Babraham Bioinformatics

Exercises:
Exploring ChIP-Seq data

Version 2018-03-3

Licence

Introduction

In this session we will go through how to visualise and explore the ChIP data we processed in the first exercise.

Software

The software which will be used in this session is listed below. Software which requires a linux environment is indicated by an asterisk*:

- SeqMonk (http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/)

Data

The data in this practical comes from ENCODE accession ENCSR260AKX
C. elegans genome data and GTF models come from Ensembl (https://www.ensembl.org/Caenorhabditis_elegans/)

Exercise 1: Visualisation and Exploration

After mapping our data we're going to look at it in SeqMonk and work out the nature of the enrichment and find any quality issues in the data which might not have been evident from the raw read level.

Step 1.1 Importing mapped data into SeqMonk

Open SeqMonk on your machine either by double clicking the SeqMonk icon or by typing '`seqmonk`' in a linux command shell.

We are going to create a new project into which we can import the BAM files we generated before. This needs to use the same version of the Worm genome that we mapped the data to. To start the project select

File > New Project

The genome we need is *Caenorhabditis elegans* / WBcel235_v91. If you don't see this in your initial list of available genomes then you will need to use the "Import Genome from Server" option to copy the genome onto your machine.

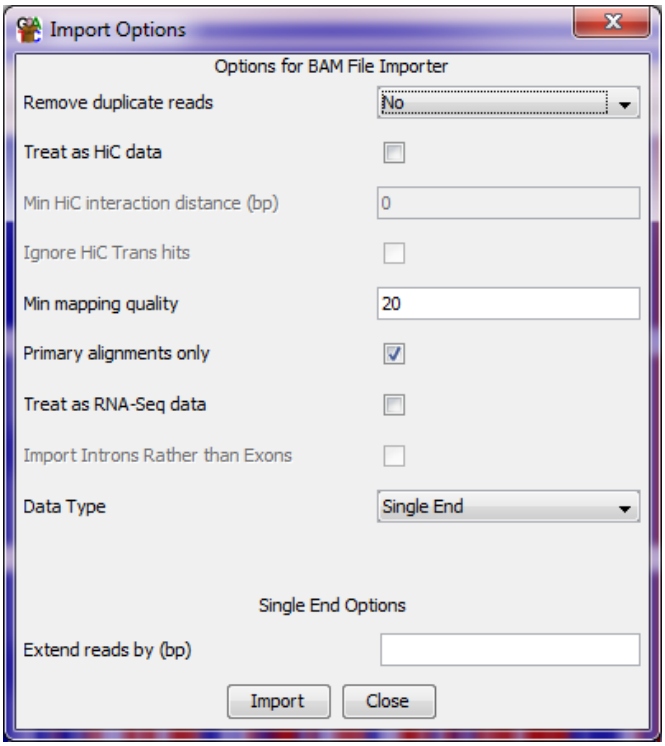
Once the project is open you can import your BAM files by selecting:

File > Import Data > BAM/SAM

Then selecting the 3 BAM files you made in the earlier part of the exercise.

The import options should have been detected automatically, but check the options which have been set to make sure they make sense.

These default options will cause some filtering of the data to be carried out. The minimum mapping quality of 20 means that only high quality reads will be imported.



We won't extend the reads at this point, so you can see what the raw positions look like, but in a real analysis you would probably want to extend your single end reads up to the insert size in your library during import.

Step 1.2 Examination of the raw data

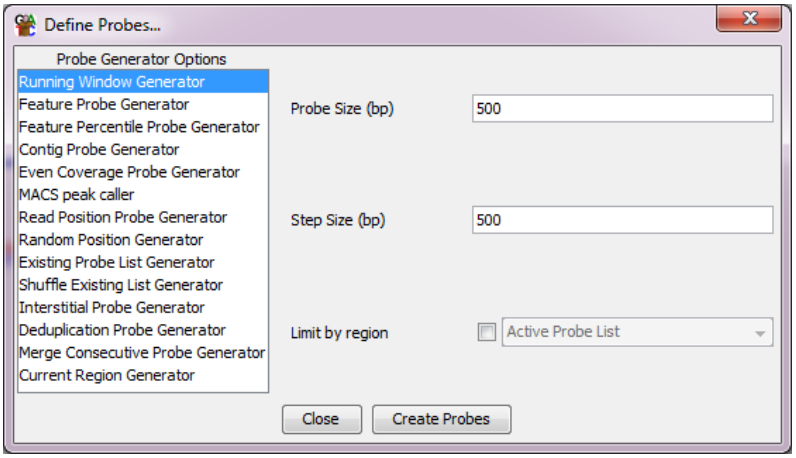
In any sequencing analysis the first step in your analysis should be to spend a couple of minutes looking through the raw data to see if you can see any obvious problems. Look around the genome and see if you think this data is OK. Questions to focus on would be:

