

# An Introduction to SeqMonk

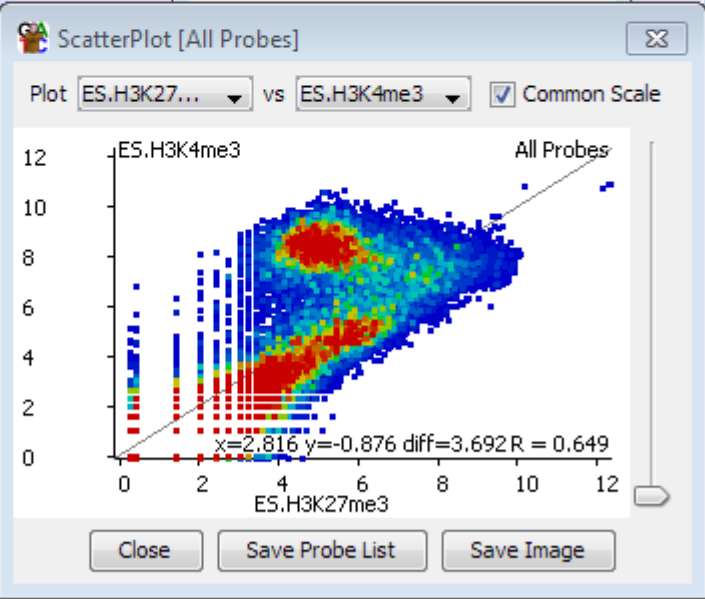
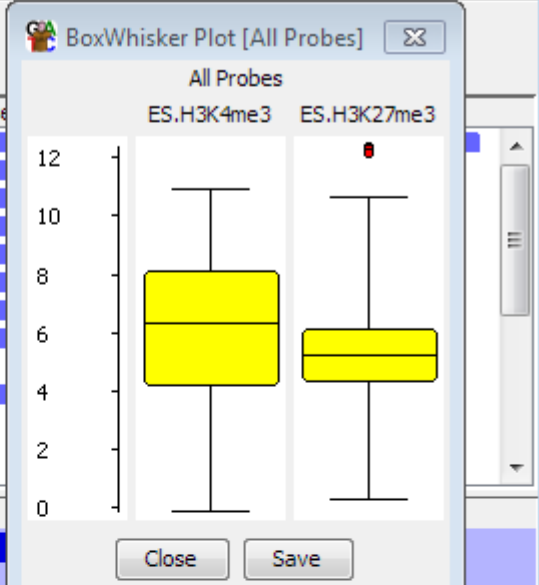
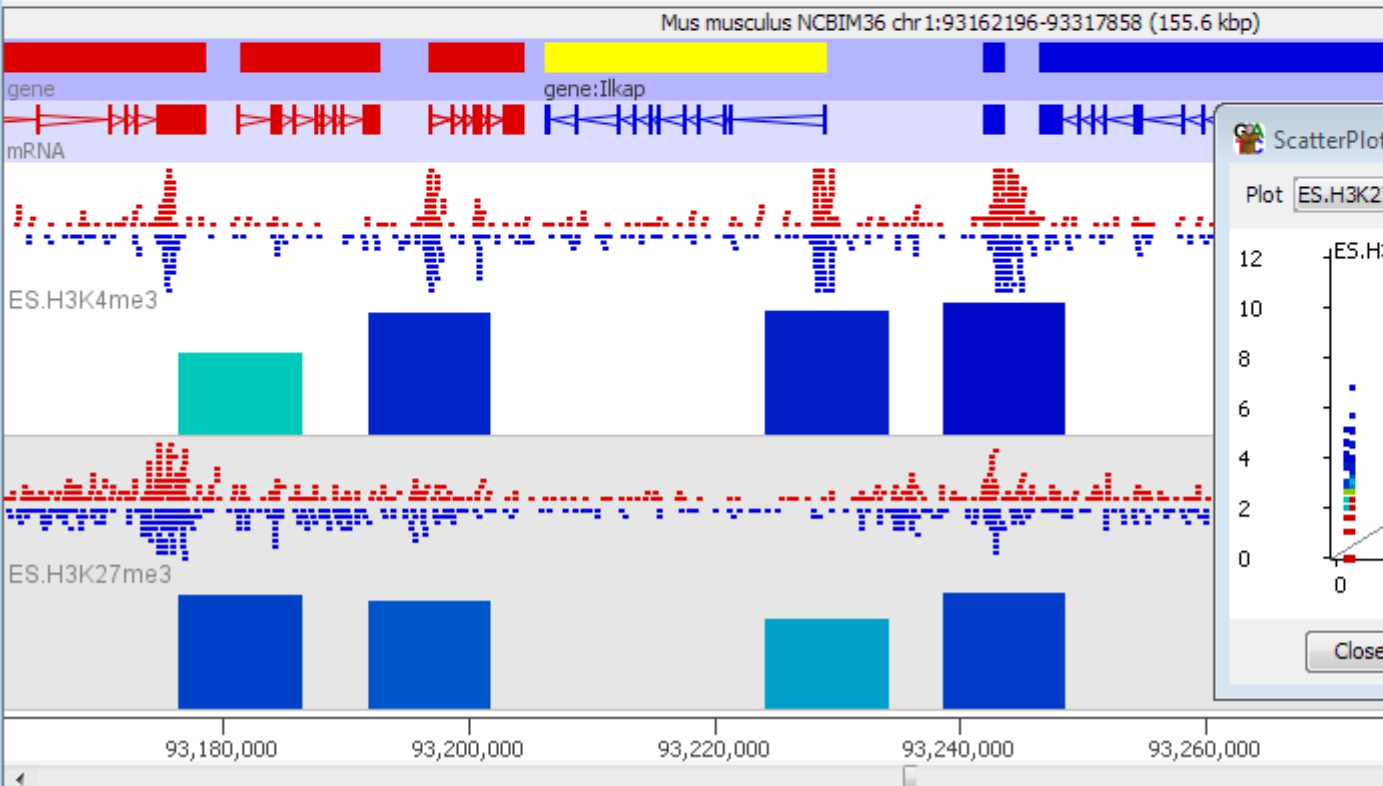
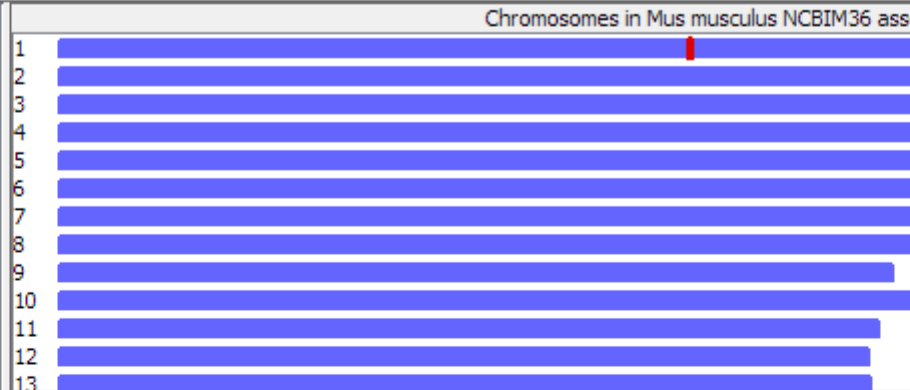
Simon Andrews

[simon.andrews@babraham.ac.uk](mailto:simon.andrews@babraham.ac.uk)

v2023-06



- Mus musculus NCBI36
  - Annotation Sets
  - Data Sets
    - ES.H3K27me3
    - ES.H3K4me3
  - Data Groups
  - Replicate Sets
  - Probe Lists
    - All Probes (28069)
    - Chr 1 (1582)
    - Difference above 4 (114)



# Course Programme

- Installing SeqMonk and Dependencies
- Creating a Project and Importing Data
- UI layout and basic controls
- Probes and Quantitation
- Plotting Figures
- Filtering Probes
- Saving, Reporting and Vistories

# Installing SeqMonk and Dependencies

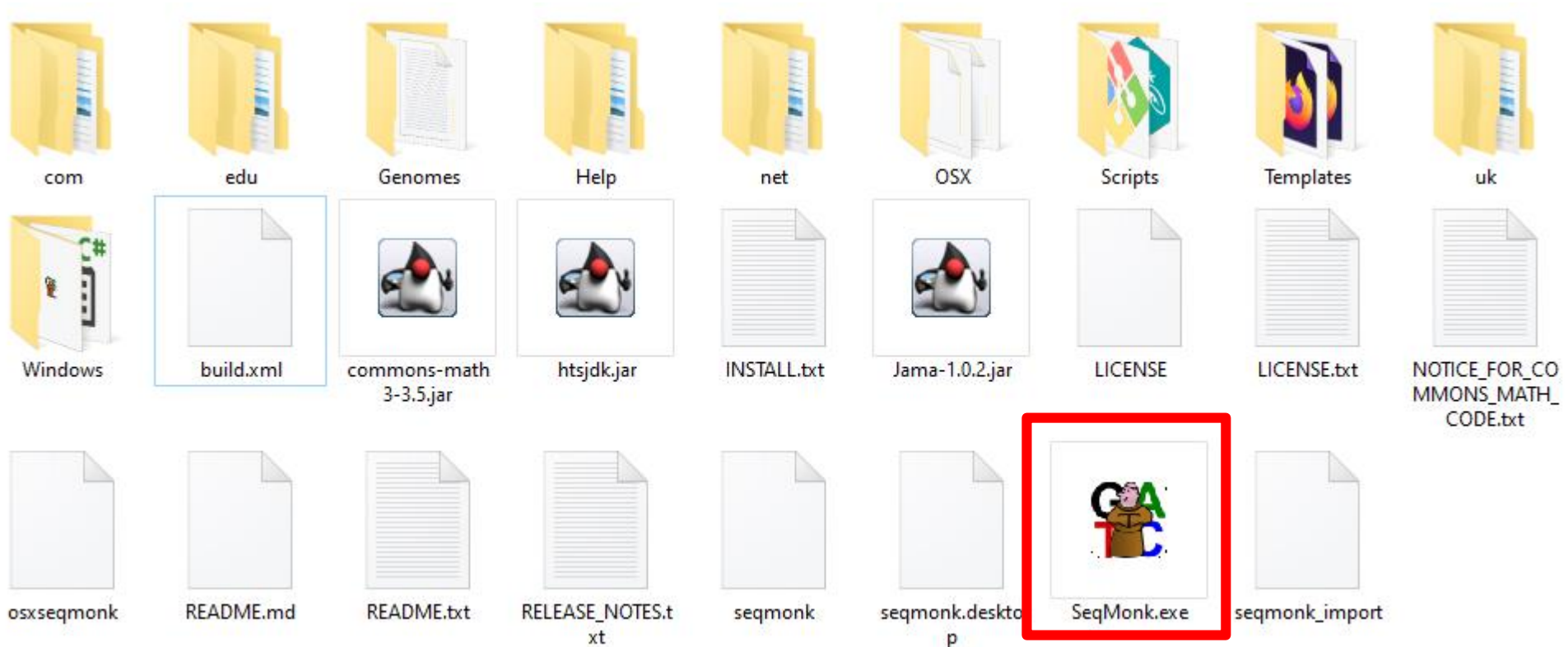
<https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/>

## **SeqMonk Mapped Sequence Analysis Tool**

- [README](#)
- [INSTALL](#) Installation instructions for the program.
- [Release Notes](#) Please read these before using the program.
- [SeqMonk v1.48.1 for 64-bit Windows](#)
- [SeqMonk v1.48.1 for 64-bit Linux](#)
- [SeqMonk v1.48.1 for 64-bit Mac OSX](#)

# Windows

- Unzip zip file



# Mac OSX

- Download and run the DMG file
- Copy App to the Applications folder

# Mac OS X





# Linux

- Download tar file
- Run the launcher

```
student@ip-172-31-23-170:~$ wget --quiet https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/seqmonk_v1.48.1_linux64.tar.gz
student@ip-172-31-23-170:~$ tar -xzf seqmonk_v1.48.1_linux64.tar.gz
student@ip-172-31-23-170:~$ cd SeqMonk/
student@ip-172-31-23-170:~/SeqMonk$ ./seqmonk
CLASSPATH is : /home/student/SeqMonk:/home/student/SeqMonk/htsjdk.jar:/home/student/SeqMonk/Jama-1.0.2.jar:/home/student/SeqMonk/comm
ons-math3-3.5.jar
Java interpreter is '/home/student/SeqMonk/jre/bin/java'
openjdk version "13.0.2" 2020-01-14
OpenJDK Runtime Environment AdoptOpenJDK (build 13.0.2+8)
OpenJDK 64-Bit Server VM AdoptOpenJDK (build 13.0.2+8, mixed mode, sharing)

Prefs file is at: /home/student/seqmonk_prefs.txt
Set memory to 0 from prefs file
Memory ceiling is 10240
Raw physical memory is 7847
Using 5231 MB of RAM to launch seqmonk
Correcting for VM actual requested allocation for 5231 is 5230
Command is: /home/student/SeqMonk/jre/bin/java -Xss4m -Xmx5230m -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true uk.ac.babraham.S
eqMonk.SeqMonkApplication
```

