

# **Exercises: Visualising and Exploring BS-Seq**

## Licence

This manual is © 2014-17, Simon Andrews.

This manual is distributed under the creative commons Attribution-Non-Commercial-Share Alike 2.0 licence. This means that you are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following conditions:

- Attribution. You must give the original author credit.
- Non-Commercial. You may not use this work for commercial purposes.
- Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under a licence identical to this one.

Please note that:

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.
- Nothing in this license impairs or restricts the author's moral rights.

Full details of this licence can be found at

<http://creativecommons.org/licenses/by-nc-sa/2.0/uk/legalcode>

## Introduction

In this session we will go through the initial steps of data import and QC followed by methylation quantitation and general exploration. The practical starts from the data produced by the Bismark methylation extractor and uses the graphical SeqMonk program for the analysis.

The basic steps we're going to include are:

- Importing Bismark methylation calls into SeqMonk
- Creating data groups to get a single data set per sample
- Viewing the raw data
- Finding and removing coverage outliers
- Unbiased methylation visualisation and quantitation
- Comparing methylation between samples
- Targeted quantitation and comparison
- Constructing trend plots

## Software

This practical is based on SeqMonk which is a cross platform graphical application. The instructions given here should work on any platform supported by SeqMonk (windows, mac, unix).

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

## Data

The data in this practical comes from GEO accession GSE56879. Specifically, the samples being used are the bulk data for MII Oocytes (GSM1370534) and Bulk Serum ESCs (GSM1370575). All of the processed data using in this practical can be downloaded from the Babraham Bioinformatics web site (<http://www.bioinformatics.babraham.ac.uk/training.html>).

## Exercise 1 – Importing Data

The data we're using for this section is in the Visualising\_and\_Exploring subfolder of your main data folder. You should be able to see two files listed there:

`CpG_NCBIM37_Oocyte_bismark_pe.deduplicated.cov.gz`

`CpG_NCBIM37_Serum_bismark_pe.deduplicated.cov.gz`

In this exercise we're only going to look at the methylation calls in CpG context, which is often the most pragmatic solution to the problem of how to cope with the huge amounts of data produced by a full BS-Seq run.

To load this data into SeqMonk you first need to create a project into which the data can be imported.

Projects are based around the genome assembly to which the data was mapped, so in this case you need to create a project based on the *Mus musculus* NCBIM37 genome.

[File > New Project](#)

If you can see NCBIM37 listed then select it, if not then press the “**Import**” button, download it, then open it.

Once the project is open you can import the data. The Bismark data is a coverage file format which SeqMonk understands so you can just directly import it with:

[File > Import Data > Bismark \(cov\)](#)

From the dialog box which opens select both coverage files from the Visualising\_and\_Exploring folder. Press OK and the data should import.

Once the data has loaded you should be able to see the methylation calls in the chromosome view. Have a look at the raw data to see the structure of the calls. You should be able to see that the calls occur at distinct positions and that there are repeated measures at the same position.

Try zooming out so you can see large chunks of a chromosome and look to see if you can find regions where the coverage suddenly increases. These are likely sites where data has been mis-mapped and the calls in these regions cannot be trusted.

## Exercise 2: Identifying Coverage Outliers

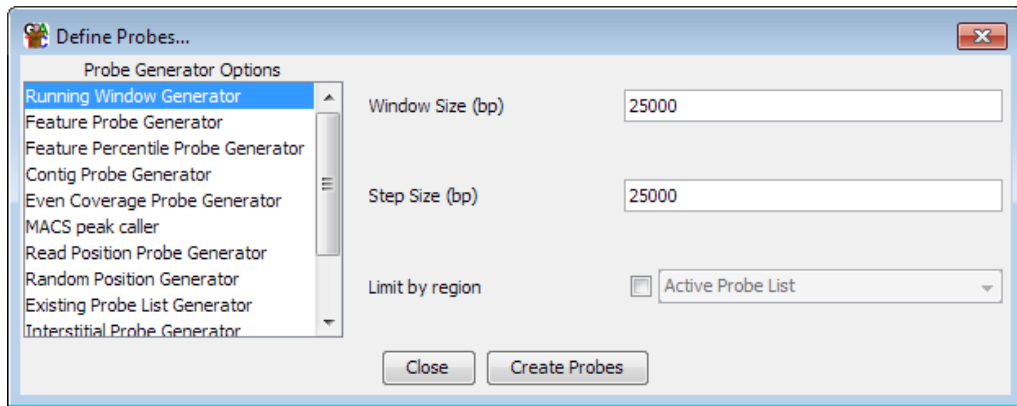
To identify parts of the genome whose coverage is so high that the calls in those regions are immediately suspect we're going to do a coarse quantitation followed by a simple detection of outliers.