- Do we see enrichment in any of the samples? If so is it the ChIP or Input?
- Is there any evidence of enrichment in the input?
- Is enrichment similar between the two ChIP replicates?
- What is the nature of the enrichment – sharp peaks or broad regions?
- What is the strand bias of the reads around enriched regions? Does this make sense?
- If we were to extend the reads, how long an extension should we use?
- Is there evidence of technical duplication (PCR artefacts)?

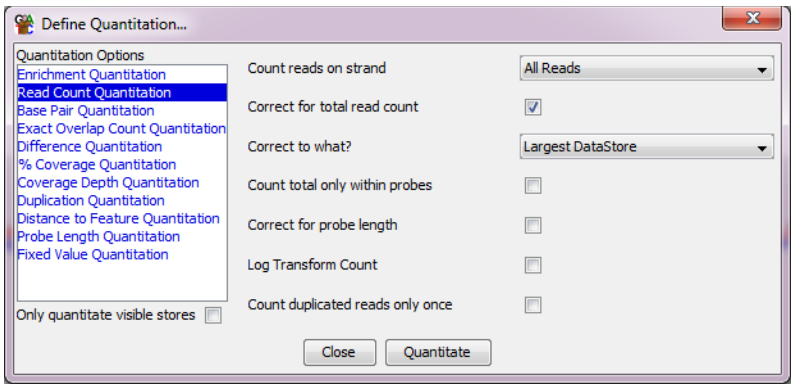
Step 1.3 Quantitation

To get a more quantitative view of our data we can do a simple running window quantitation of the whole genome.

To make the probes we use **Data > Define Probes > Running Window Probes**. We will make 500bp probes tiled end to end across the genome.



For the quantitation (which should open automatically) we will do a simple globally normalised linear quantitation. Note that you need to turn off the option to log transform.



You can now look at a quantitative view across your genome. You might need to adjust the data zoom (**Control + Z**) to get the tops of the peaks to be visible.

Step 1.4 QC

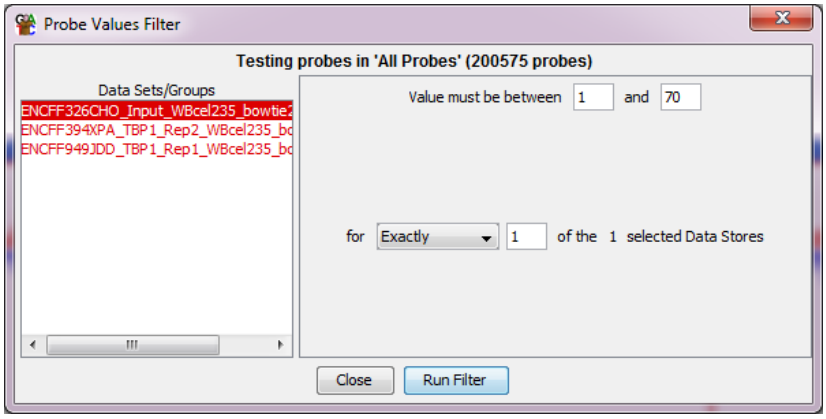
We can do some initial checks on the quality of the data to see if it all looks good. Firstly we can look at the distribution of values in the input sample, so that we can identify and remove any places which have unusually high input signal.

To do this select the input sample in the data view (top left window) and then do **Plots > Probe Value Histogram**. This will show the range of values in the input. Due to some large outliers you will need to zoom in (by clicking and dragging) to the far left of the plot.

Have a look at what the normal range of values is in the input.

Next we will filter out probes with unusual input values. Hopefully you saw that by the time you hit a quantitated value of 70 then there were virtually no regions in the genome with that level of signal. We'll also get rid of regions with no data at all in the input as they're probably just holes in the assembly.

