Babraham Bioinformatics

Exercises: Visualising and Exploring BS-Seq

Version 2021-04

Licence

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
- Viewing the raw data
- Finding and annotating 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 `Meth_Course_Data` 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 *Mus musculus* > NCBIM37 listed then select it, if not then press the "Import Genome From Server" button, download it, then open it by pressing the "Start New Project" button.

Once the project is open you can import the data. The Bismark data is in coverage files, a 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.

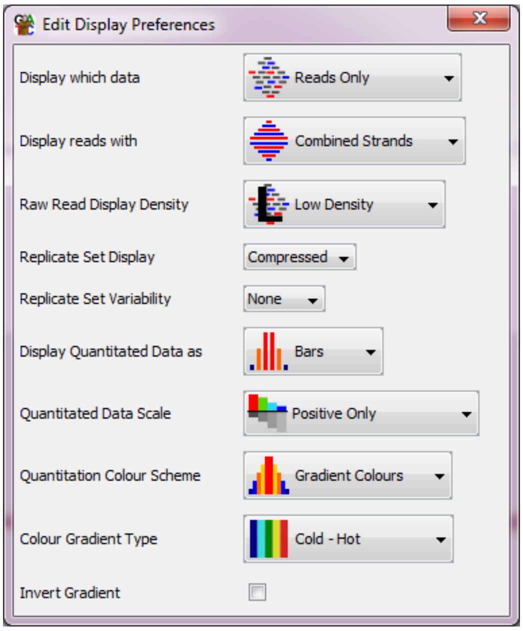
Finally in this section, when the new data imports it will be called by the name of the file from which it came, so if you expand your "Data Sets" section of the Data View (top left window in the GUI) you should see the file names you just imported. These file names are often quite long and unwieldy, so we'll shorten them to something more informative.

If you right click on the CpG_NCBIM37_Oocyte_bismark_pe.deduplicated.cov.gz Data Set and rename it to just "Oocyte" and then do the same thing to the other dataset and name it "Serum". This should make your name display a little clearer for any subsequent work.

Exercise 2 – Viewing the raw data

The data you've been given is whole genome bisulphite sequencing. Think about where you expect the methylation calls to be and the relative depths of coverage for different parts of the genome.

Have a good look at the raw data in detail. Look at the data at various levels of zoom from a few hundred bases to a whole chromosome. Check whether you can see any potential problems with the data.



Edit Display Preferences

Display which data	Reads Only
Display reads with	Combined Strands
Raw Read Display Density	Low Density
Replicate Set Display	Compressed
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	□

To help you look at the data there are some changes you can make to the way the raw data is displayed. You can open the display preferences under View > Data Track Display.

You can change the Raw Read Display Density to get more or less data on screen at the same time.

You can change the Display Reads With option to "Separated strands" to put the methylated (red) and unmethyated (blue) reads into different parts of the track which makes it easier to assess the methylation level.

You can use View > Set Annotation Tracks to add CpG islands as a new track to the display. Have a look at the regions under some islands. See what the presence of a CpG island does to both the data density and the methylation levels. Note that the CpG islands shown here are computationally predicted, and actually miss a large number of weaker islands in mouse. We will import a more complete set of CpG islands later on.

Try to get a general feel for the patterning of methylation levels you see in the two data sets. Are they obviously different? What sort of structure do you think exists in the methylation profiles of the two samples? Whilst we can get some idea of what's going on from looking at the raw calls. This will be much easier when we later use quantitated methylation levels.

Exercise 3: Identifying Coverage Outliers

We are going to make up a new annotation track of dodgy regions in the genome which we don't trust. These are going to be based on places where the coverage of data is suspiciously high.

