BioTrain.TV
#Babraham Bioinformatics

Exercise

Visualising and Exploring Methylation Data in SeqMonk

Version 2026-06

Licence

Introduction

In this exercise we will explore some DNA methylation data using SeqMonk. We want to use the quantitation and visualisation tools to understand what is happening in the samples we're looking at. The practical starts from the coverage files produced by the Bismark methylation extractor.

The basic steps we're going to include are:

- Importing Bismark methylation calls into SeqMonk
- Viewing the raw data
- Unbiased methylation visualisation and quantitation
- Comparing methylation between samples
- 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 – Create a Project, Import Data

The data we're using for this section is in the `Visualising_and_Exploring` subfolder of your main `Meth_Course_Data` data folder. There will be two files 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

You should see ***Mus musculus* > NCBIM37** listed and be able to **select it**.

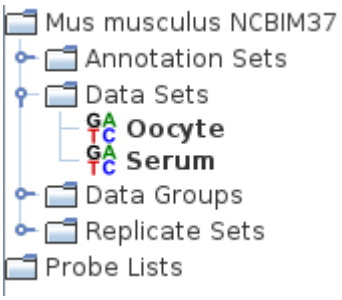
Create the project by pressing the "Start New Project" button.

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 `Meth_Course_data/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.

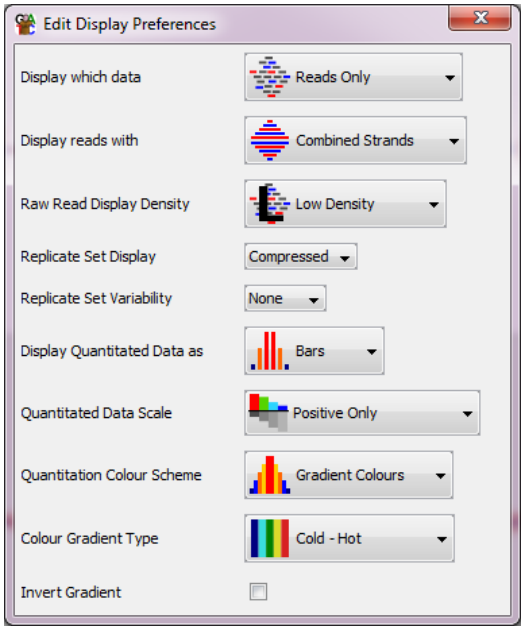


If you **right click** on the **CpG_NCBIM37_Oocyte_bismark_pe.deduplicated.cov.gz** Data Set and **rename** it to just "Oocyte" and then do the same thing to the other dataset and name it "**Serum**". This should make your name display a little clearer for any subsequent work.

Exercise 2 – Viewing the raw data

The data you've been given is whole genome bisulphite sequencing.

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.



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.

