR changes

SNPEff – All positions

- Needs a reference genome with GFF3 file to determine the impact of mutations
- Exclude introns has no real impact
- Application is near the bottom of the application pane
- Select "Pileup" as input format
- Results overwhelming, calls all minor variants as MNPs
- Better to filter first



Filter pileup

Select dataset:

80: VAC CMV 21950R Pileup

which contains:

Pileup with ten columns (with consensus)

See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:

20

No variants with quality below this value will be reported

Do not report positions with coverage lower than:

100

Pileup lines with coverage lower than this value will be skipped

Only report variants?:

Yes

See "Examples 1 and 2" below for explanation

Convert coordinates to intervals?:

No

See "Output format" below for explanation

Print total number of differences?:

No

See "Example 3" below for explanation

Print quality and base string?:

Yes

See "Example 4" below for explanation

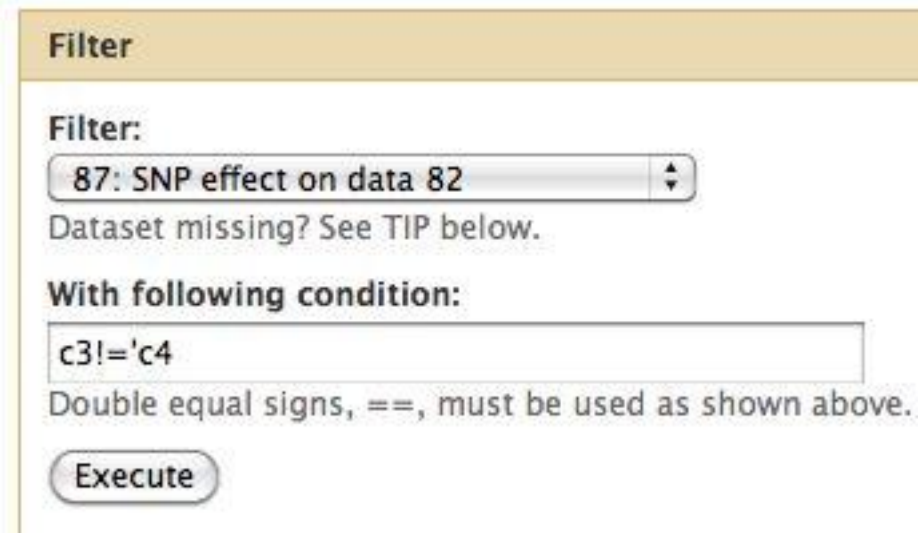
Filter Pileup

Remove low coverage areas, I choose 100 reads as a cutoff

Select 10 column format

If you don't have results, you can go back to the library and pull out already computed results

+ Pileup Results



Filter

Filter:

87: SNP effect on data 82

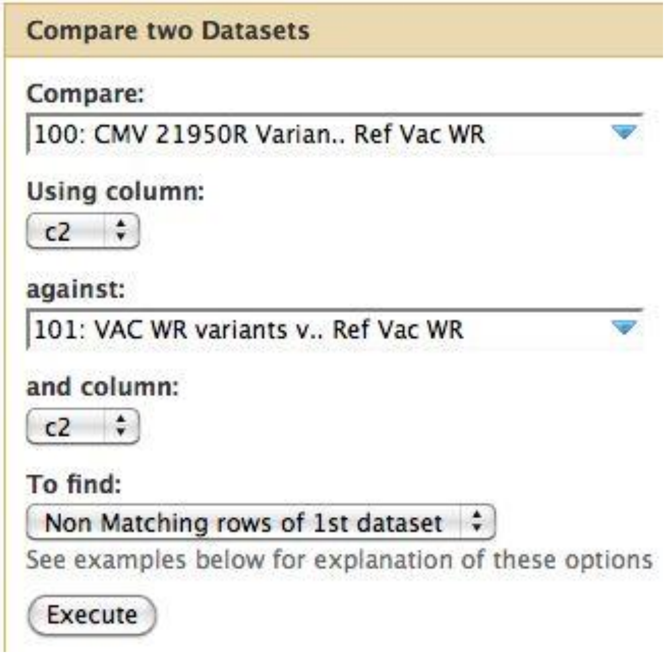
Dataset missing? See TIP below.

With following condition:

c3!='c4

Double equal signs, ==, must be used as shown above.