# Installing R

- <https://cran.r-project.org/>

## Download and Install R

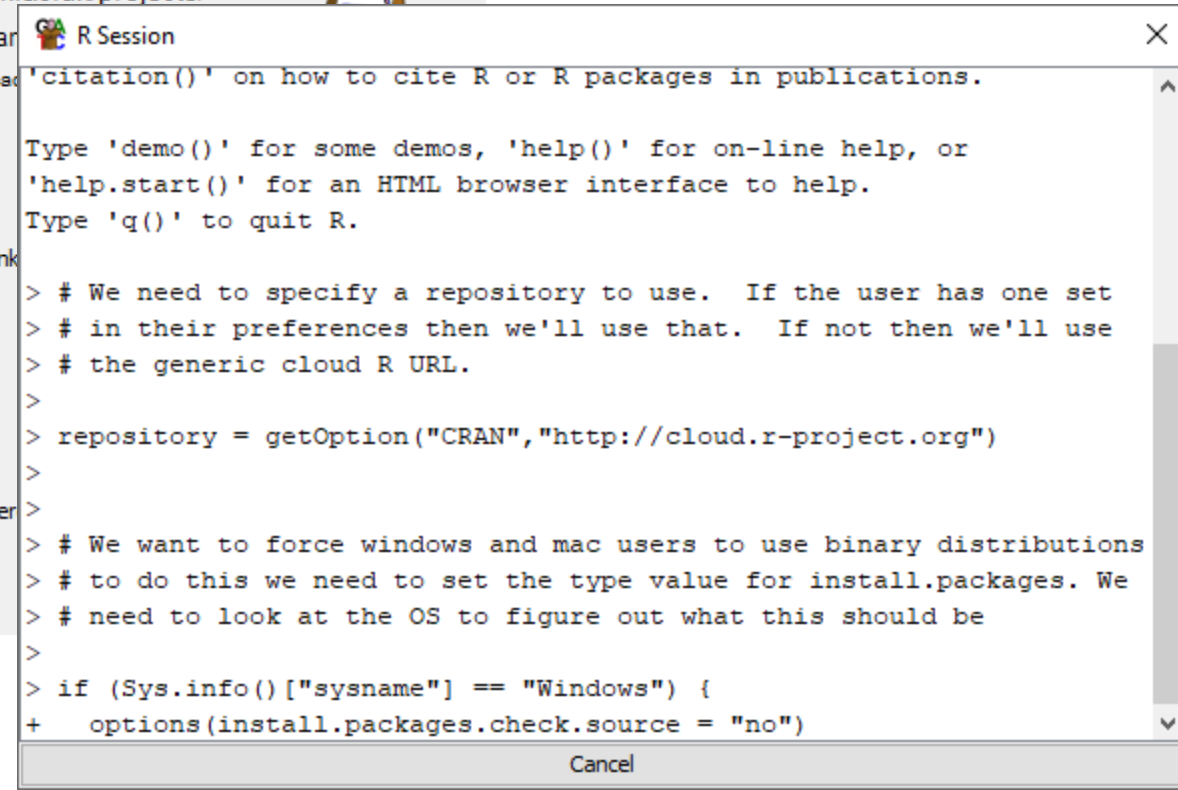
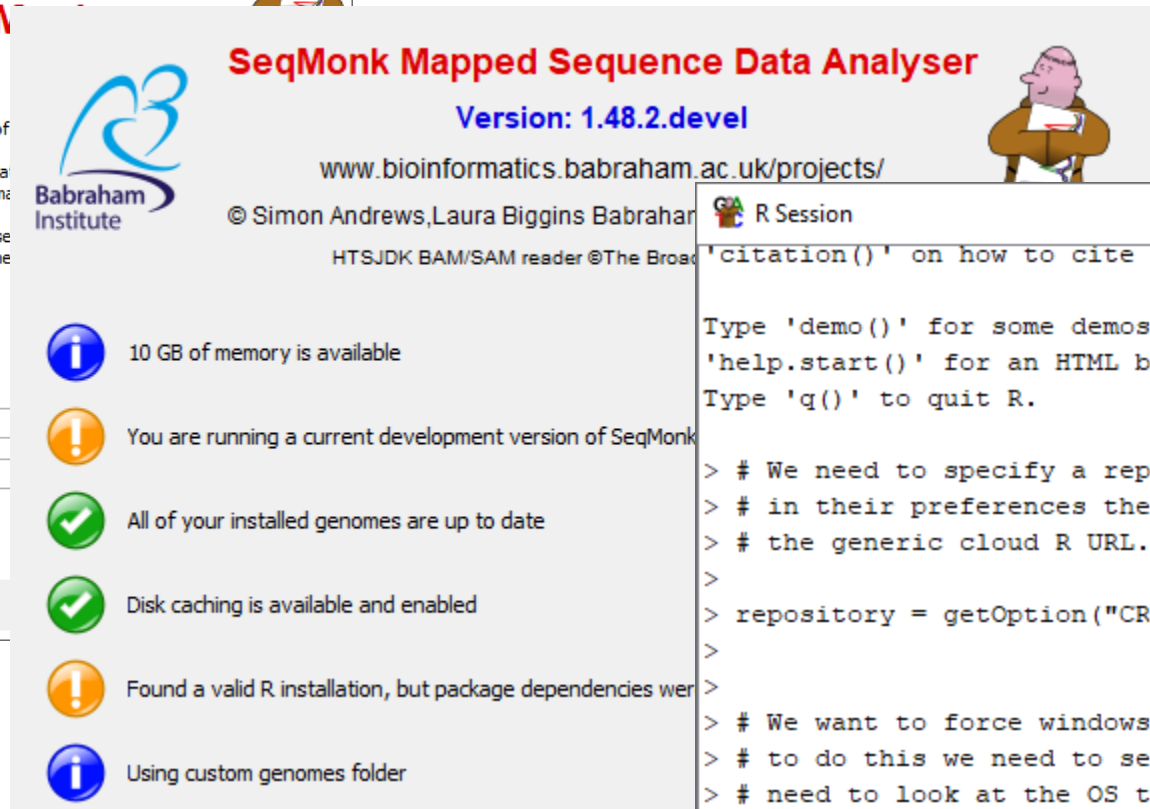
Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux \(Debian, Fedora/Redhat, Ubuntu\)](#)
- [Download R for macOS](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

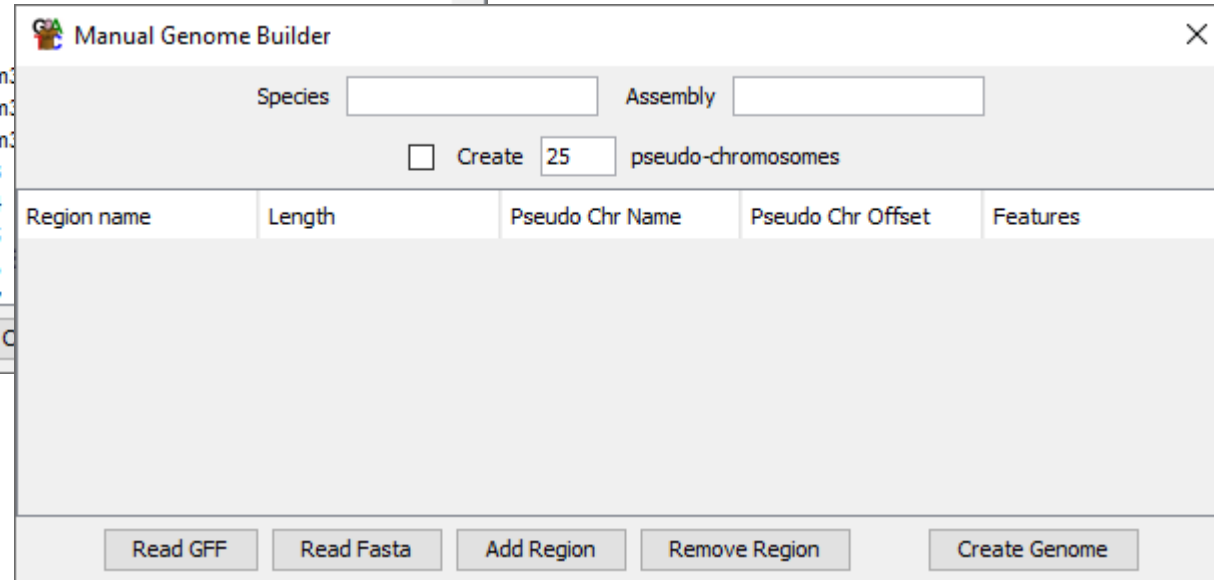
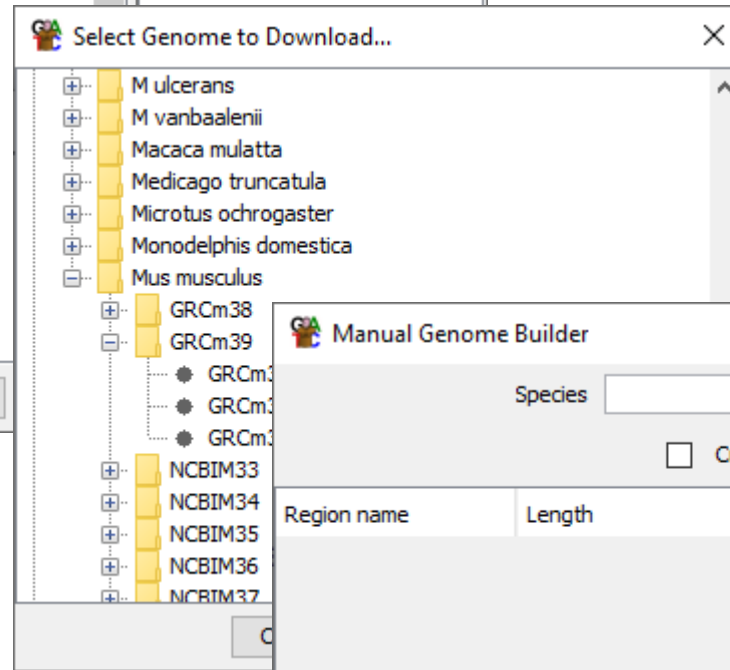
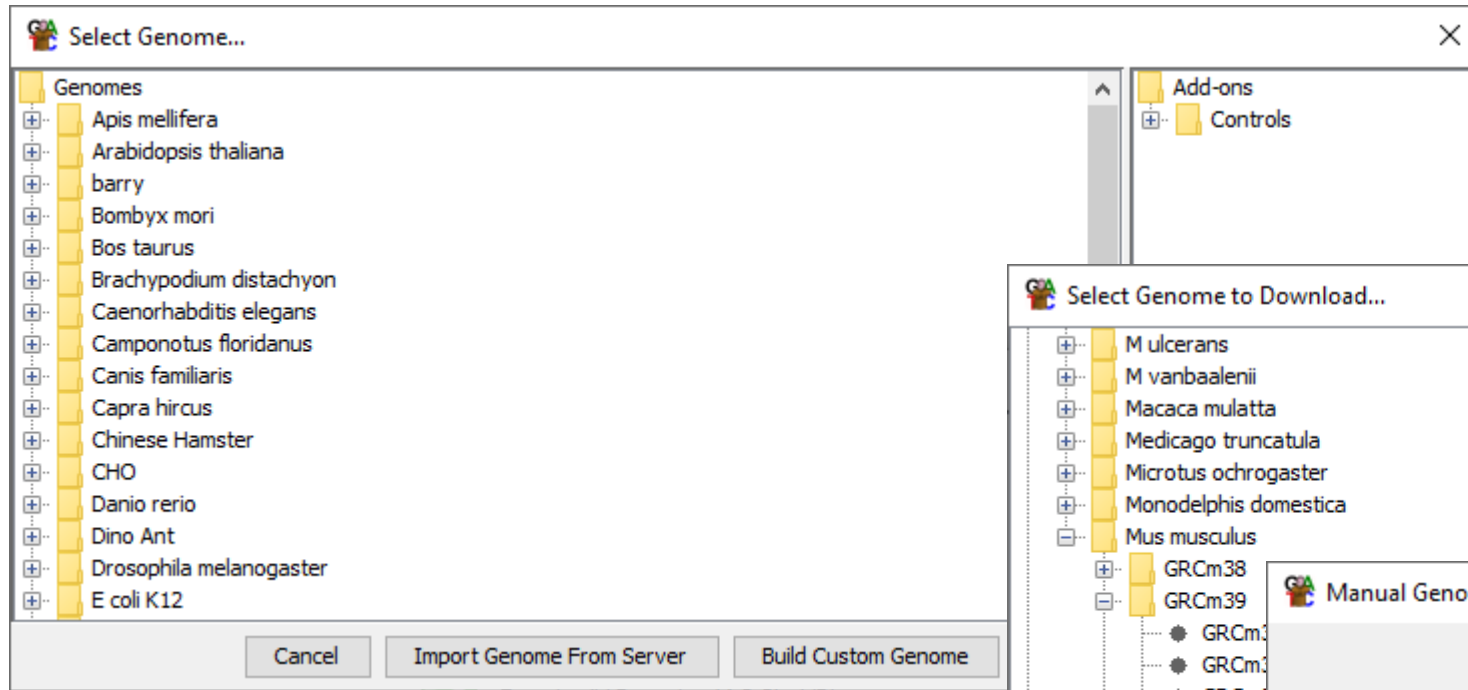
- Linux users need to install development versions of libssl, libcurl4, libxml2