To get a quick impression of the coverage we can make windows of 25kb over the whole genome and count the number of calls in each window.

To make the 25kb windows select:

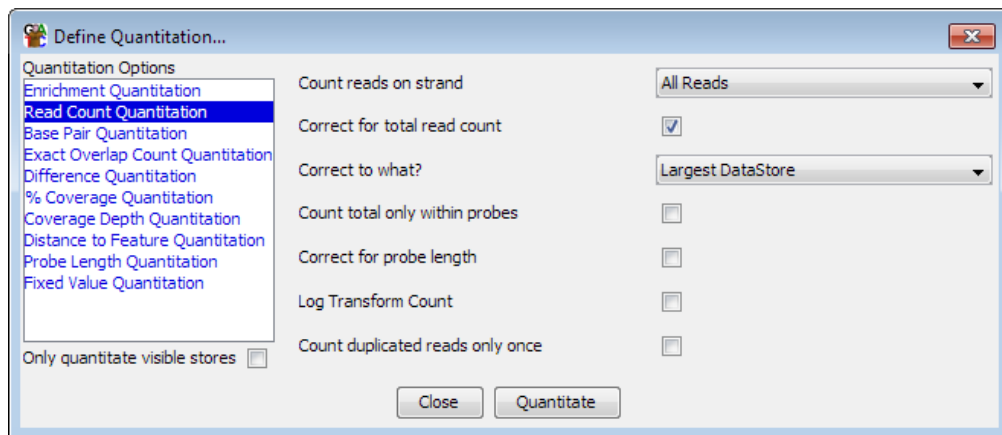
[Data > Define Probes > Running Window Generator](#)

Set both the window size and the step size to 25000, then press “**Create Probes**”.

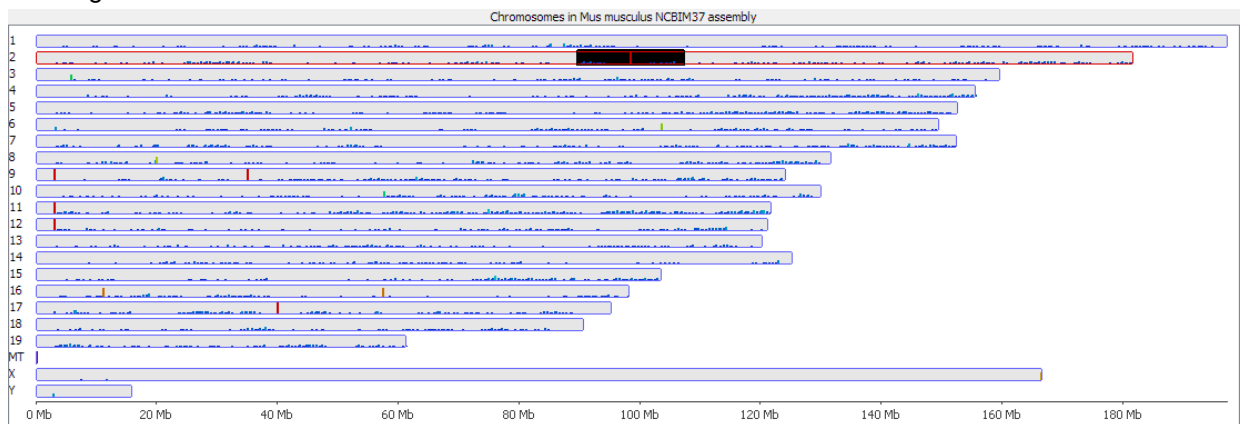


A new “Define Quantitation” dialog should automatically display (if it doesn’t then you can use [Data > Quantitate Existing Probes](#) to bring it up).