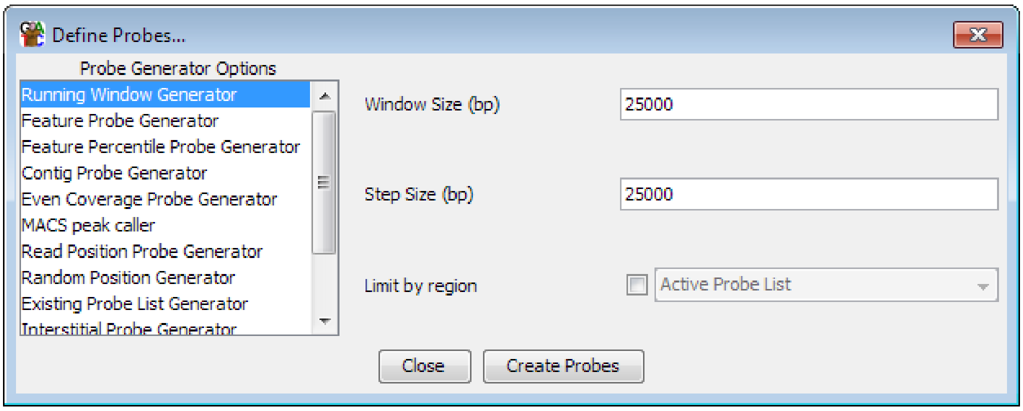
To identify parts of the genome whose coverage is so high that the calls in those regions are immediately suspect we're going to do a coarse quantitation followed by a simple detection of outliers.

To get a quick impression of the coverage we can make probes (measurement regions) of 25kb over the whole genome and then count the number of calls in each window.

We'll start by defining the regions (called Probes) over which we want to make measurements. These will be 25kb tiled probes over the whole genome. To make these select:

Data > Define Probes > Running Window Generator

Set both the window size and the step size to 25000, then press "Create Probes".


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
Existing Probe List Generator
Interstitial Probe Generator
Close
Create Probes

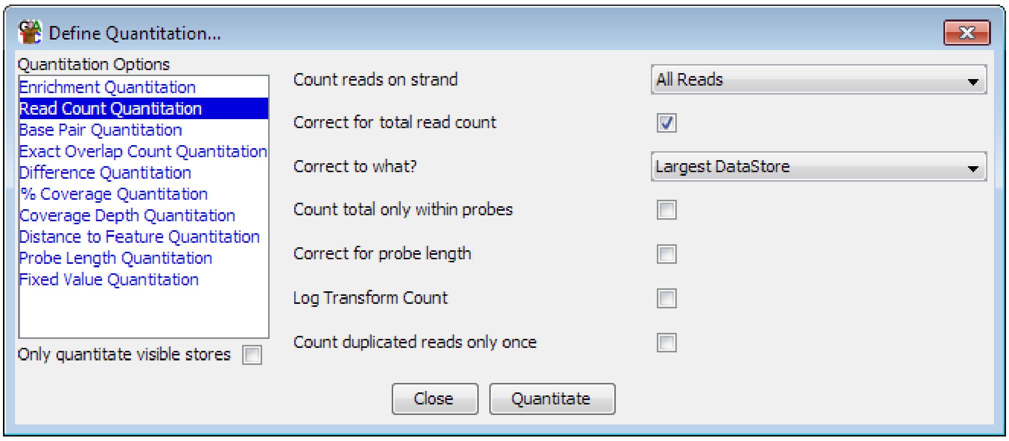
Window Size (bp) 25000

Step Size (bp) 25000

Limit by region □ Active Probe List

The probes should be created, and a new "Define Quantitation" dialog should automatically display (if it doesn't then you can use Data > Quantitate Existing Probes to bring it up).

We now want to quantitate our data. This will associate a numerical value with each Probe in each data set. The quantitation we're going to use to start with is a simple count of the number of methylation calls which fall into each Probe. 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.