# First Launch

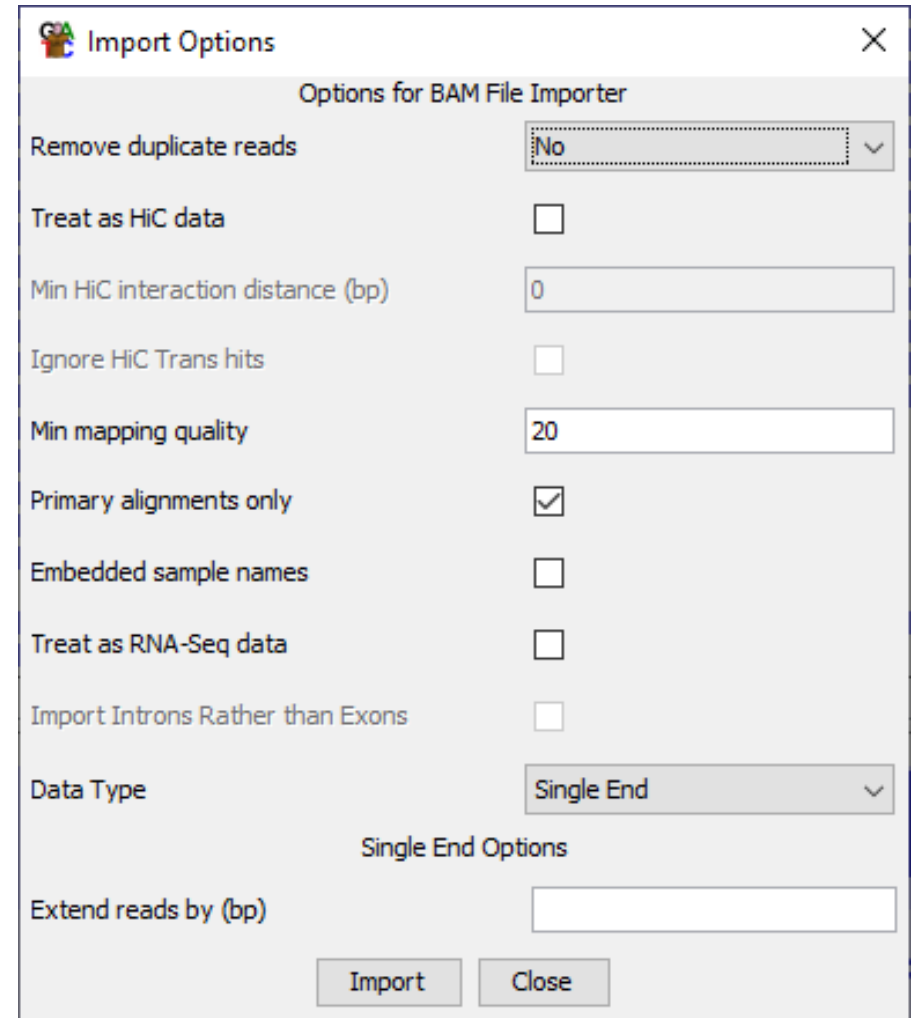
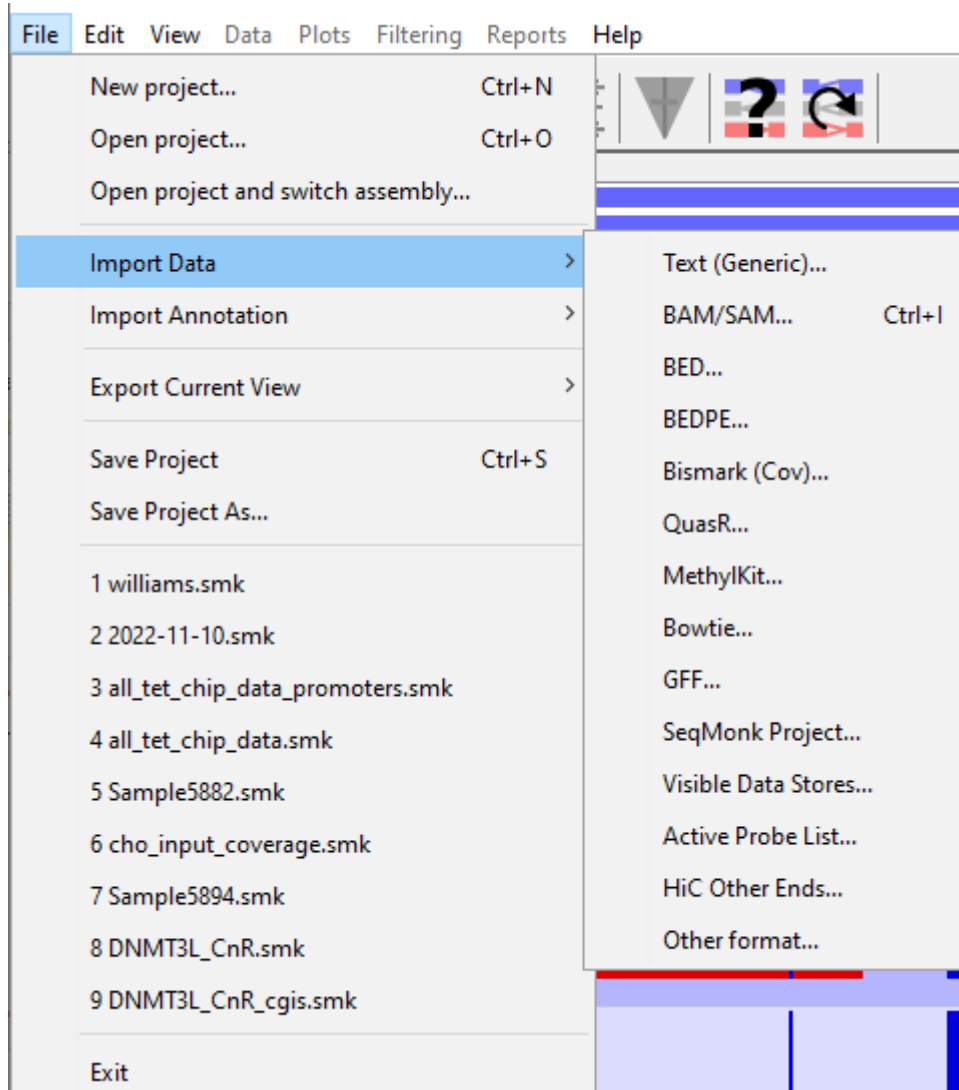


# Creating a Project and Importing Data

# File > New Project

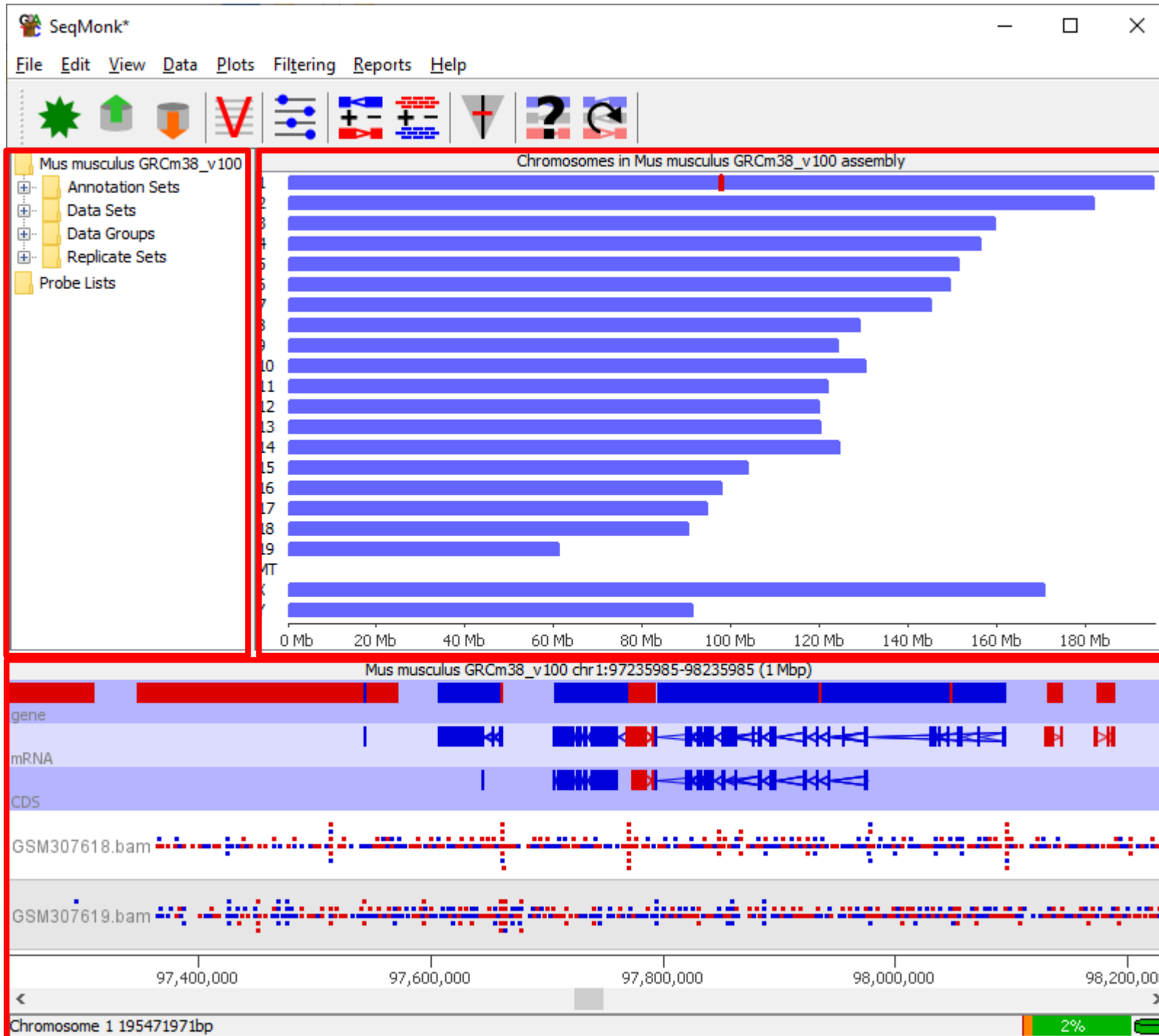


# File > Import Data



# UI Layout and Basic Controls

Data View



Genome View

Chromosome View



# Data Types

## Annotation

- Annotation collection
  - All of the annotation in the project
- Annotation set
  - A collection of features of varying types which came from the same source
- Annotation track
  - A set of features of the same type which might be drawn from several annotation sets

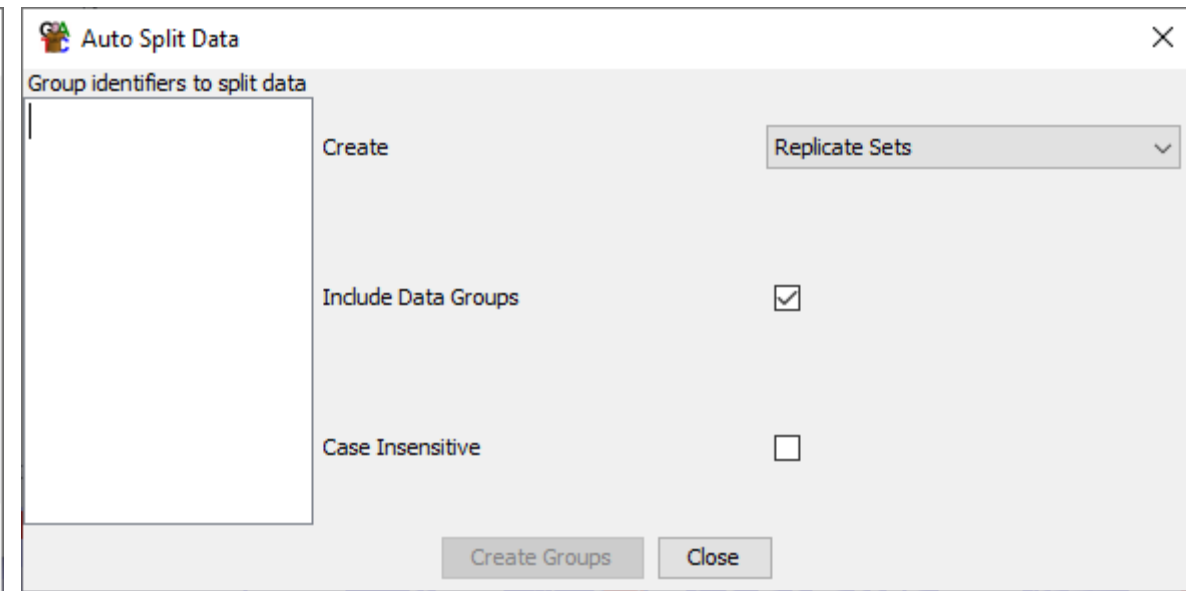
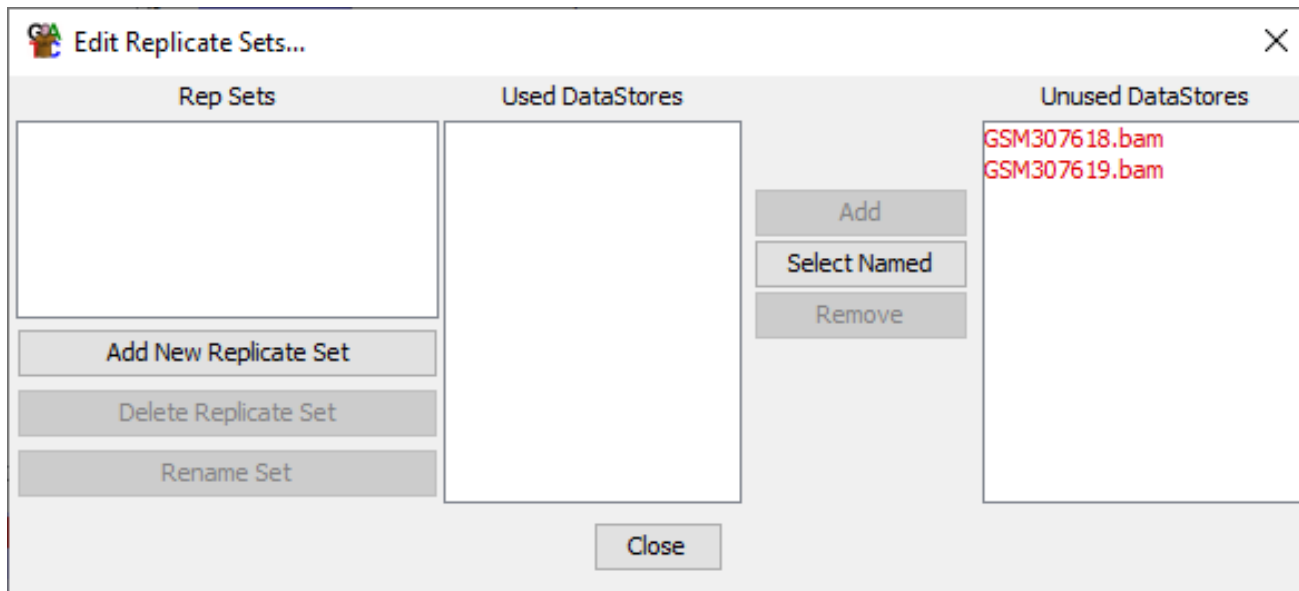
## Reads