To quantitate the number of calls in each window select “[Read Count Quantitation](#)” and count all reads, correcting for total read count, but turn **OFF** the log transformation.

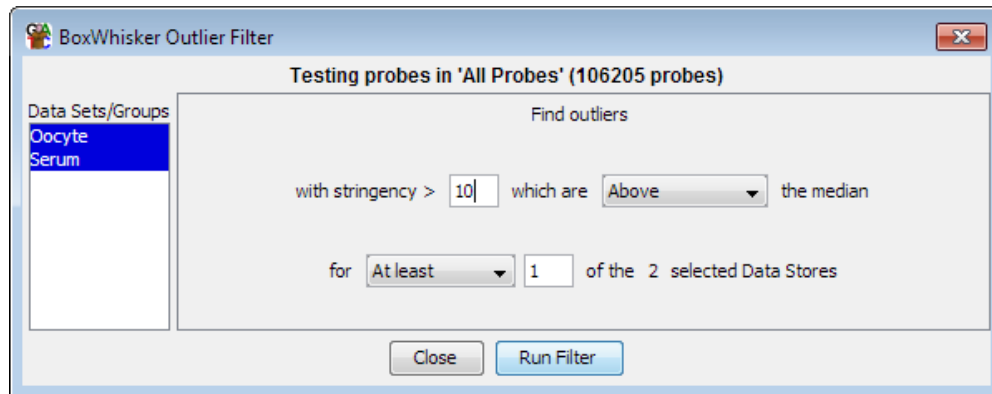


You can now look at your quantitated data in the chromosome view. If you select one of your data groups in the Data Panel (top left) then you can also see the quantitation over the whole genome. By playing around with the data zoom ([View > Set Data Zoom Level](#)) you can see where the really large coverage outliers are.



To select these outliers we can use a box whisker filter. Select: [Filtering > Filter by Statistical Test > Box Whisker](#)

In the options which appear, select your two blue Data Groups from the panel on the left, and then opt to select outliers with stringency > 10 Above the median for at least 1 of the 2 selected Data Stores.



Run the filter and save the probe list which is generated. You can then select this probe list by clicking on the + next to “All Probes” in the data view, and selecting the list you created. You can now see which probes were considered to be outliers.

Finally we can create an annotation track of the outlier positions so that we can remember where they were found so that we can later choose to ignore them if interesting hits seem to overlap these positions. To do this simply right click on the filtered probe list you just created and select “Convert to annotation track”. Call the new track “Coverage outliers”

### Exercise 3: Unbiased quantitation

Initially we'd like to do an unbiased quantitation over the genome. To try to get a stable measure of methylation we need to perform our quantitation in windows. The simplest approach would be to use fixed size windows, but these might well show considerable variation in coverage. To get an idea of how this might vary we can make some fixed windows and look at the coverage variability.

Use the same type of probe generation and quantitation as you used in Exercise 2 to make 2.5kb tiled probes over the genome and quantitate these with normalised linear (not log transformed) counts).

[Data > Define Probes > Running Window Generator](#)

Followed by:

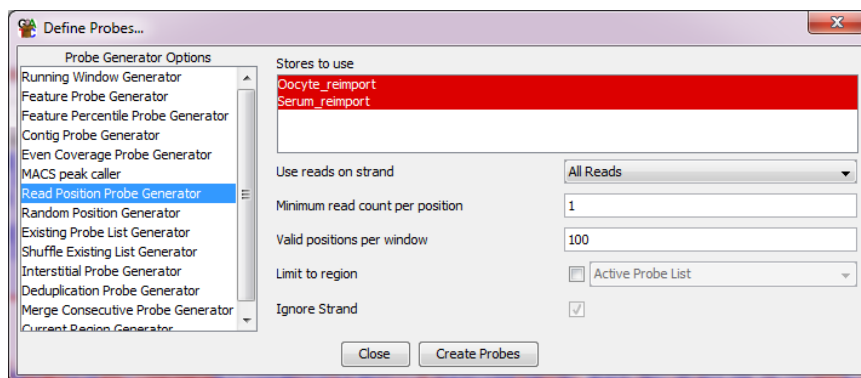
[Data > Quantitate Existing Probes > Read Count Quantitation](#)