Only quantitate visible stores □

Count reads on strand All Reads

Correct for total read count ☑

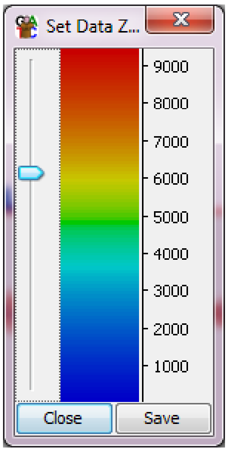
Correct to what? Largest DataStore

Count total only within probes □

Correct for probe length □

Log Transform Count □

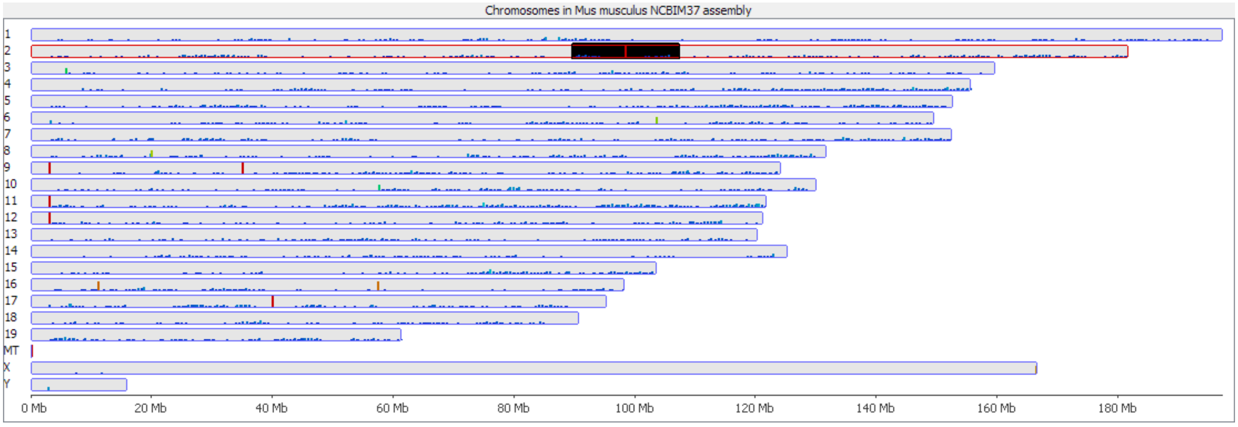
Count duplicated reads only once □



You can now look at your quantitated data in the chromosome view. The height of the bars you see relates to the number of methylation calls falling into that part of the genome.

If you select one of your Data Sets in the Data Panel (top left window in the main display) then you can also see the quantitation over the whole genome. By playing around with the scale on the y-axis of the quantitative part of the chromosome view (View > Set Data Zoom Level) you can see where the really large coverage outliers are.

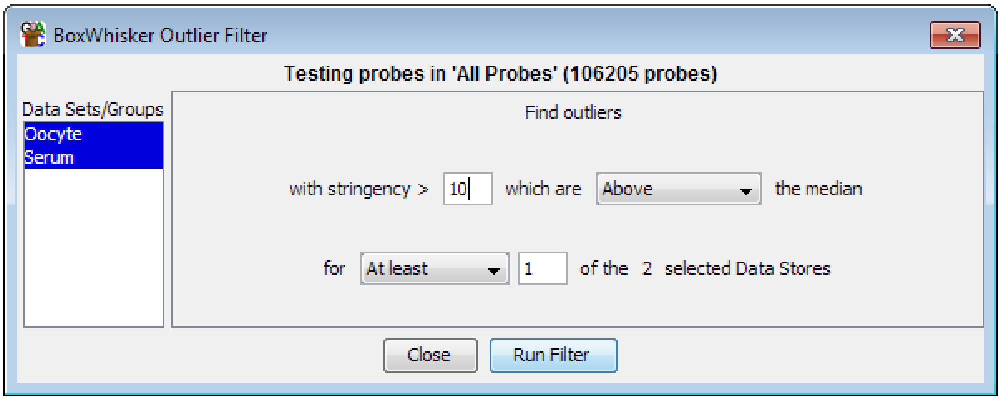
Drag a box over some of the big outliers in the genome view and have a look at the data under them.