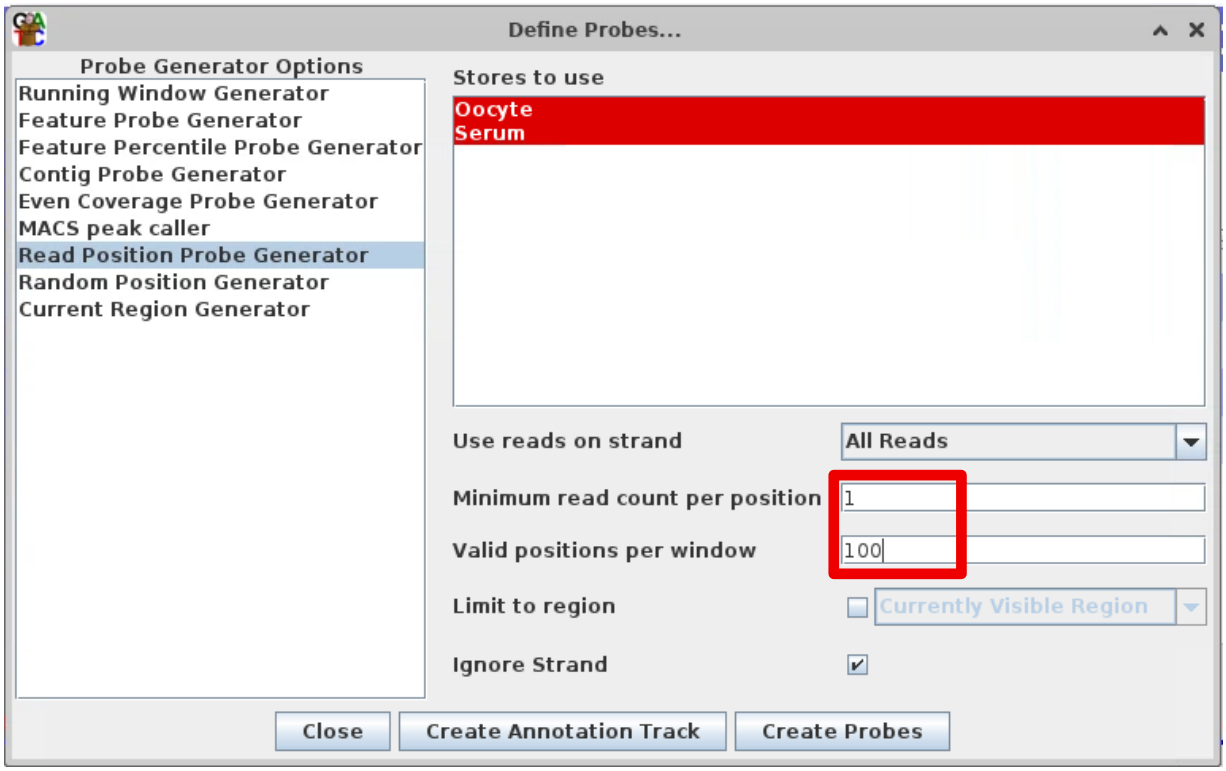
Try to get a general feel for the patterning of methylation levels you see in the two data sets. Are they obviously different? Is one more methylated that the other overall? If so does this look like it happens everywhere in the genome?

Exercise 3: Unbiased Methylation Quantitation

Next 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

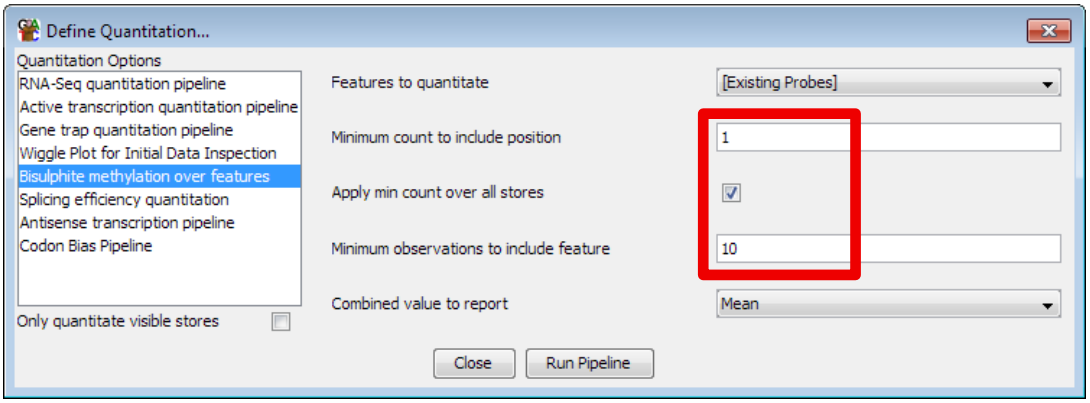
You need to select both data sets in the box at the top, and then choose to use all CpGs with at least 1 read, but then group the valid positions into windows of 100 CpGs. Press "**Create Probes**" to make the probes.



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).



Press “Run Pipeline” to perform the quantitation.

You should now be able to see bars in your main view where the height of the bars represents the level of methylation. Browse around the genome and have a look at the patterns of methylation.

A few questions to answer from this.

- Are the methylation levels for the two samples are similar
- Whether the methylation appears to be biased around any specific types of feature
- 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 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 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.