To do this go to **Filtering > Filter on Values > Individual Probes**. We'll select probes with a values between 1 and 70 in the input.



You can call the list which is generated "Sensible Input". You can see it by expanding the "All Probes" list in the data view. Select this newly created list and see how many probes were removed by this filtering.

Other things we want to check are the distributions of values. To do this we are going to use:

Plots > Cumulative Distribution Plot > Visible Data Stores

Plots > QQ Distribution > Visible Data Stores

Try running these plots from both the All Probes and Sensible Input lists to compare the results. Do you see obvious evidence of enrichment, and is this consistent between the two ChIP replicates?

We can also check the duplication of our samples. To do this start from the "Sensible Input" list and then select:

Plots > Duplication Plot

If we see universally low duplication, or if the duplication is proportional with the read density then we are fine. If we see high duplication not related to high overall read density then we potentially have a PCR duplication problem.

Do these samples look like they have a duplication problem?

Finally we can use a scatterplot to compare the signal in the input with the ChIP samples, and the two ChIPs with each other to see whether there is enrichment and how consistent it is.

Plots > Scatterplot

Is there obvious enrichment over input?

Is the enrichment level for the two ChIPs consistent?

Step 1.5 Profiling positional enrichment

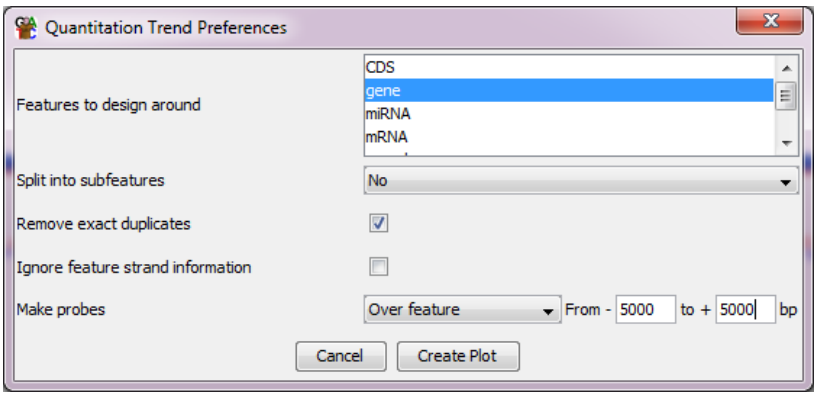
In the final part of the exploration of this data we will look to see whether we need to do peak calling on our data or whether we can identify regions of the genome defined by annotation features which will capture the enrichment efficiently.

Before we do an automated analysis of the enrichment positions, have a look at the data manually and see if you can see a relationship between the enrichment position and the gene features in the data.

To look at the positional enrichment of our data we will use a quantitation trend plot which looks at the average signal over different parts of a type of feature.

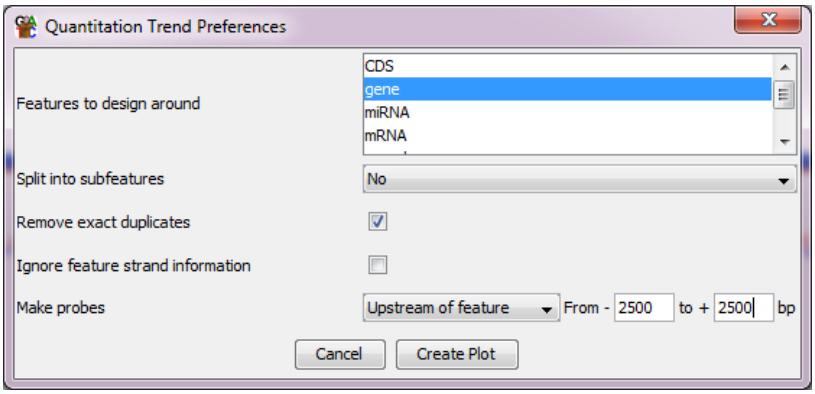
Plots > Quantitation Trend Plot

We can start by looking at the signal over the whole of a gene +/- 5kb



What does this suggest about the location of the enrichment? Could the signal here be confounded by the size of genes we commonly see in worm?