We want to identify these coverage outliers, to do this we're going to perform a filter, which will selectively remove some of our original set of Probes. To select the outliers we can use a box whisker filter, which is a way of statistically detecting outliers in any distribution. Select:

Filtering > Filter by Statistical Test > Outlier Statistics > Box Whisker Outlier Detection

In the options which appear, select your two Data Sets 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.


BoxWhisker Outlier Filter

Testing probes in 'All Probes' (106205 probes)

Data Sets/Groups: Oocyte, Serum

Find outliers

with stringency > 10 which are Above the median

for At least 1 of the 2 selected Data Stores

The value of 10 here is somewhat arbitrary, but the outliers are so far away from the main distribution it doesn't matter too much what value is selected. Traditionally BoxWhisker outliers use a stringency of around 2, so 10 is already only picking huge outliers.

Run the filter and save the probe list which is generated. You can then select this probe list by clicking on the little icon to the left of the "All Probes" entry in the Data View tree, 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"

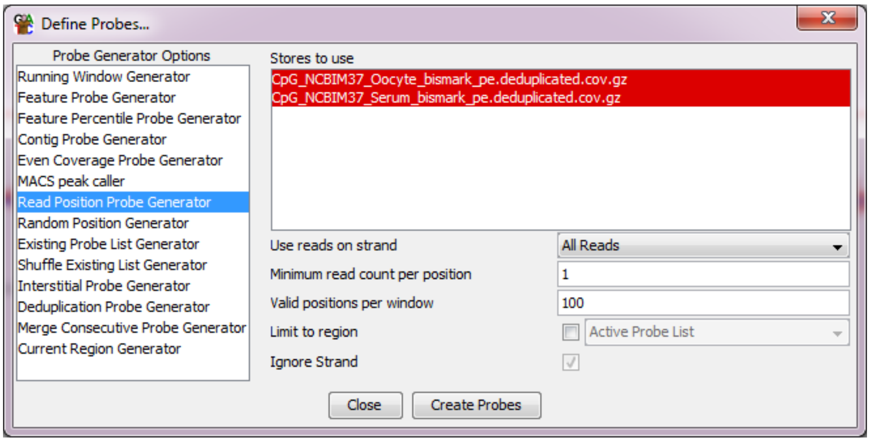
Find some of these outlier regions in the view and have a look at the data underneath them to see if you can see why we might want to ignore these.

Exercise 4: Unbiased Methylation Quantitation

Now we can start to look quantitatively at the methylation levels in our datasets. Initially we'd like to do an unbiased quantitation over the genome. We don't have enough data to quantitate every cytosine so we're going to quantitate in windows, and to equalise our noise and power we're going to make windows with fixed numbers of CpGs in them.

To make probes with a fixed number of observed CpGs in them 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.


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
Existing Probe List Generator
Shuffle Existing List Generator
Interstitial Probe Generator
Deduplication Probe Generator
Merge Consecutive Probe Generator
Current Region Generator
Close
Create Probes

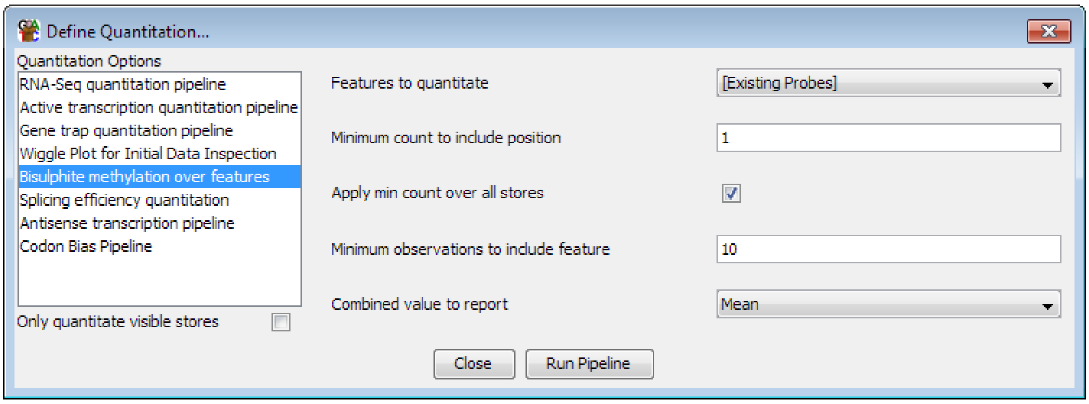
Stores to use
CpG_NCBIM37_Oocyte_bismark_pe.deduplicated.cov.gz
CpG_NCBIM37_Serum_bismark_pe.deduplicated.cov.gz