Comparative Genomics



Compare two Datasets

Compare: 100: CMV 21950R Varian.. Ref Vac WR

Using column: c2

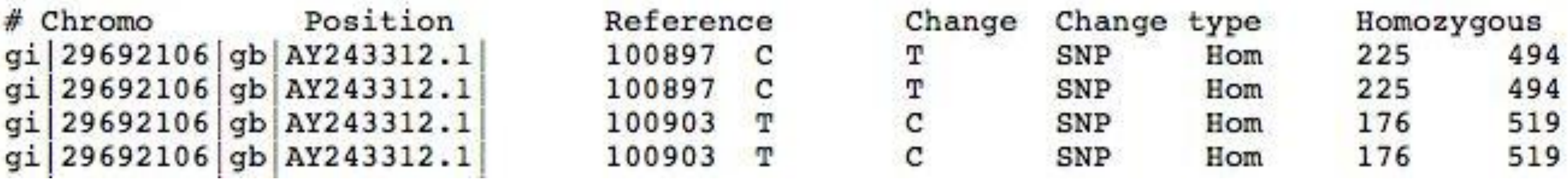
against: 101: VAC WR variants v.. Ref Vac WR

and column: c2

To find: Non Matching rows of 1st dataset

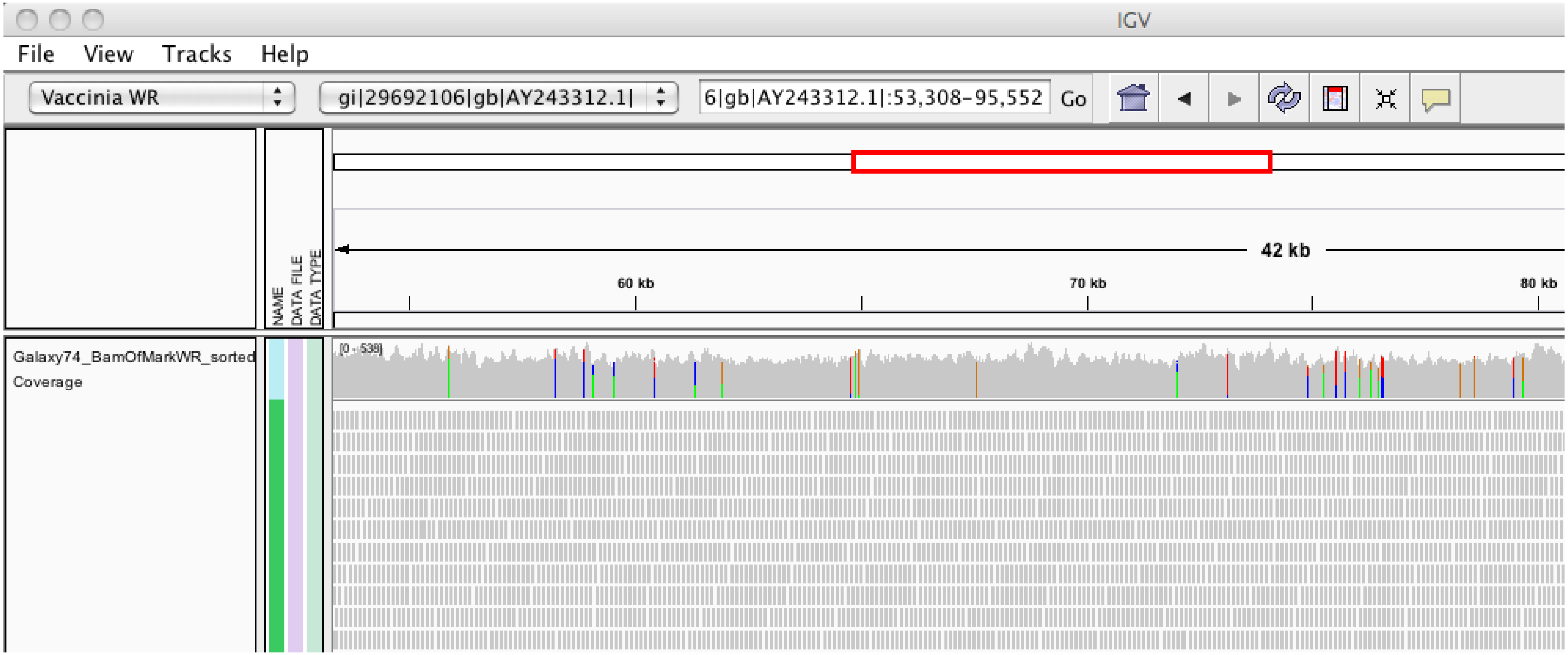
See examples below for explanation of these options