- Data Set
  - A set of reads which came from one source (usually file)
- Data Group
  - A set of reads merged together from multiple datasets.
- Replicate Set
  - A collection of data sets / groups which come from the same biological condition

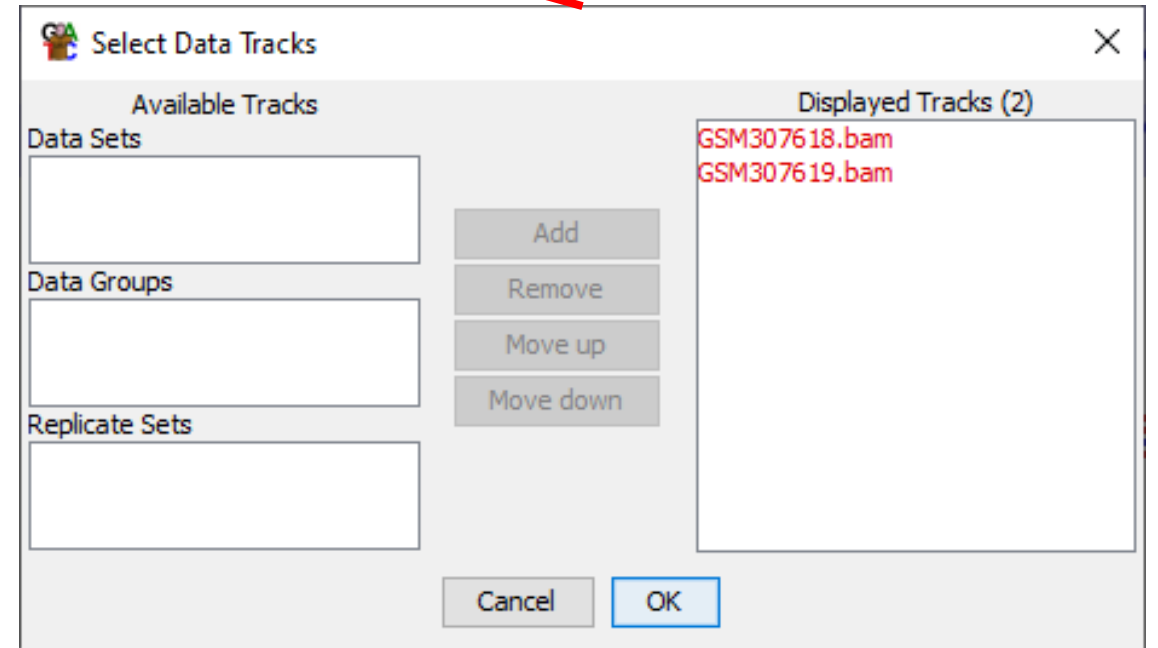
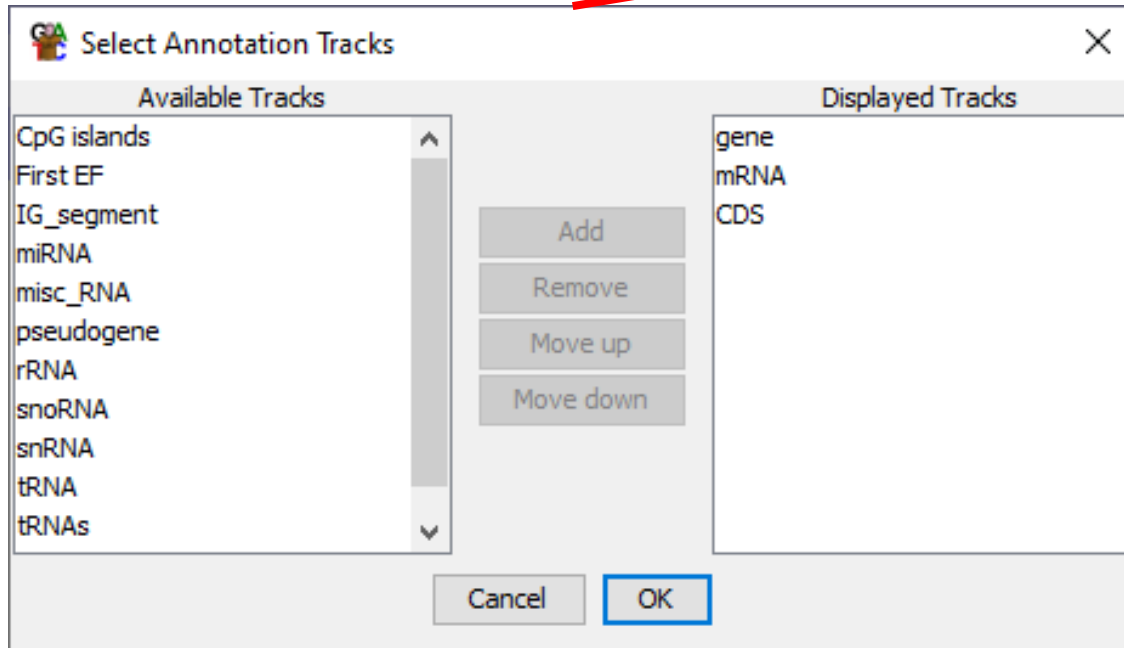
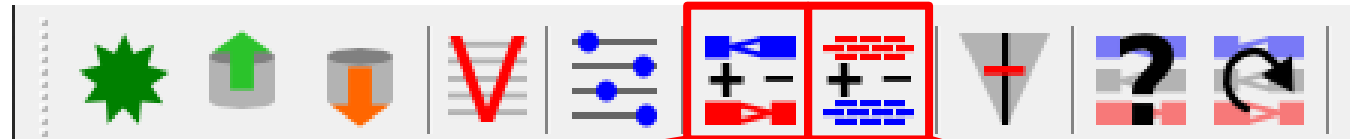
# Creating Data Groups / Replicate Sets

Data > Edit Replicate Sets


Data > Auto Create Groups/Sets





# Changing Chromosome Tracks




# Changing Display Preferences

 Edit Display Preferences ✕

Display which data  Reads Only ▾


Display reads with  Combined Strands ▾


Raw Read Display Density  Low Density ▾


Replicate Set Display Compressed ▾


Rep Set NA Exclusion Excluded ▾

Replicate Set Variability None ▾

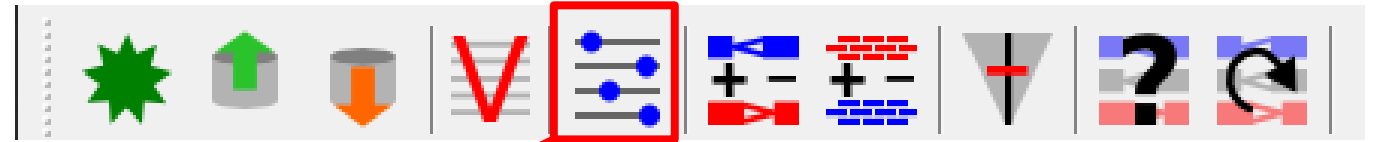
Display Quantitated Data as  Bars ▾

Quantitated Data Scale  Positive Only ▾

Quantitation Colour Scheme  Gradient Colours ▾

Colour Gradient Type  Cold - Hot ▾

Invert Gradient



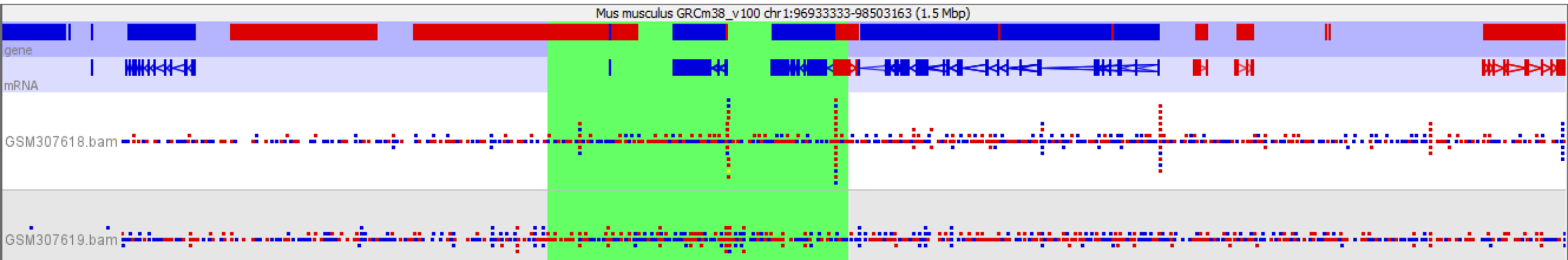
# Movement Controls

## Mouse

- Scroll Wheel to move left / right
- Click and drag to zoom in
- Right click to zoom out
- Double click for feature details

## Keyboard

- Up arrow to zoom in
- Down arrow to zoom out
- Left / Right arrows to move along
- Control +F to search (find)
- Control +G to jump to position (goto)



# Probes and Quantitation

# Terminology for Quantitation

- **Probe**
  - A region of the genome where a measurement will be made. Has a start and end, and optionally a strand
- **Probe Set**
  - The full set of probes currently being used for quantitation (eg all the promoters in the genome)
- **Probe List**
  - A subset of probes drawn from within the current Probe set (eg all of the promoters on chromosome 1)

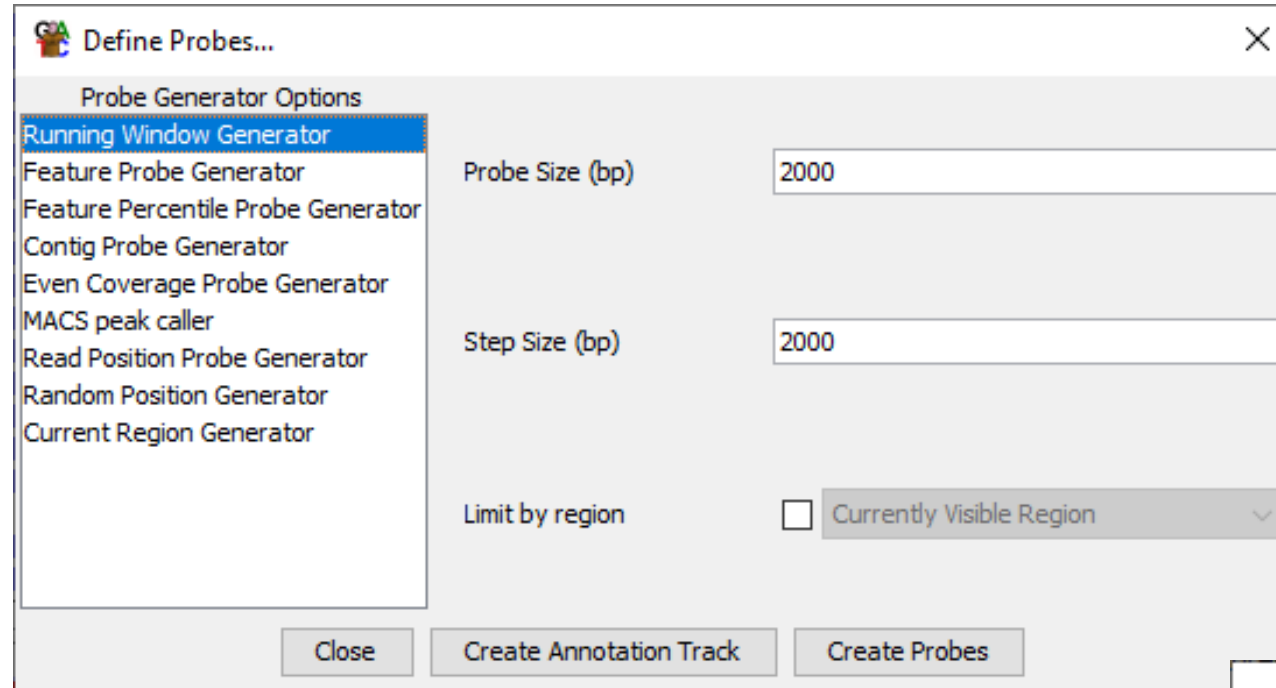
# Quantitation Rules

- A project can only have a single probe set, and the same probe set is used to quantitate all data
- Each probe has a quantitative value associated with it in every Data Set and Data Group
- Replicate Sets show the mean quantitation of the Data Sets within them
- The chromosome view will show only the currently selected probe list, and most plots only use data from the current probe list