Use reads on strand: All Reads
Minimum read count per position: 1
Valid positions per window: 100
Limit to region: □ Active Probe List
Ignore Strand: ☑

After creating the probes you will again get up the standard quantitation options. We're going to do a slightly more complicated methylation quantitation, so you can close the default quantitation window.

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 CpG position, but we'll insist that the same positions are used in both data sets, and we'll require 10 observations per probe (which we should generally pass since we know that the combined data contains 100 positions per probe).


Define Quantitation...
Quantitation Options
RNA-Seq quantitation pipeline
Active transcription quantitation pipeline
Gene trap quantitation pipeline
Wiggle Plot for Initial Data Inspection
Bisulphite methylation over features
Splicing efficiency quantitation
Antisense transcription pipeline
Codon Bias Pipeline
Close
Run Pipeline

Only quantitate visible stores □

Features to quantitate: [Existing Probes]
Minimum count to include position: 1
Apply min count over all stores: ☑
Minimum observations to include feature: 10
Combined value to report: Mean

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:

- How big your probes (measurement regions) are. This will be reported in the status bar at the bottom of the program when you mouse over a probe in the chromosome view
- 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 5: Looking at methylation distributions

One simple thing to look at is whether the distributions of values between your datasets are 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 each of them and in the popup menu that appears select
Show probe value histogram

You can do this for both datasets 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 6: 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.

The colours in the scatterplot reflect the density of points which fall into that part of the plot (since you don't have enough pixels to show every probe you have). Hotter colours have more probes than colder colours.

Take a look at the plot and see what combinations of methylation levels in oocyte and serum emerge most clearly. Does this match with what you can see in the chromosome view of the methylation? What sort of biological questions do you think would make sense given the global difference in methylation between the two samples which you can see.

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.

Exercise 7: Plotting trend plots

Now we 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.

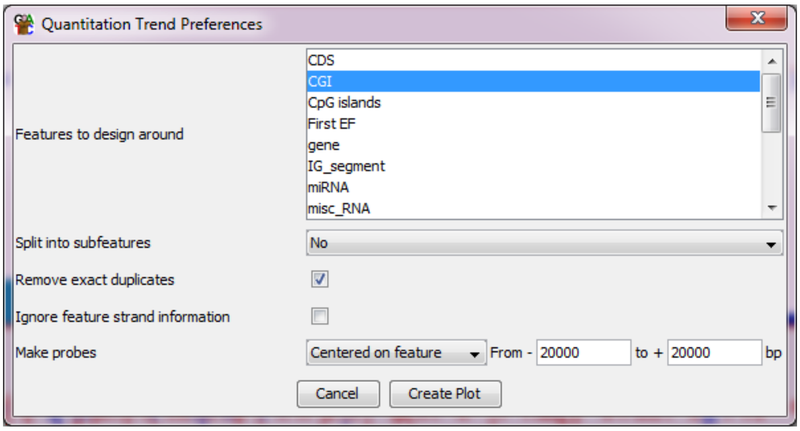
To import a new annotation track with the CpG islands in it select:
File > Import Annotation > GFF/GTF

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. Compare this to the “CpG islands” track of computationally predicted islands which we used before, and see that there are more islands in the experimentally determined set.