We only want differences between the mutant strains and the wild type.
Here we filter out all mutants in common between the CMV 21950R and the parental VAC WR strain

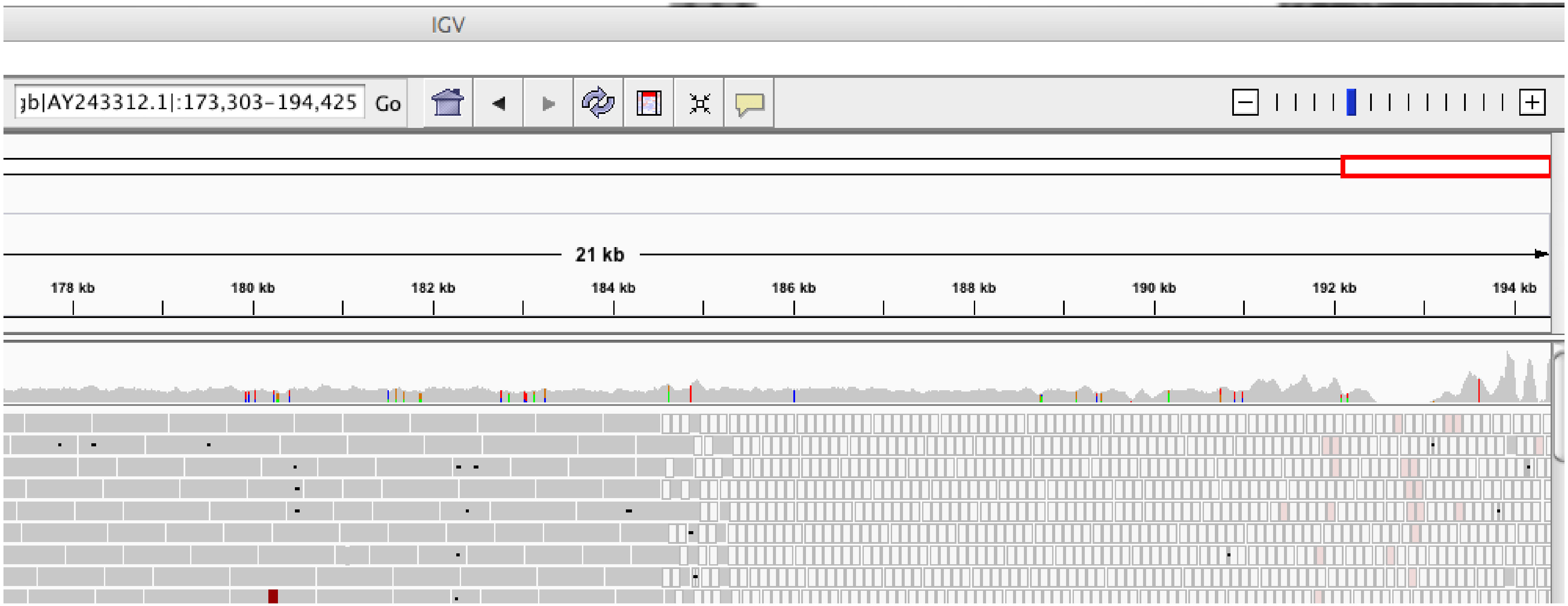


```
# Chromo           Position          Reference         Change   Change type       Homozygous
gi|29692106|gb|AY243312.1|           100897   C        T        SNP      Hom      225      494
gi|29692106|gb|AY243312.1|           100897   C        T        SNP      Hom      225      494
gi|29692106|gb|AY243312.1|           100903   T        C        SNP      Hom      176      519
gi|29692106|gb|AY243312.1|           100903   T        C        SNP      Hom      176      519
```

IGV – Central segment of Vaccinia WR



IGV – Vaccinia WR Right End



IGV
gb|AY243312.1|:173,303-194,425
Go
21 kb
178 kb
180 kb
182 kb
184 kb
186 kb
188 kb
190 kb
192 kb
194 kb