Once the data are quantitated look at the variability in the coverage you have across the genome. You will see regions with virtually no calls, and other regions with lots of data. This variability will pass through to the accuracy with which your methylation can be calculated, so we should look for a better way to define our windows so that we can try to equalise the amount of data within each window.

To make more evenly sized probes we can create windows with a fixed number of observed CpGs in them. To do this select:

[Data > Define Probes > Read Position Probe Generator](#)

You need to select both data sets in the box at the top, and then choose to use all CpGs with at least 1 read, but then group the valid positions into windows of 100 CpGs.



If you now quantitate these new windows by read count you should see that the coverage is much more even, and that the windows you get are variable in size. You can look at the distribution of probe sizes by using:

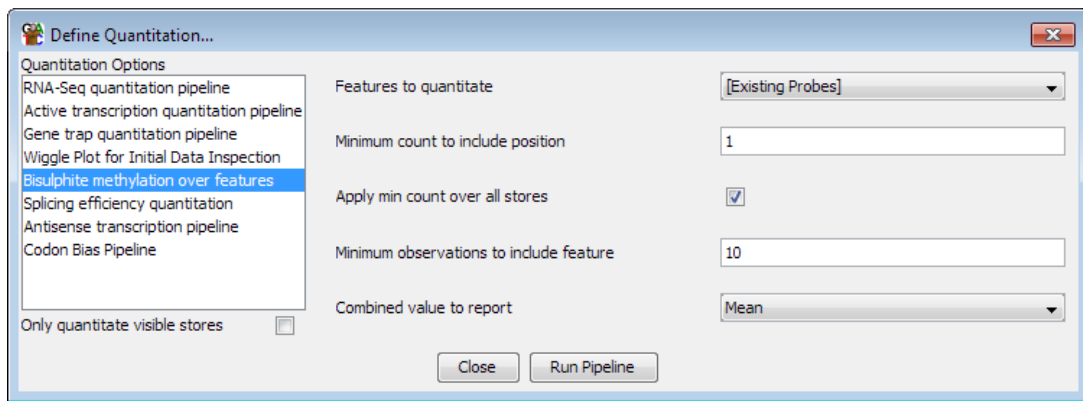
[Plots > Probe length histogram](#)

You should see that you have a few huge outliers (these are where your probes span over large gaps in the assembly), but that the vast majority of probes have a size between around 5kb and 20kb. To see the low end of the histogram you can drag a box to zoom into part of the plot.

You can now quantitate the methylation within the windows you have defined. To do this we're going to use the methylation quantitation pipeline. You can access this from:

[Data > Quantitation Pipelines > Bisulphite Methylation over Features](#)

Rather than specifying features to make new probes we're going to use the existing probes we've already made. Since our overall coverage is low we're going to only require 1 read to include a position, but we'll insist that the positions are the same in both data sets, and we'll require 10 observations per window (which we should generally pass since we know that the combined data contains 100 positions per window).



You should now be able to manually browse around the genome looking at the patterns of methylation.

Spend a few minutes looking over the data to see what you can see. Try to get an impression for:

- Whether the methylation levels for the two samples are similar
- Whether the methylation appears to be biased around any specific annotations
- Whether the methylation is even over the whole genome

#### Exercise 4: Looking at methylation distributions

One simple thing to look at is whether the distributions of values between your datasets is the same.

There are a few ways to look at these in SeqMonk, but the most straightforward is to draw a histogram of the methylation values in each sample. To do this find your datasets in the Data View (top left of the display). You can then right-click on it and in the popup menu that appears select

[Show probe value histogram](#)

You can repeat this for the other dataset so you can see both distributions. Within the distributions window you can drag the 'Divisions' slider to increase or decrease the resolution of the plot.