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 the middle of 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.



Quantitation Trend Preferences

Features to design around: CDS, **CGI**, CpG islands, First EF, gene, IG_segment, miRNA, misc_RNA

Split into subfeatures: No

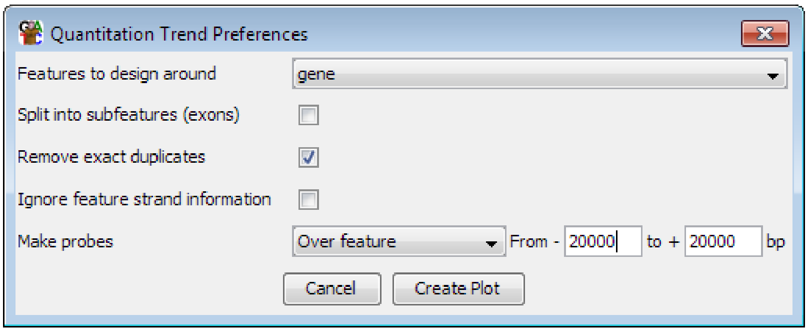
Remove exact duplicates ☑

Ignore feature strand information ☐

Make probes: Centered on feature From - 20000 to + 20000 bp

Once you’ve created the plot have a look at the profiles for the two different datasets and see what conclusions you would draw. Does it look like the CpG islands have an effect on the methylation level in these samples, and does it look any different between serum and oocyte?

For a more complicated trend plot we will next generate a plot over all gene bodies, 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.



Quantitation Trend Preferences

Features to design around: gene

Split into subfeatures (exons): □

Remove exact duplicates: ☑

Ignore feature strand information: □

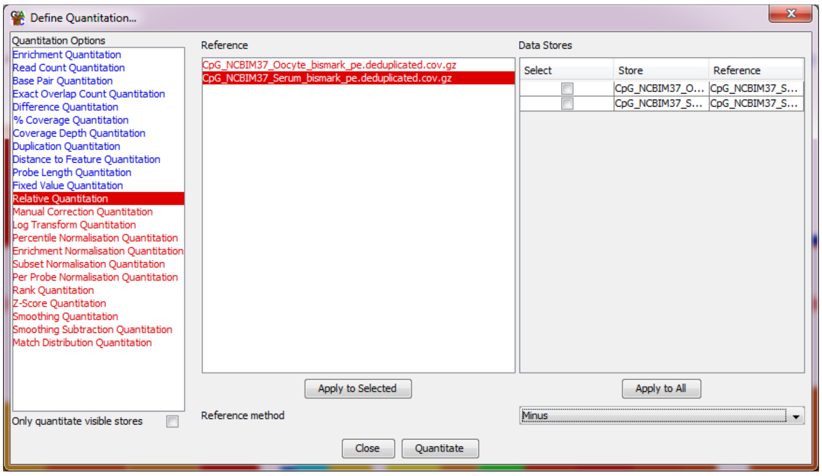
Make probes: Over feature From - 20000 to + 20000 bp

Specifically look at the way the trend plot changes when you enter and leave the gene body, and see if that looks any different between the two samples.

Finally, to try to make the positional enrichment between the samples clearer we will plot our trend plot a different way. Instead of plotting the absolute amount of methylation in the oocyte and serum separately, we will plot out the *difference* in methylation between the two samples.

To do this we need to change our quantitation to be the difference between the two samples. We can do this 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".



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
Relative Quantitation
Manual Correction Quantitation
Log Transform Quantitation
Percentile Normalisation Quantitation
Enrichment Normalisation Quantitation
Subset Normalisation Quantitation
Per Probe Normalisation Quantitation
Rank Quantitation
Z-Score Quantitation
Smoothing Quantitation
Smoothing Subtraction Quantitation
Match Distribution Quantitation

Reference

CpG_NCBIM37_Oocyte_bismark_pe.deduplicated.cov.gz
CpG_NCBIM37_Serum_bismark_pe.deduplicated.cov.gz

Data Stores

Select	Store	Reference
□	CpG_NCBIM37_O...	CpG_NCBIM37_S...
□	CpG_NCBIM37_S...	CpG_NCBIM37_S...