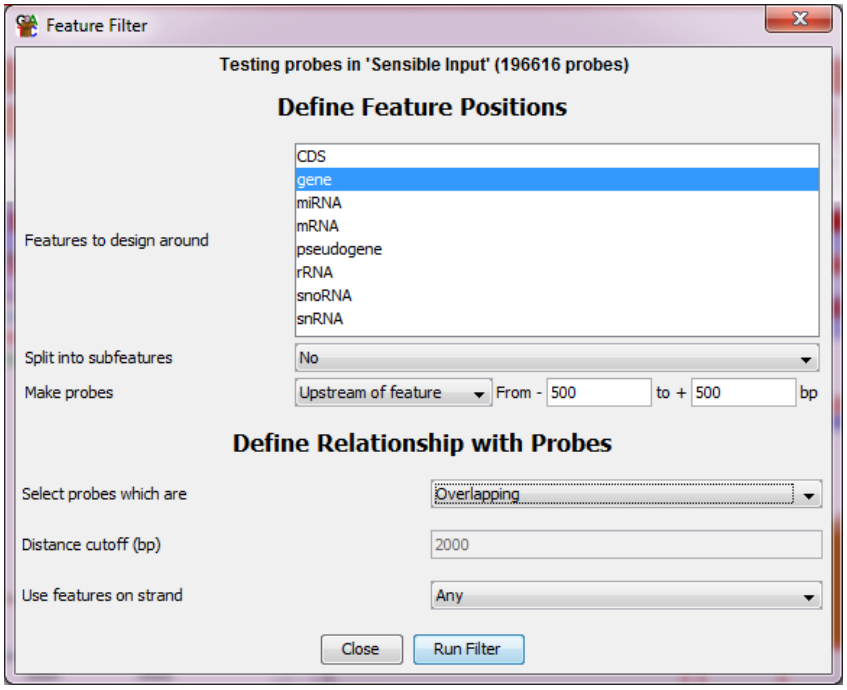
We can do a second plot looking specifically at the promoters of genes.



Does this change our view about where the enrichment is?

It looks like promoters might be a good candidate for defining our enrichment. We can see whether restricting our analysis to promoters will capture the vast majority of the signal in our data. To do this we'll filter out all of the probes which overlap promoters (start of gene +/- 500bp), and then see whether there is significant enrichment outside those regions.