To get a combined view you could try running a beanplot of the two datasets ([Plots > Bean Plot > Visible Data Stores](#))

#### Exercise 5: Visualising differences

A simple way to look at the differences between two conditions is to plot a scatterplot of the two. You can create a scatterplot using:

[Plots > Scatter Plot](#)

Then select your two groups from the drop down lists at the top.

Take a look at the plot and see what groups of patterns emerge most clearly. Does this match with what you see in the chromosome view of the methylation?

Try putting your mouse over some of the outlier points at the edge of the scatterplot and double clicking. This will take the chromosome view to the equivalent position in the genome so you can see the raw data which underlies those points.

Are you able to see why the methylation difference has been called?

Do you see any cases where you would doubt whether the methylation difference was real?



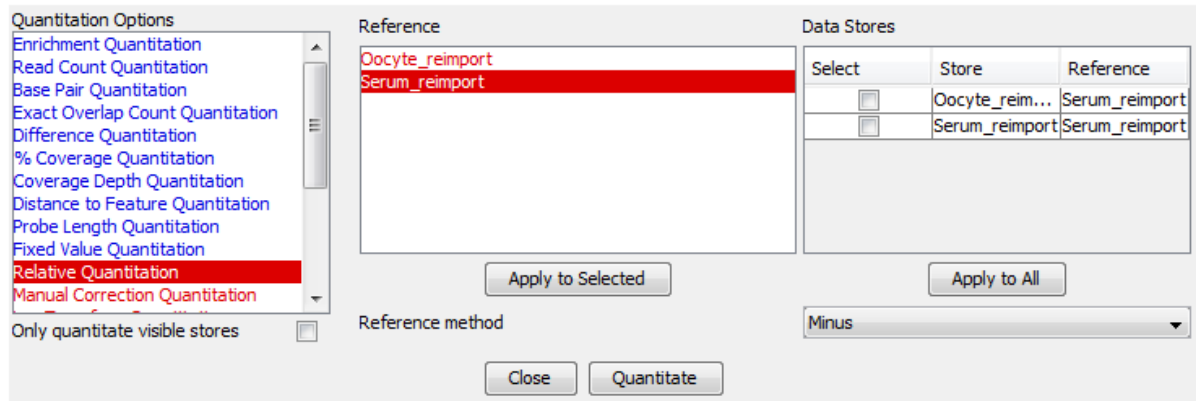
Do you see unusual behaviour at the positions you previously identified as coverage outliers?

Another way to look at the differences would be to express the methylation in one of your samples as the difference to the other sample, in effect performing a relative methylation quantitation.

You can perform a relative quantitation on your data by doing the following:

[Data > Quantitate Existing Probes > Relative Quantitation](#)

Select the Serum Sample as the reference, then press “[Apply to All](#)”. The reference method should be “[Minus](#)”.



If you then quantitate the data you should see the serum values all become zero, and the Oocyte data shows the difference to the serum. This might give a clearer impression of the pattern of differences between your samples. You should also be able to plot out the distribution of differences by selecting the Oocyte dataset in the Data Panel then choosing:

[Plots > Probe Value Histogram](#)

You should see a fairly clear bimodal distribution of differences.

## Exercise 6: Plotting trend plots

We now want to have a look at the pattern of methylation around different types of features. One of the feature types we want to use are CpG islands, but unfortunately in mouse the computationally predicted CpG islands are not very good, so we prefer to use an experimentally determined set.