Only quantitate visible stores □

Reference method: Minus

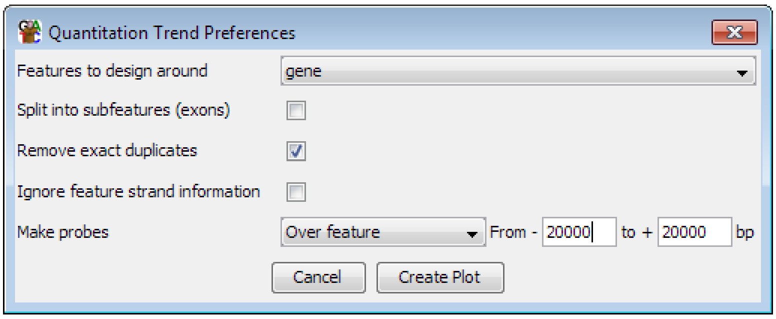
Apply to Selected
Apply to All
Close
Quantitate

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.

If you have a look at the values in the chromosome view you should also see the blocks of increasing and decreasing methylation more clearly than before.

You can now re-draw the quantitation trend plot over genes and you should now see a clearer rationale for using the gene body as the basis for a feature based analysis of this data.

Plots > Quantitation Trend Plot



Quantitation Trend Preferences

Features to design around: gene

Split into subfeatures (exons) □

Remove exact duplicates ☑

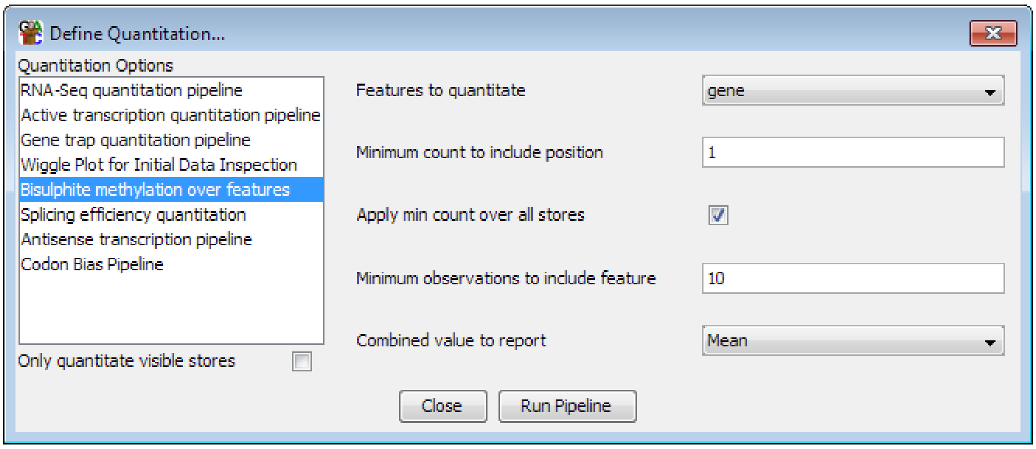
Ignore feature strand information □

Make probes: Over feature From - 20000 to + 20000 bp

Exercise 8: Targeted quantitation

Hopefully by the end of Exercise 7 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.

Data > Quantitation Pipelines > Bisulphite Methylation over Features



Define Quantitation...

Quantitation Options

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

Only quantitate visible stores □

Features to quantitate: gene

Minimum count to include position: 1

Apply min count over all stores ☑

Minimum observations to include feature: 10

Combined value to report: Mean

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. Try exploring the main clusters in the plot, but also look at some examples of genes which do not fit the general trend since these can often also be interesting.

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.