Filtering > Filter by Features

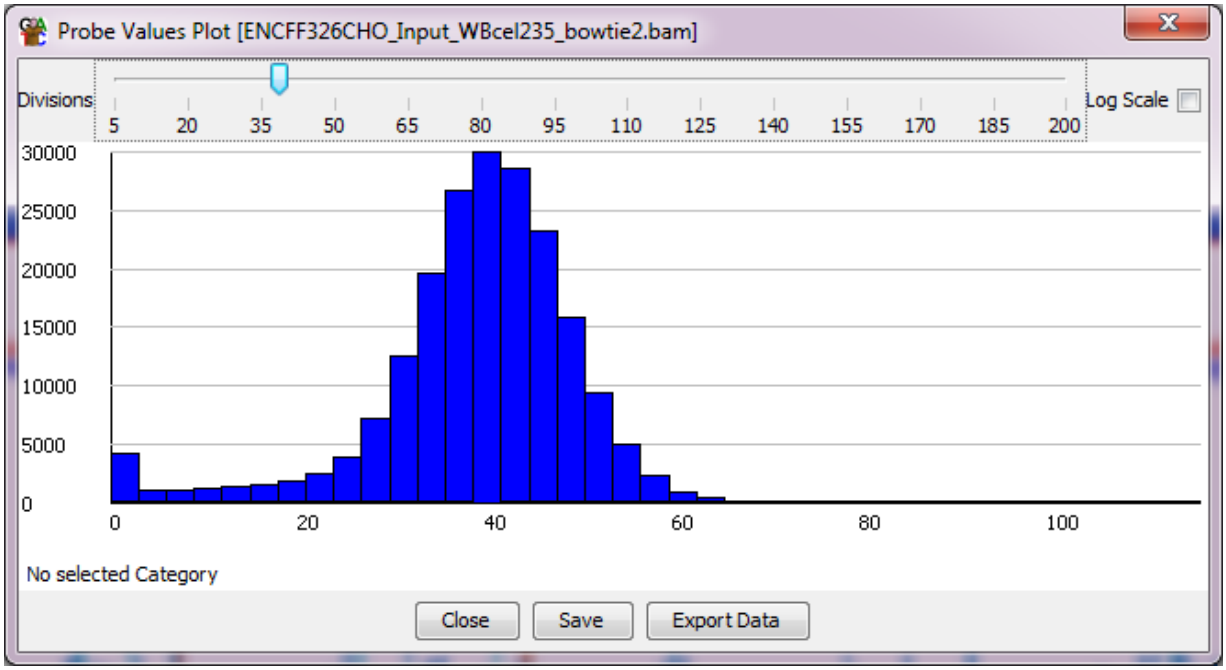


Call the list you generate "Promoter Regions" it should be a sub-list of "Sensible Input". You can repeat the filter but selecting "Not overlapping" as the relationship with the feature to get the non-promoter probes.

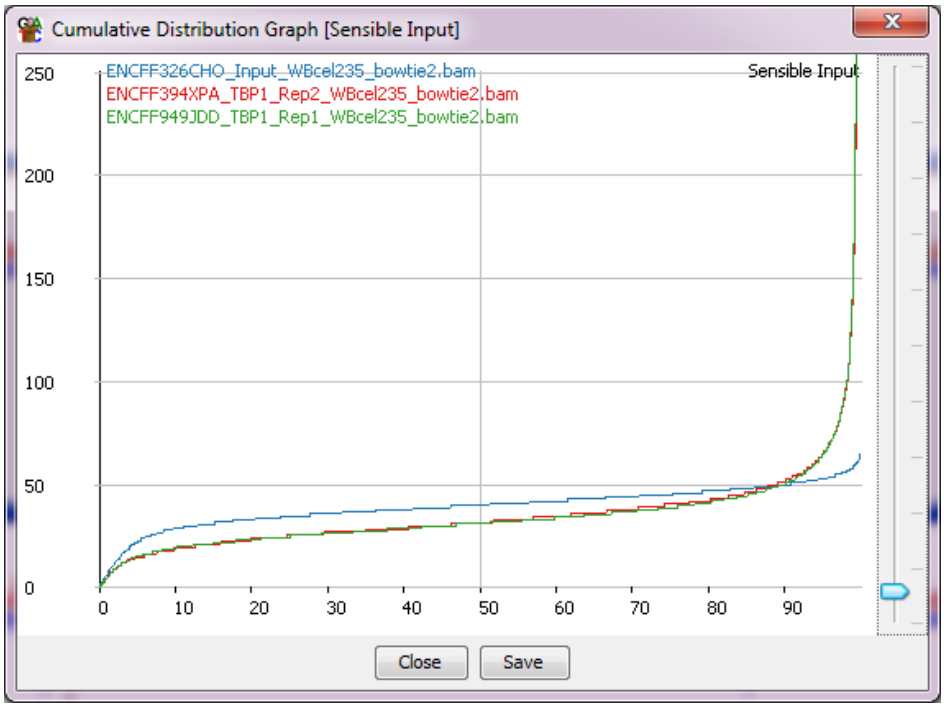
To see how well the promoters capture the enrichment signal we can plot out a scatterplot of the "Sensible Input" list for the input vs one of the ChIP samples (it doesn't really matter which one). We can then use the "Highlight Sublists" option to highlight on the scatterplot the promoter and non-promoter regions. You can turn off the "Common Scale" option to make better use of the available space in the plot.

Does this suggest that an analysis of promoters might be an effective way to analyse this data?