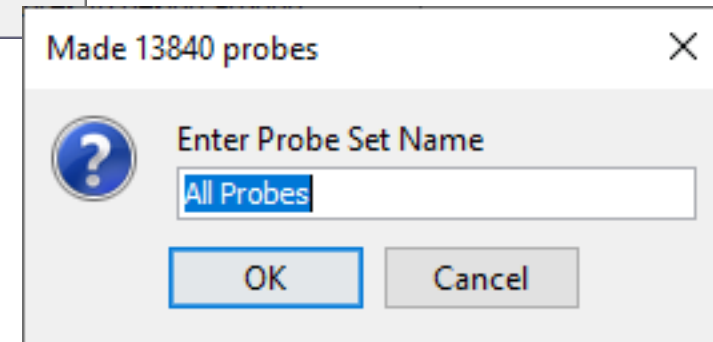


# Data > Define Probes

**Probe Generators**  
(Different ways of defining a probe set)

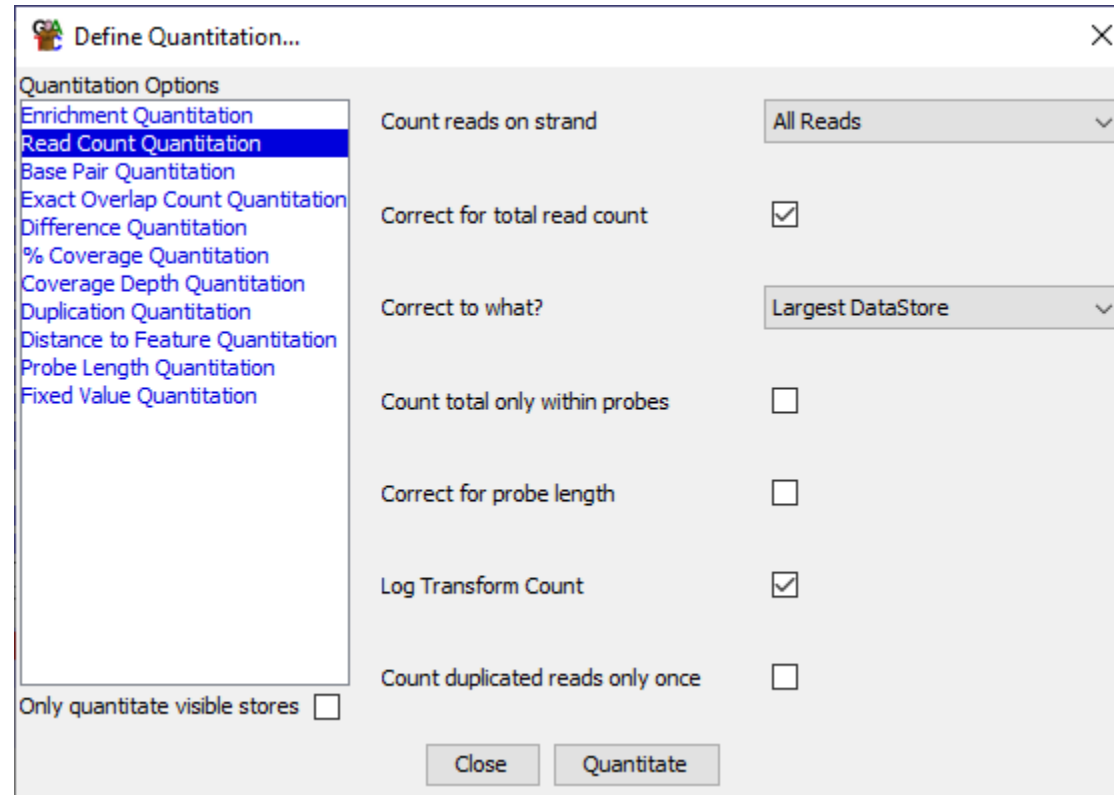


**Generator Options**  
(Options specific to the currently selected generator)



# Data > Quantitate Existing Probes

- Opens automatically after defining new probes
- Can be rerun on existing probes without changing them



**Quantitation Methods**  
(Different ways of assigning  
a value to a probe)

**Quantitation Options**  
(Options specific to the  
currently selected  
quantitation method)

# Quantitation Example

**Define Probes...**

Probe Generator Options

- Running Window Generator
- Feature Probe Generator**
- Feature Percentile Probe Generator
- Contig Probe Generator
- Even Coverage Probe Generator
- MACS peak caller
- Read Position Probe Generator
- Random Position Generator
- Current Region Generator

Features to design around

- CDS
- CpG islands**
- First EF
- gene
- IG\_segment
- miRNA
- misc\_RNA
- mRNA
- pseudogene
- rRNA

Split into subfeatures: No

Remove exact duplicates:

Ignore feature strand information:

Make probes: Centered on feature (dropdown) From - 2000 to + 2000 bp

Buttons: Close, Create Annotation Track, Create Probes

**Define Quantitation...**

Quantitation Options

- Enrichment Quantitation
- Read Count Quantitation**
- Base Pair Quantitation
- Exact Overlap Count Quantitation
- Difference Quantitation
- % Coverage Quantitation
- Coverage Depth Quantitation
- Duplication Quantitation
- Distance to Feature Quantitation
- Probe Length Quantitation
- Fixed Value Quantitation

Count reads on strand: All Reads (dropdown)

Correct for total read count:

Correct to what?: Per Million Reads (dropdown)

Count total only within probes:

Correct for probe length:

Log Transform Count:

Count duplicated reads only once:

Only quantitate visible stores:

Buttons: Close, Quantitate

**Made 13840 probes**

Enter Probe Set Name

CpG Islands

Buttons: OK, Cancel

# Quantitation Example

Probe Lists  
CpG Islands (13840)

CpG Islands (13840 probes)

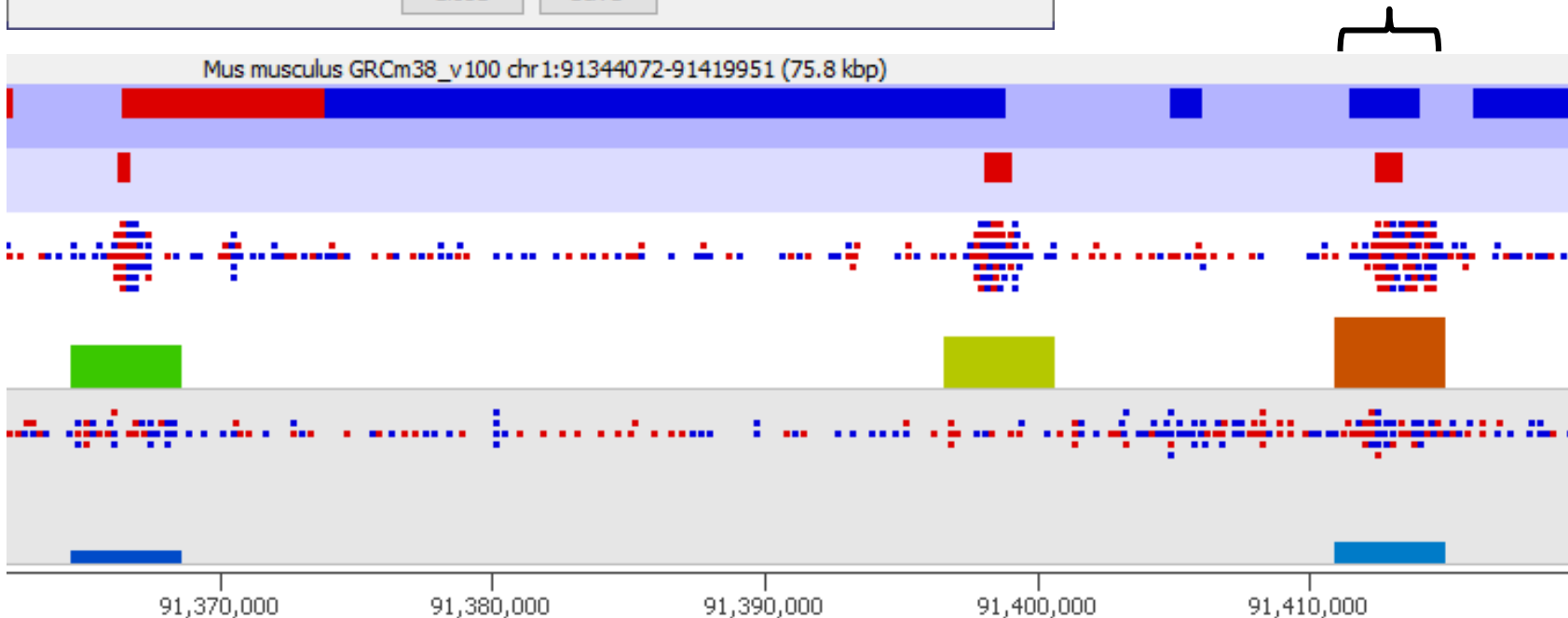
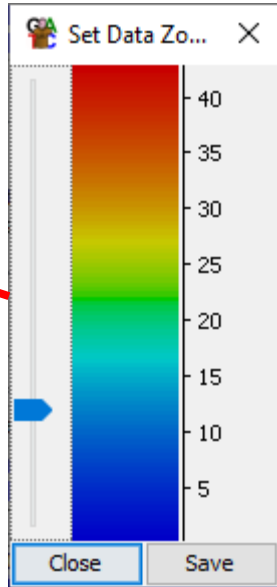
Description:  
Feature generator using CpG islands duplicates removed Centered on feature from 2000-2000. Quantitated with Read Count Quantitation using All Reads correcting for total count per million reads

Probe	Chr	Start	End
oe = 0.70 Chr1:366...	1	3669388	3673388
oe = 0.77 Chr1:449...	1	4490687	4494687

Close Save



**Bars = Probes**  
Check the positions match with what you expect



**Height = Value = Colour**  
Capped at 95<sup>th</sup> Percentile by default

# Quantitation Pipelines

- Data > Quantitation Pipelines
- Combine Probe Generation and Quantitation

**Pipeline**

Define Quantitation...

Quantitation Options

- RNA-Seq quantitation pipeline
- Active transcription quantitation pipeline
- Intron regression pipeline
- Gene trap quantitation pipeline
- Wiggle Plot for Initial Data Inspection
- Bisulphite methylation over features
- Splicing efficiency quantitation
- Antisense transcription pipeline
- Codon Bias Pipeline
- Transcription termination pipeline

Only quantitate visible stores

Transcript features mRNA

Library type Non-strand specific

Libraries are paired end

Merge transcript isoforms

Generate Raw Counts

Log transform

Apply transcript length correction

Don't quantitate probes with no counts

Correct for DNA Contamination

Correct for Duplication

Close Run Pipeline

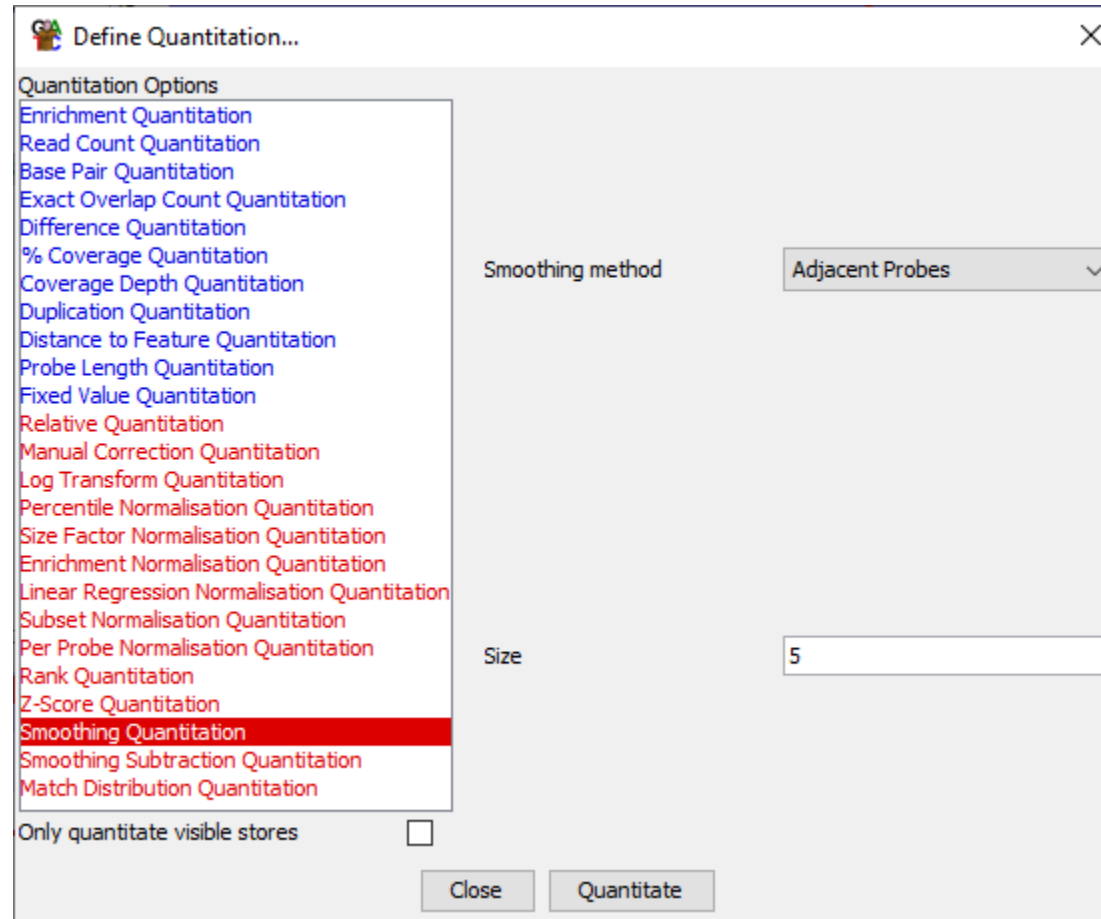
**Pipeline Options**

# Quantitation Adjustment

- Additional Options once you have a quantitation

**Blue**  
Fresh quantitation from the  
raw read data

**Red**  
Methods to normalise /  
scale / adjust the existing  
quantitation



# Plotting Figures

Probe Value Histogram...