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 mean there is more data in that part of the plot 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.

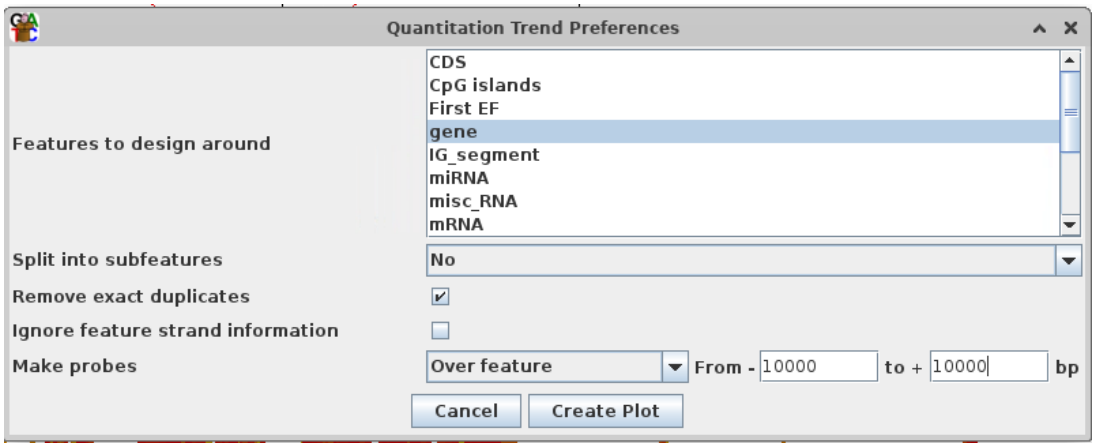
Exercise 6: Plotting trend plots

Now we want to have a look at the pattern of methylation around genes. We can use a quantitation trend plot to do this.

To do this select

Plots > Quantitation Trend Plot > Current Probe List

We'll make a plot around all genes with an additional 10kb upstream and downstream of the gene features.



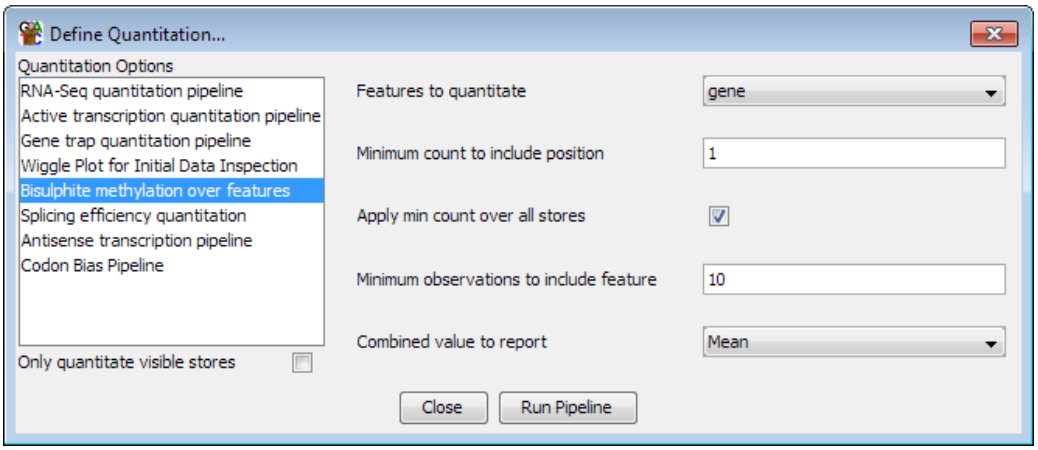
Once you've created the plot have a look at the profiles for the two different datasets and see what conclusions you would draw. Pay particular attention to the methylation level before entering the gene and the methylation level within the gene. Are there differences between Oocyte and Serum?

Take a look at the main display and see if you can see that something specific happens within genes in the Oocyte sample.

Exercise 7: Targeted quantitation

Hopefully by the end of Exercise 6 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



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.