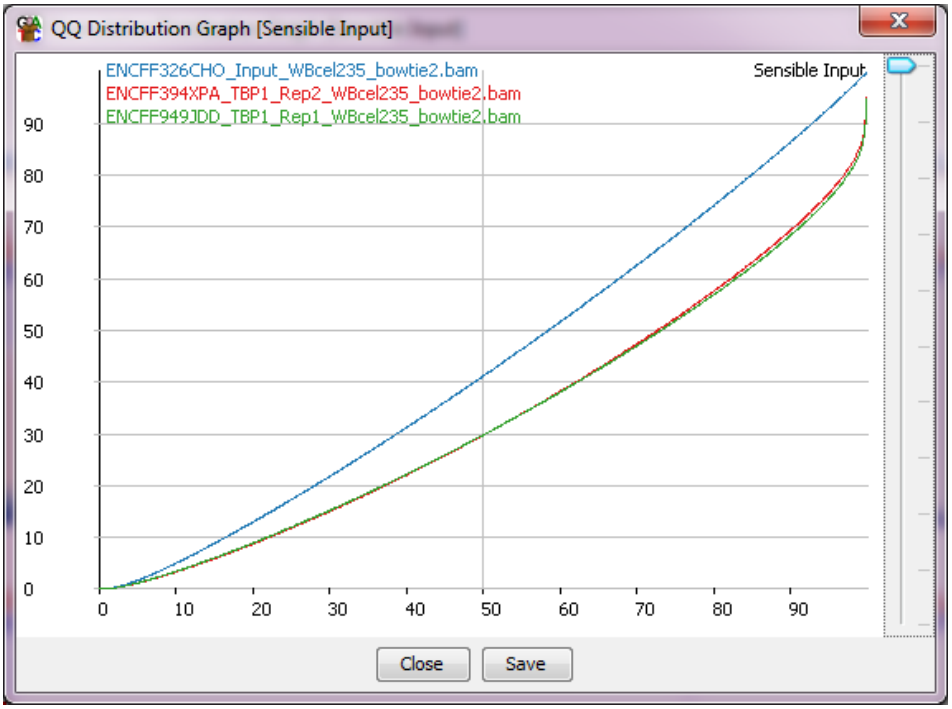
Step 1.4 Input Value Range Histogram



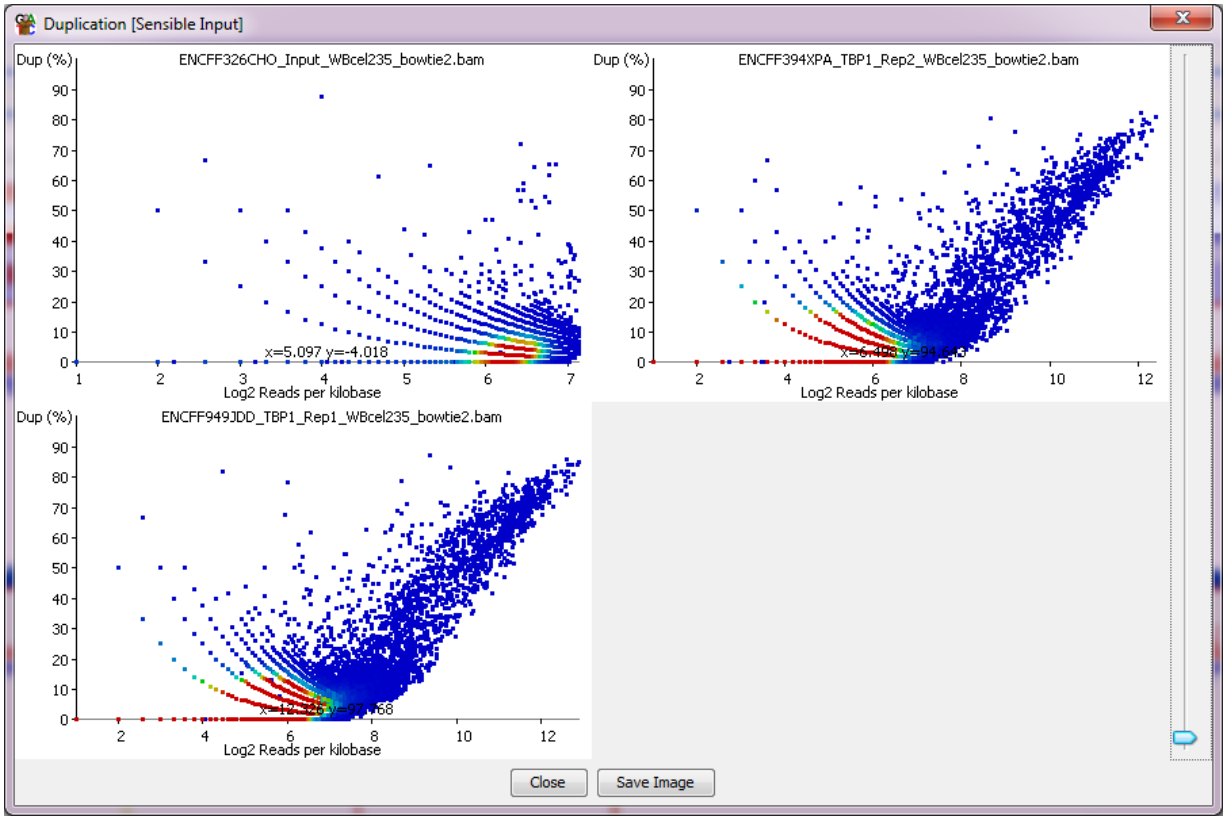
Step 1.4 Cumulative distribution plot