Read Length Histogram...

Probe Length Histogram...

Data Store Similarity &gt;

Probe List Overlap &gt;

Domainogram...

Hierarchical Clusters &gt;

Cumulative Distribution Plot &gt;

QQ Distribution Plot &gt;

Bean Plot &gt;

Box Whisker Plot &gt;

Star Wars Plot &gt;

Probe Trend Plot... &gt;

Aligned Probes Plot &gt;

Quantitation Trend Plots.. &gt;

HiC Heatmap &gt;

HiC Cis/Trans Scatterplot

HiC Length Histogram &gt;

Scatter Plot...

MA Plot...

Volcano Plot...

Duplication Plot...

Variation Plot...

Strand Bias Plot...

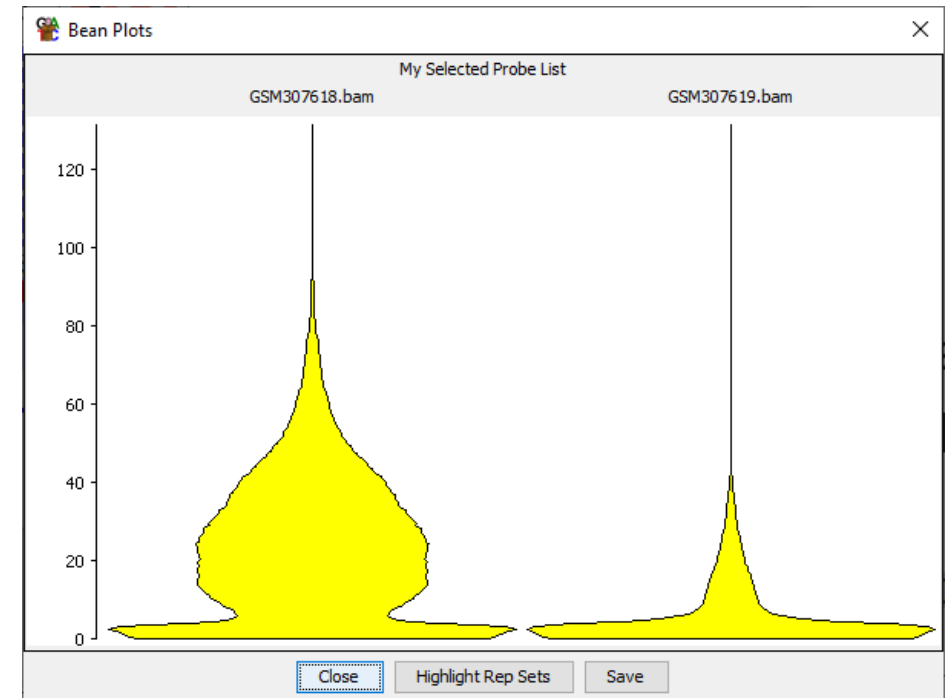
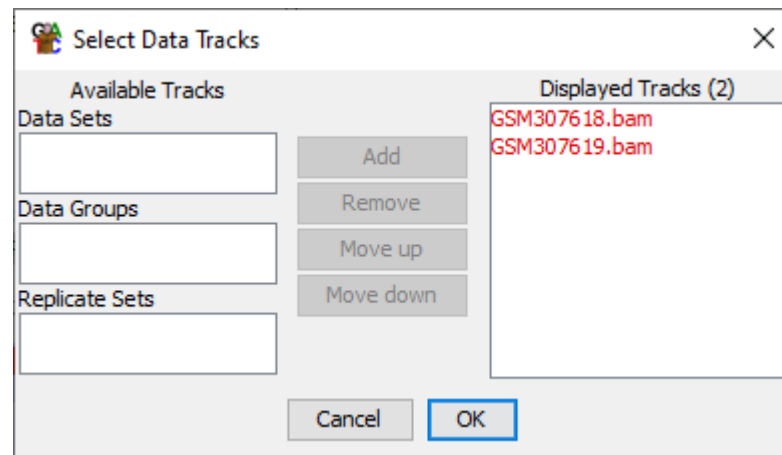
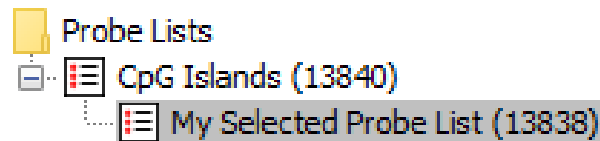
Line Graph... &gt;

RNA-Seq QC Plot...

Small RNA QC Plot...

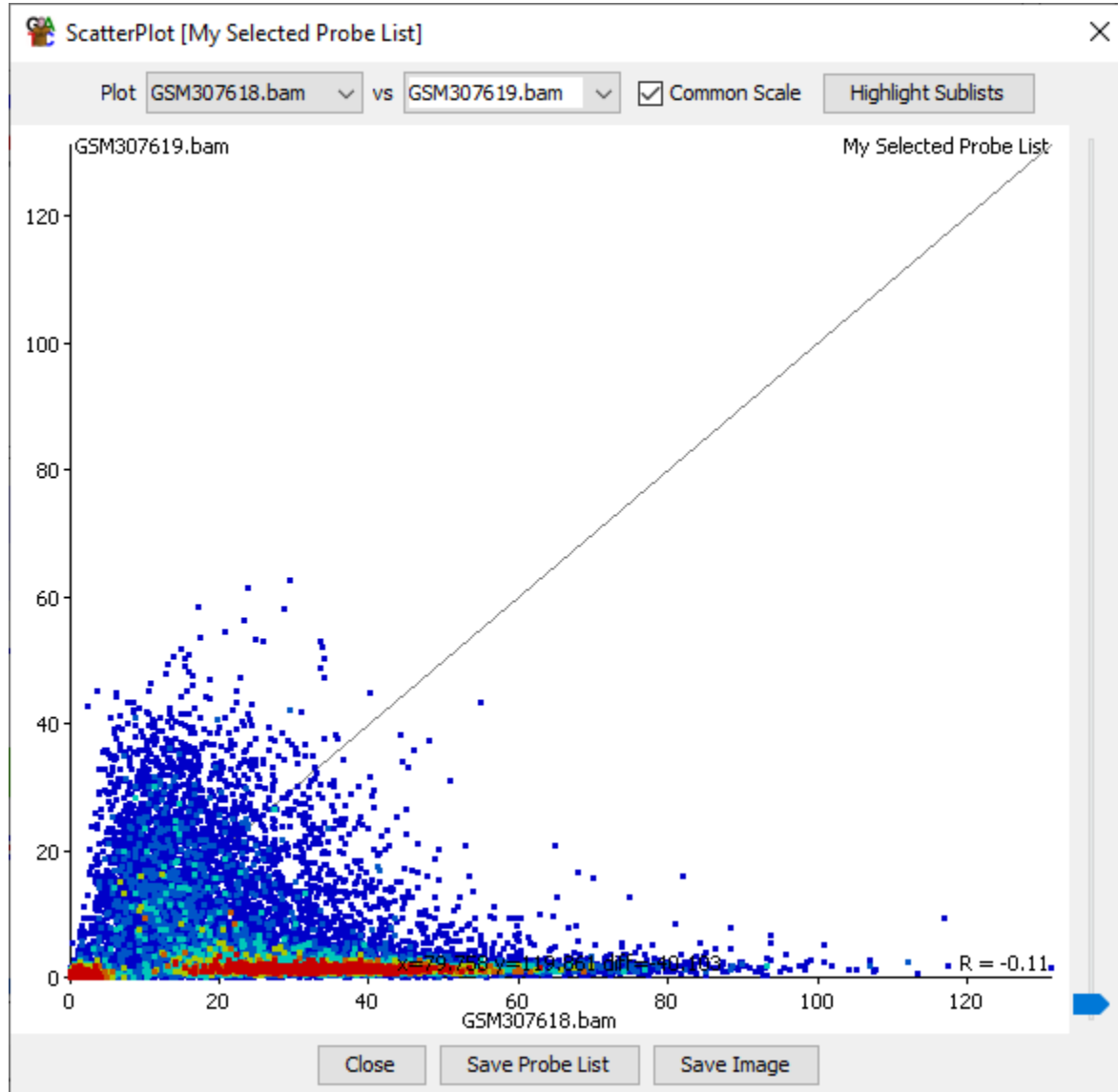
# Plotting

- By Default
  - Uses the data stores shown in the chromosome view
  - Uses the probes in the selected probe list





# Some Plots are Interactive



- Hover to see label
- Click to fix label
- Double click to show probe in Chromosome View
- Triple Click to clear labels

# Some Plots can be Duplicated

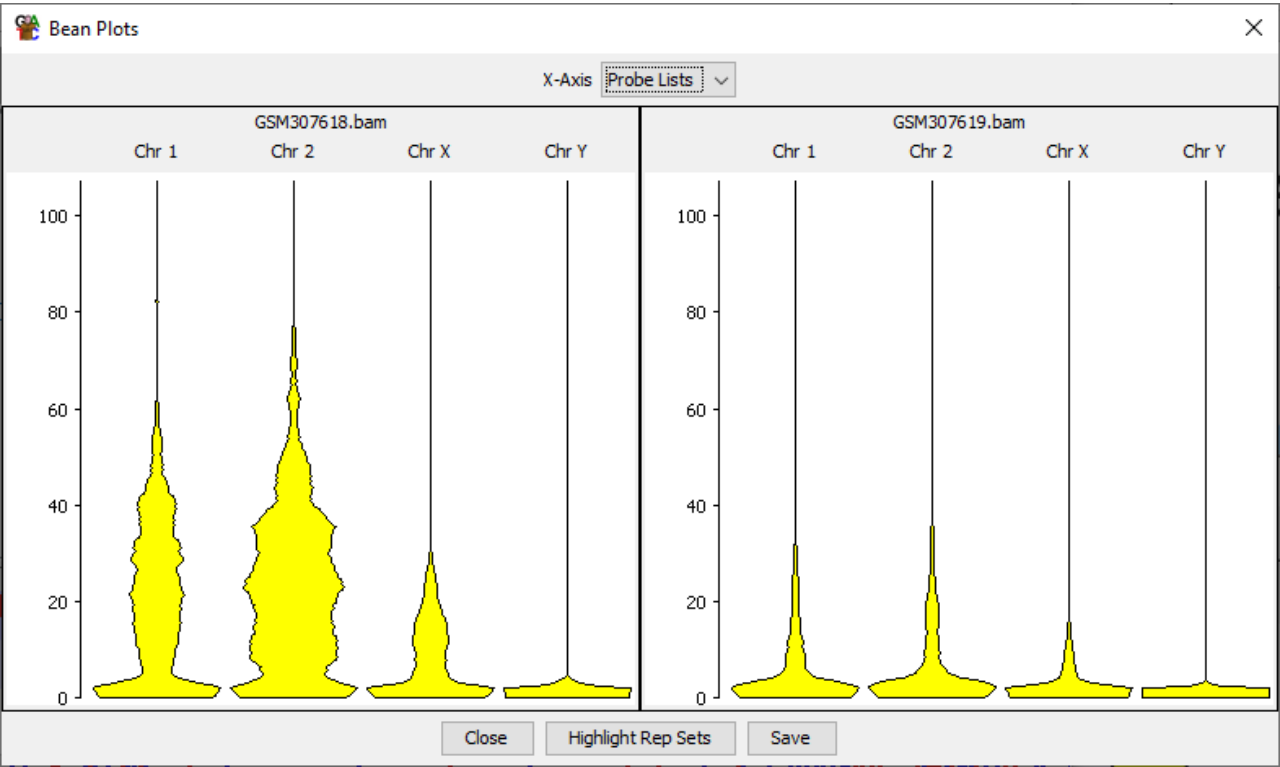
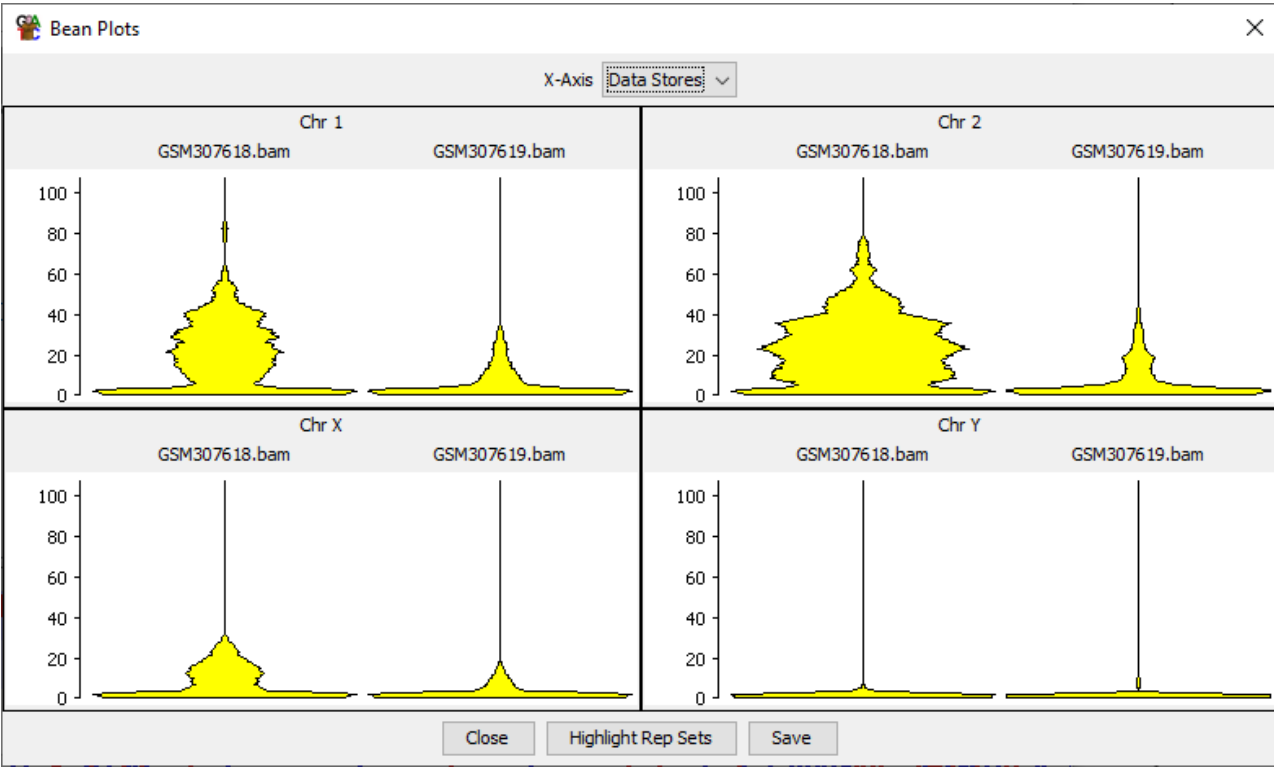
Bean Plot >

- Visible Data Stores...
- Multiple Probe Lists...
- Multiple Probe Lists and DataStores...