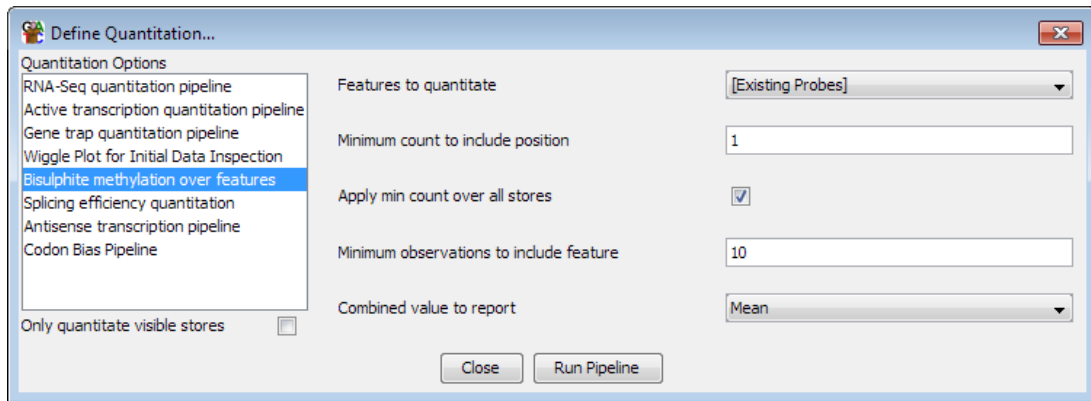
To import a new annotation track with the CpG islands in it select:

[File > Import Annotation > GFF/GTF](#)

In the folder which contained your original data should be a file called `experimental_cpg_islands.gff`. Select this as the file to import. You can choose to add a prefix to your features, but you can leave this blank.

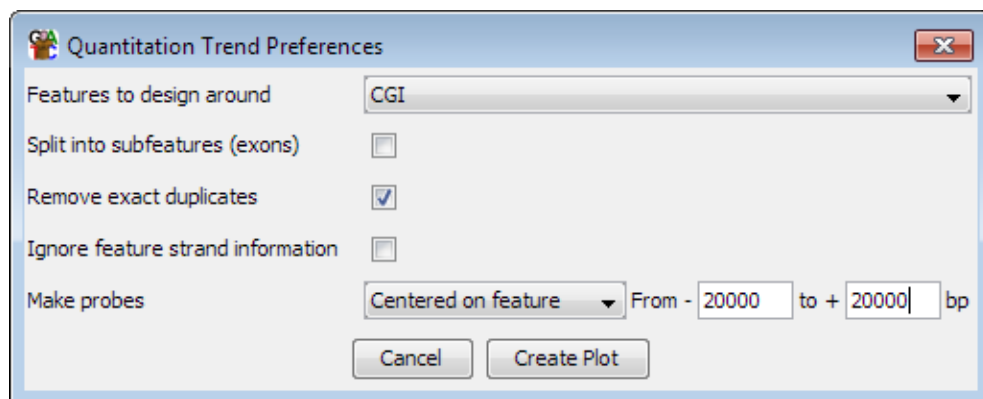
Once the import is complete you should see a new annotation track called “CGI” which contains the new set of features you just imported.

Before doing the trend plots we’ll need to re-run the normal bisulphite quantitation pipeline to get normal methylation values for both of our samples. Please therefore re-run the quantitation the way we did in Exercise 3:



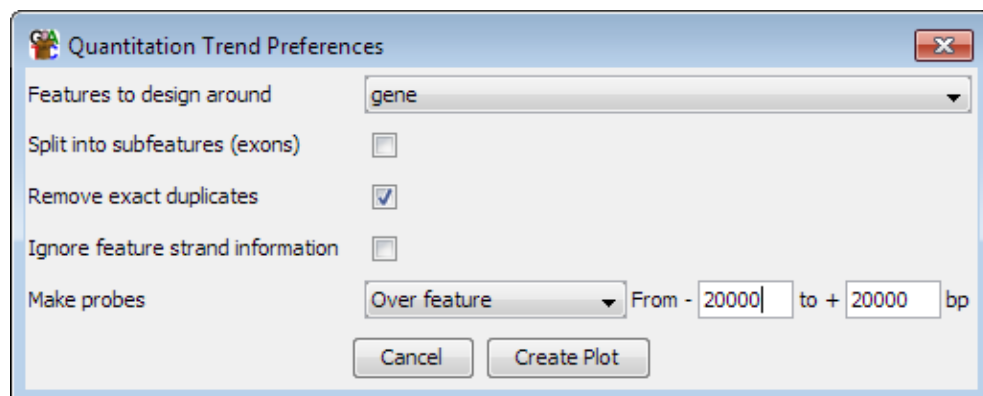
We'll start off by doing a simple fixed width quantitation around CpG islands. To do this select [Plots > Quantitation Trend Plot](#)

We're going to design in windows centred on all CGI features. We'll add 20kb of context to each side to give plenty of score to see the full range of the effect.