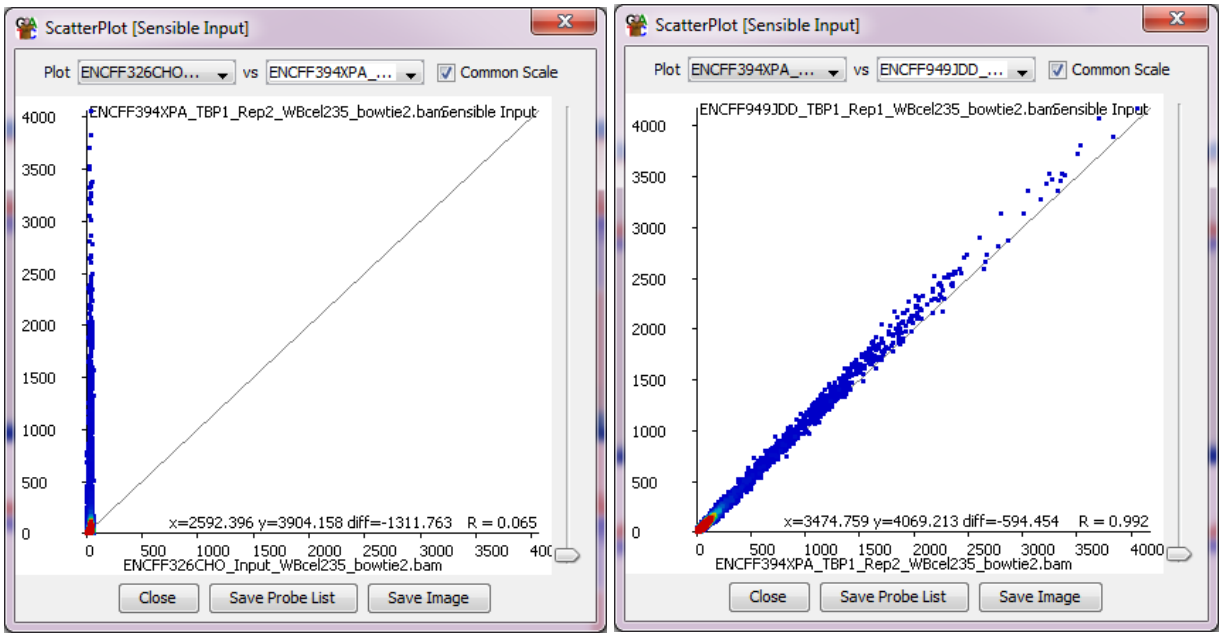
Step 1.4 QQ plot



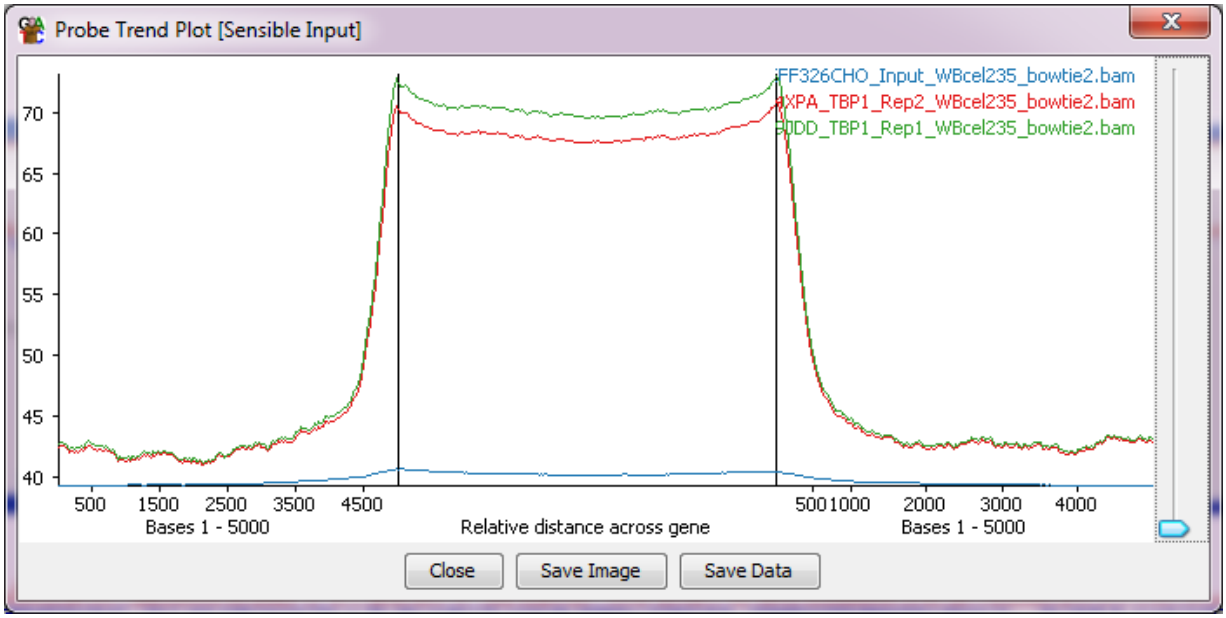
Step 1.4 Duplication plot