Select Probe Lists to use

- Probe Lists
  - CpG Islands (13840)
  - My Selected Probe List (13838)
    - Chr 1 (833)
    - Chr 2 (1058)
    - Chr X (554)
    - Chr Y (73)

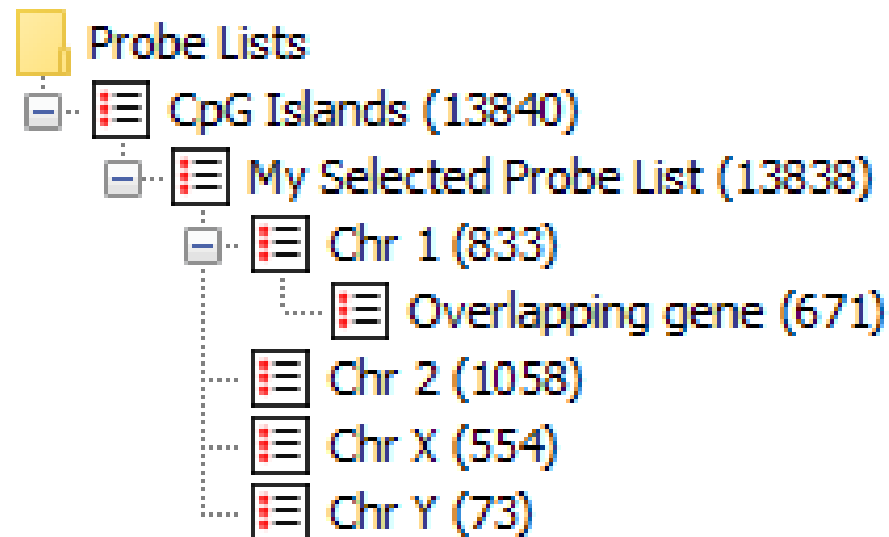
Close Select



# Filtering Probes

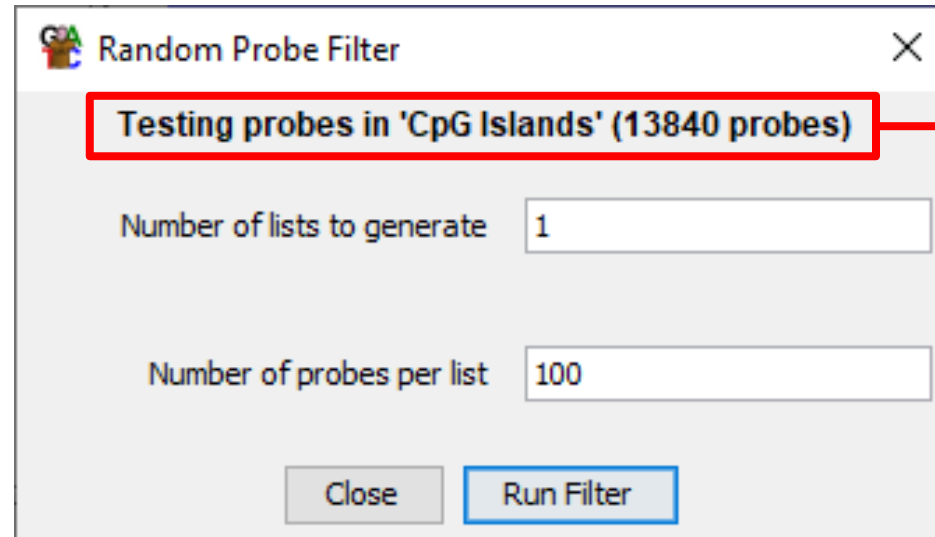
# Filtering Concepts

- Start from an existing Probe List
- Run a Filter to select a subset of those probes
- Create a new Probe List as a child of the original list
- Build up a tree of filtered Probe Lists



# Filters

- Filtering
  - Filter on Values >
  - Filter on Value Differences >
  - Filter on Variance >
  - Filter by Statistical Test >
  - Filter by Correlation >
  - Filter by Segmentation
  - Filter by Position...
  - Deduplication Filter...
  - Filter by Probe Length...
  - Filter by Features...
  - Filter by Feature names...
  - Filter by Probe names...
  - Filter Random Probes...
  - Duplicate Existing List
  - Combine existing lists >



Random Probe Filter

Testing probes in 'CpG Islands' (13840 probes)

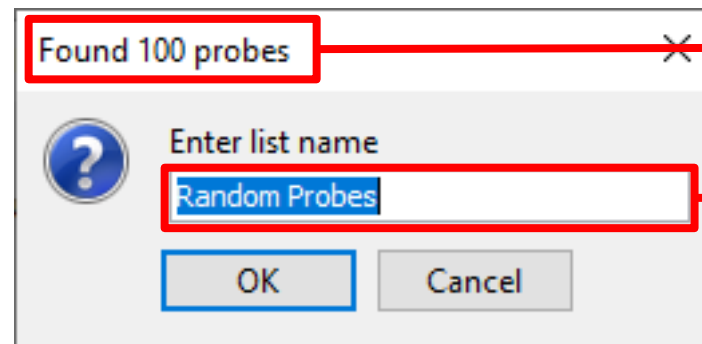
Number of lists to generate

Number of probes per list

Close Run Filter

**Starting List**

Check this is what you expect



Found 100 probes

Enter list name

Random Probes

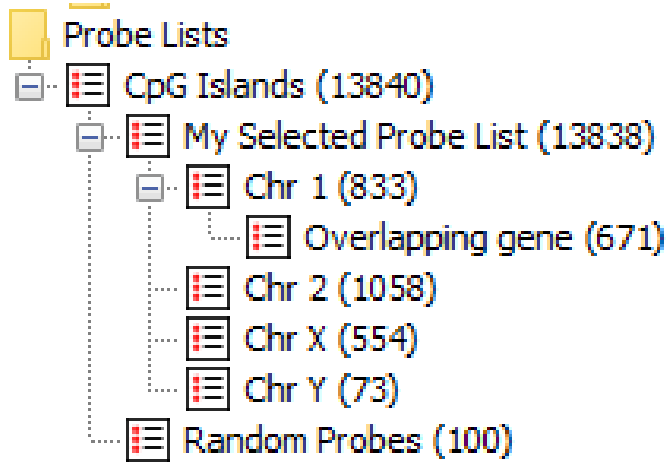
OK Cancel

**Number of passed probes**

**New List Name**

# Filter Details

- Rename list (doesn't change contents)
- View list (shows options used)
- Delete list (and any children)



- View List
- Make List Report
- Convert to annotation track
- Similar Lists
- Show Probe Length Histogram
- Rename
- Edit comments
- Delete

Random Probes (100 probes)

Description:  
[Random Probe Filter] A random subset of 100 probes from CpG Islands

Comments:

Probe	Chr	Start	End
oe = 0.91 Chr 1:309...	1	30947521	30951521
oe = 0.96 Chr 1:364...	1	36443092	36447092
oe = 0.78 Chr 1:545...	1	54555530	54559530
oe = 0.88 Chr 1:709...	1	70984541	70988541
oe = 0.91 Chr 1:977...	1	97768111	97772111
oe = 0.86 Chr 1:127...	1	127203285	127207285
oe = 0.87 Chr 1:135...	1	135281827	135285827
oe = 1.01 Chr 1:179...	1	179801482	179805482
oe = 0.75 Chr 2:180...	2	18037920	18041920

Close Save

# Saving, Reporting and Vistories

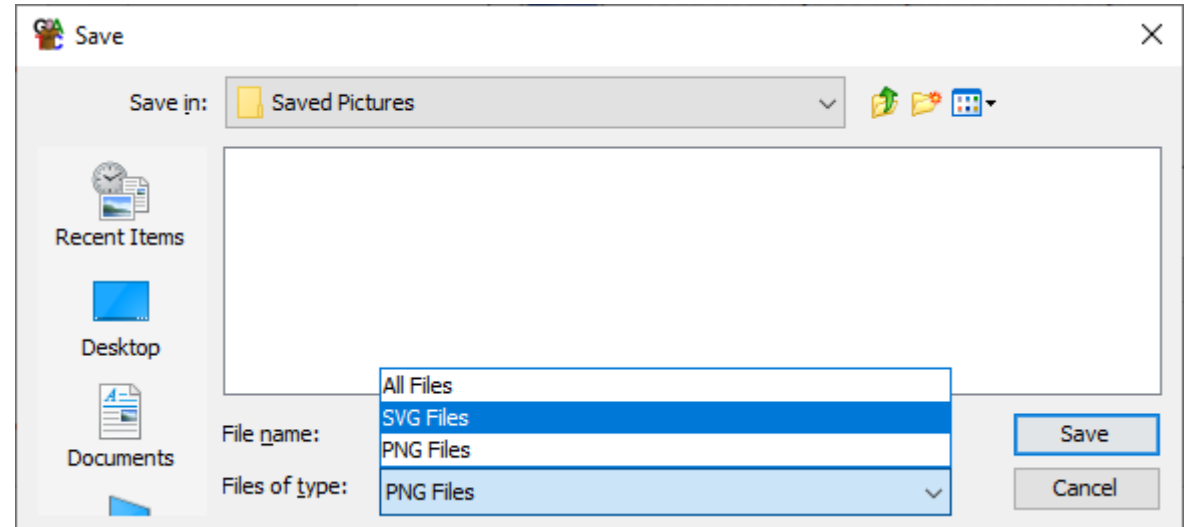
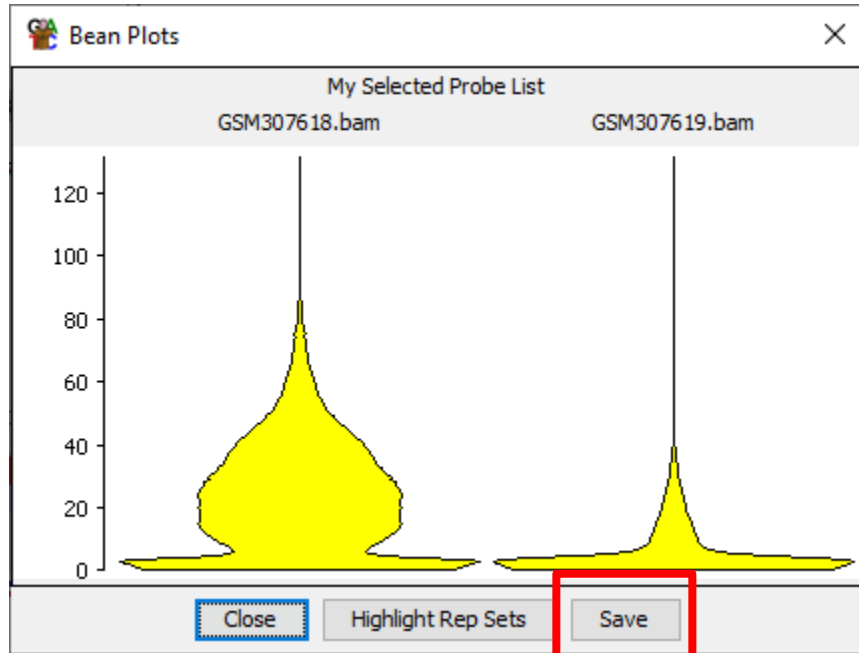
# Saving SeqMonk Projects

- File > Save Project
- Saves Everything
  - Data
  - Quantitations
  - Probes / Filters
  - Current View
- Single file with .smk file extension
- Can be moved to another machine and opened\*

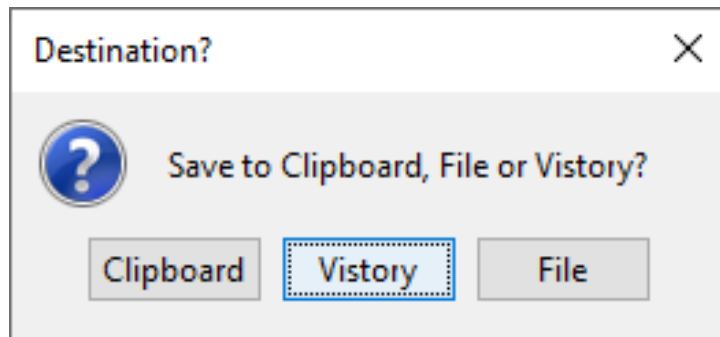
\*Unless using a custom genome – you need to copy that separately