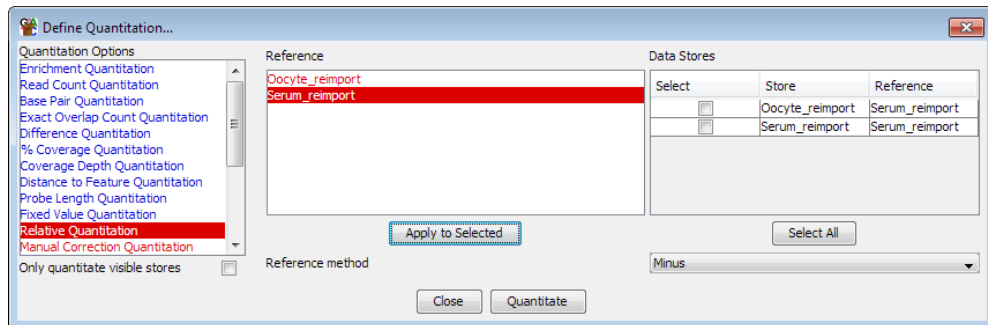


Once you've created the plot have a look at the profiles for the two different datasets and see what conclusions you would draw.

For a more complicated trend plot you could generate a plot over all genes adding 20kb of context to both ends. Have a look at this plot and see what the profiles look like and if you can relate any differences to what you see in the raw data.

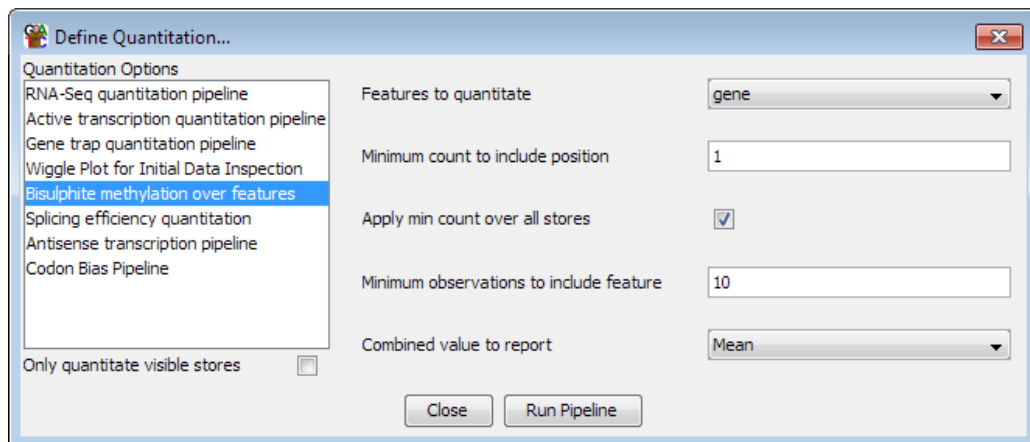


Finally you could try re-applying the relative quantitation we did at the end of Exercise 4 and running the same trend plot over genes for just the Oocyte sample to see how that looks.



### Exercise 7: Targeted quantitation

Hopefully by the end of Exercise 5 you might have seen that it looked like something interesting was happening over gene bodies, so let's go back and target those specifically. We can run the bisulphite quantitation pipeline, but this time we can make it create new probes over all genes.



Having re-done the quantitation you can now try to re-draw the scatterplot of the two conditions against each other to see what pattern you get from looking at a gene-centric view of your data.

In the scatterplot you can double click on any point to see the corresponding region in the chromosome view. Try looking at genes from different parts of the plot and seeing if the original data obviously supports the difference you see, or if there might be any technical artefacts which might be causing problems.