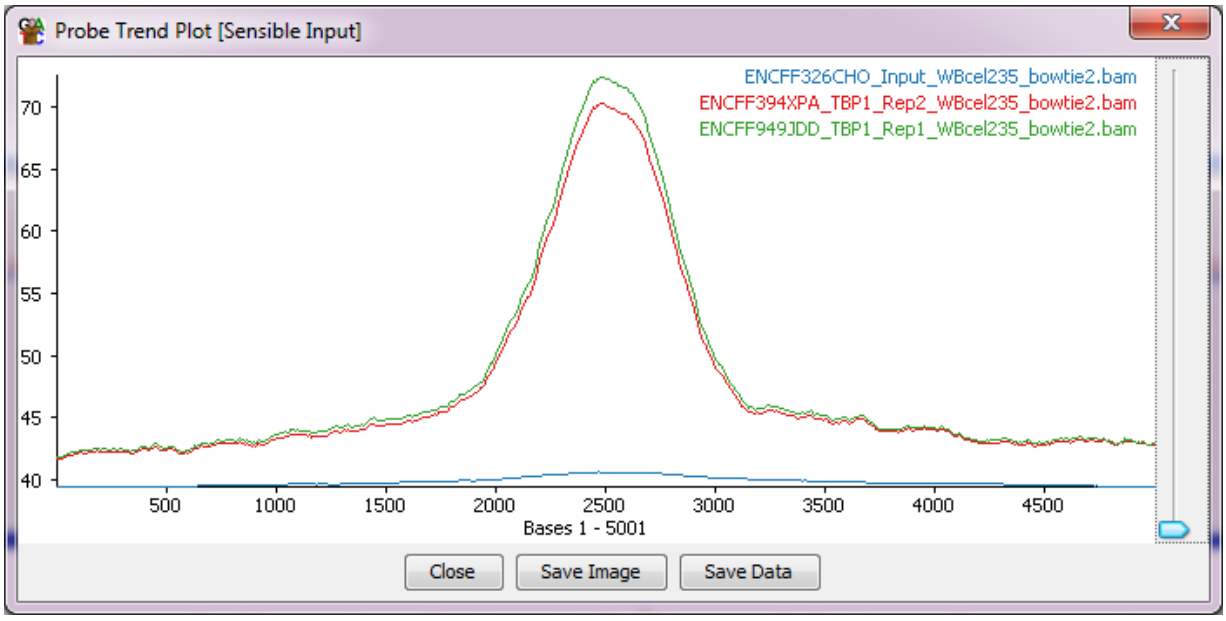


Step 1.4 Scatterplots



Step 1.5 Positional Enrichment





Step 1.5 Scatterplot highlighting promoters