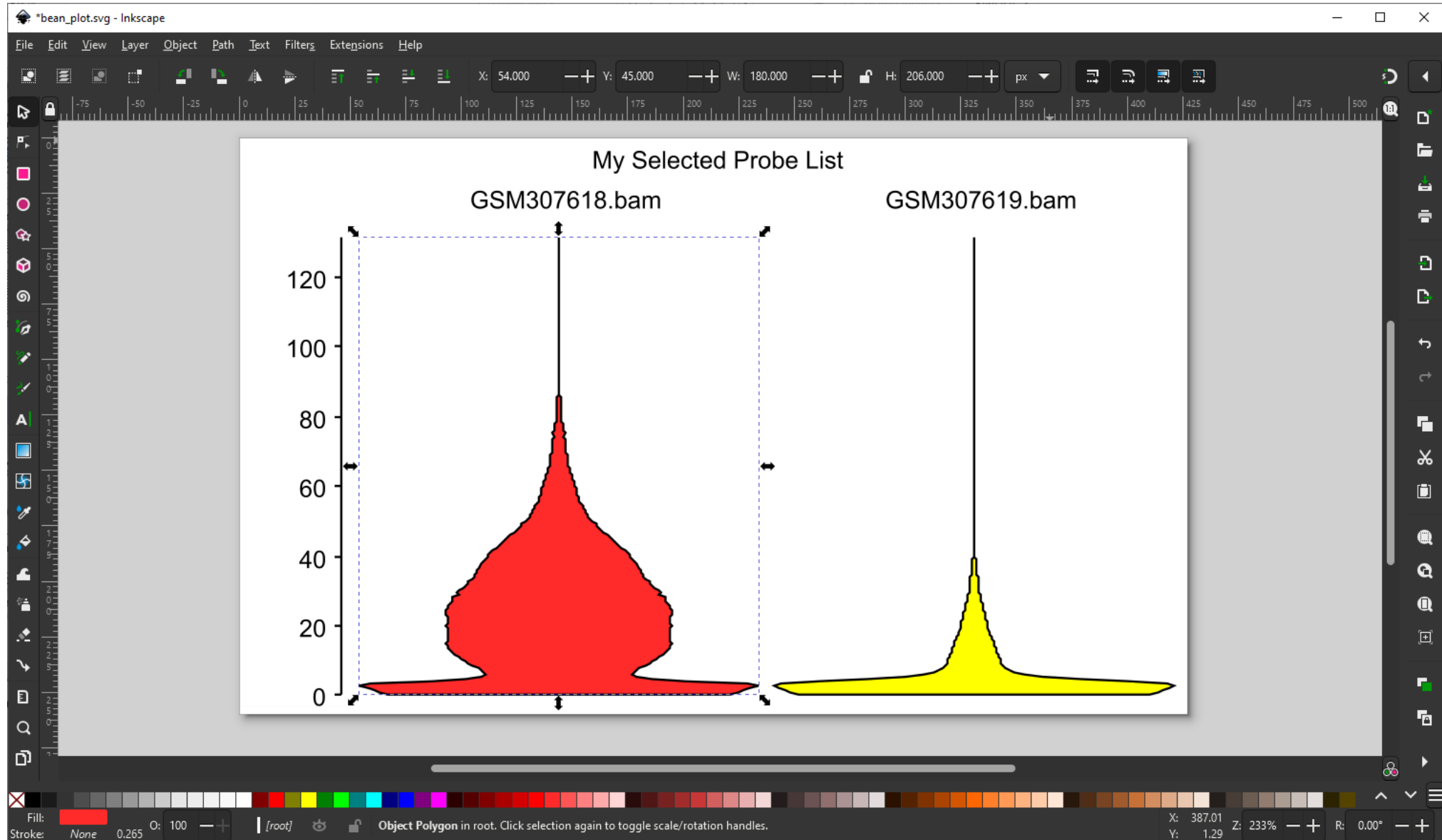
# Saving Images



- PNG – Bitmap – Screenshot
- SVG – Vector – Editable



# Editing SVG Images



# Creating Reports

- **Annotated Probe Report**
  - Generates a report for every probe in a probe list. Can annotate it with a feature from a chosen annotation track
- **Probe Group Report**
  - Generates a report from a probe list but can group together probes which are close to each other
- **Feature Report**
  - Generates a report for all features in an annotation track. Relates them to probes in a probe list
- **Data Store Summary**
  - Gives statistics about the data and quantitation in the currently selected probe list.

# Creating Reports

Options for Annotated Probe Report for My Selected Probe List

Reporting on probes in 'My Selected Probe List' (13838 probes)

Annotate with

Annotation distance cutoff  bp

unannotated probes

data for currently visible stores

0 annotations selected

- Sort by clicking headers
- Double click to change view
- Save to file (tab delimited text)

Annotated Probe Report for My Selected Probe List

Probe	Chromosome	Start	End	Probe Strand	Feature	ID	Description	Feature Str...	Type	Feature Ori...	Distance	GSM307618...	GSM30761...
oe = 0.70	1	3669388	3673388	+	Xkr4	ENSMUSG00...	X-linked Kx bl...	-	gene	overlapping	0	22.471	11.923
oe = 0.77	1	4490687	4494687	+	Sox17	ENSMUSG00...	SRY (sex det...	-	gene	overlapping	0	33.279	22.162
oe = 1.00	1	4495278	4499278	+	Sox17	ENSMUSG00...	SRY (sex det...	-	gene	overlapping	0	9.845	27.346
oe = 0.61	1	4558325	4562325	+						Not found	0	0.214	0
oe = 0.78	1	4565280	4569280	+						Not found	0	0.428	0.518
oe = 0.77	1	4569858	4573858	+						Not found	0	12.948	6.61
oe = 0.80	1	4783595	4787595	+	Mrpl15	ENSMUSG00...	mitochondrial...	-	gene	overlapping	0	49.223	0.778
oe = 0.92	1	4805831	4809831	+	Lypla1	ENSMUSG00...	lysophospholi...	+	gene	overlapping	0	27.073	0.648
oe = 0.81	1	4855919	4859919	+	Gm37988	ENSMUSG00...	predicted ge...	+	gene	overlapping	0	33.921	1.296
oe = 0.76	1	5017134	5021134	+	Rgs20	ENSMUSG00...	regulator of ...	-	gene	overlapping	0	27.073	25.142
oe = 0.82	1	5081288	5085288	+	Atp6v1h	ENSMUSG00...	ATPase, H+ ...	+	gene	overlapping	0	15.944	0.518
oe = 0.67	1	5248736	5252736	+						Not found	0	0.856	0.12

# Vistories

- A way to formally record your activities in SeqMonk
- Generates an HTML report
- Created automatically
- You can add commentary, images, reports and summaries
- Easy way to record and share your analysis

# Vistories

The screenshot shows the Vistory application window with a toolbar containing icons for Load, Save, Save As, Title, Subtitle, Text, Paste Image, Clear, Summary, and Save HTML. The main content area displays a list of operations:

- DataSet Added**  
GSM307618.bam E:\Illumina Analysis\SeqMonk\_Course\_Data\GSM307618.bam Library type Single End Dedup=No MAPQ>=20 Primary alignments only.
- DataSet Added**  
GSM307619.bam E:\Illumina Analysis\SeqMonk\_Course\_Data\GSM307619.bam Library type Single End Dedup=No MAPQ>=20 Primary alignments only.
- New Probe Set:CpG Islands (13840 probes)**  
Feature generator using CpG islands duplicates removed Centered on feature from 2000-2000
- Probes Quantitated**  
Read Count Quantitation using All Reads correcting for total count per million reads
- New Probe List: My Selected Probe List (13838 probes)**  
[Probe Values Filter] Filter on probes in CpG Islands where exactly 2 of GSM307618.bam , GSM307619.bam had a value below 200.0. Quantitation was Read Count Quantitation using All Reads correcting for total count per million reads
- New Probe List: Chr 1 (833 probes)**  
[Position Filter] Probes from My Selected Probe List which are on chromosome 1 on strand All Probes
- New Probe List: Chr 2 (1058 probes)**  
[Position Filter] Probes from My Selected Probe List which are on chromosome 2 on strand All Probes
- New Probe List: Chr X (554 probes)**  
[Position Filter] Probes from My Selected Probe List which are on chromosome X on strand All Probes
- New Probe List: Chr Y (73 probes)**  
[Position Filter] Probes from My Selected Probe List which are on chromosome Y on strand All Probes
- New Probe List: Overlapping gene (671 probes)**  
[Feature Filter] Filter probes in Chr 1 on regions based on gene Over feature relationship is Overlapping
- New Probe List: Random Probes (100 probes)**  
[Random Probe Filter] A random subset of 100 probes from CpG Islands

The screenshot shows the Vistory application window with the same toolbar as the first image. The main content area displays an example workflow:

### Example Vistory

This is an example which shows what you can do with vistories.

#### Load the data

- DataSet Added**  
GSM307618.bam E:\Illumina Analysis\SeqMonk\_Course\_Data\GSM307618.bam Library type Single End Dedup=No MAPQ>=20 Primary alignments only.
- DataSet Added**  
GSM307619.bam E:\Illumina Analysis\SeqMonk\_Course\_Data\GSM307619.bam Library type Single End Dedup=No MAPQ>=20 Primary alignments only.

#### Make Probes

- New Probe Set:CpG Islands (13840 probes)**  
Feature generator using CpG islands duplicates removed Centered on feature from 2000-2000

#### Quantitate

- Probes Quantitated**  
Read Count Quantitation using All Reads correcting for total count per million reads

#### Filter

- New Probe List: My Selected Probe List (13838 probes)**  
[Probe Values Filter] Filter on probes in CpG Islands where exactly 2 of GSM307618.bam , GSM307619.bam had a value below 200.0. Quantitation was Read Count Quantitation using All Reads correcting for total count per million reads
- New Probe List: Chr 1 (833 probes)**  
[Position Filter] Probes from My Selected Probe List which are on chromosome 1 on strand All Probes
- New Probe List: Chr 2 (1058 probes)**  
[Position Filter] Probes from My Selected Probe List which are on chromosome 2 on strand All Probes
- New Probe List: Chr X (554 probes)**  
[Position Filter] Probes from My Selected Probe List which are on chromosome X on strand All Probes
- New Probe List: Chr Y (73 probes)**

# Vistories

**Plotting**  
We can demonstrate how to embed plots into the vistory

My Selected Probe List

GSM3076 18.bam	GSM3076 19.bam

**Reporting**

DataStore	Total R...	Forwar...	Revers...	Unknow...	Mean R...	Total R...	Fold Co...	Total Q...	Median ...	Mean Q...	Valid
GSM3076...	9345214	4674124	4671090	0	33	317580516	0.11652...	318641.2...	20.973303	23.026543	13838
GSM3076...	7716069	3858872	3857197	0	28	216436208	0.07941...	80898.19...	1.9439951	5.84609	13838

SeqMonk Vistory Report

file:///C:/Users/andrewss/Pictures/Saved Pictures/example\_vistory

**SeqMonk Vistory**

- Top
- Example Vistory
- Load the data
- Make Probes
- Quantitate
- Filter
- Plotting
- Reporting

**Example Vistory**

This is an example which shows what you can do with vistories.

**Load the data**

**DataSet Added**  
GSM307618.bam E:illumina Analysis\SeqMonk\_Course\_Data\GSM307618.bam Library type Single End Dedup=No MAPQ>=20 Primary alignments only.

**DataSet Added**  
GSM307619.bam E:illumina Analysis\SeqMonk\_Course\_Data\GSM307619.bam Library type Single End Dedup=No MAPQ>=20 Primary alignments only.

**Make Probes**

**New Probe Set:CpG islands (13840 probes)**  
Feature generator using CpG islands duplicates removed Centered on feature from 2000-2000

**Quantitate**

**Probes Quantitated**  
Read Count Quantitation using All Reads correcting for total count per million reads

- Save as .smv file (editable)
- Save as HTML (report)

https://www.bioinformatics.babraham.ac.uk/vistorydb/

**Vistory DB**  
A collection of SeqMonk Vistory files

About

### Filters

Type search term here

ngs(4) chip-seq(1) peak calling(1)  
annotation(1) promoter(1) rna-seq(2)  
statistics(3) qc(2) visualisation(2)  
reporting(1)

### Vistories

#### Calling Peaks from Replicated ChIP Data

In replicated ChIP-Seq datasets there are a few different ways to call peaks. We go through a few options for how to call peaks, explaining the differences between them and the strengths and weaknesses of each

ngs, chip-seq, peak calling

#### Calling Peaks on ChIP data with Replicates

In this vistory I'm going to look at calling peaks on some replicated ChIP data. We have two replicates and a single input and we're going to look at a couple of options for how to build a peak list from them.

At the end we want to have a single set of probes which are the peaks we would take forward for further analysis.

At the start we've just imported the data and have done a simple running window quantitation. We can see that we have strongly enriched peaks in both of the replicates and that in general the peak positions look similar.

### Top

- Calling Peaks on ChIP data with Replicates
- Project Summary
- Basic Project Info
- Data Sets
- Quantitation
- Strategies for multiple peak calling
- Creating a blacklist
- Calling Peaks
- Merged Data
- Individual Peak Calls
- Combining Peaks

SeqMonk Vistory