Babraham Bioinformatics

Exercises:
Differential Methylation

Version 2017-11

Licence

Introduction

In this session we will look at a couple of different ways to try to identify differentially methylated regions, and will look at how we can visualise and validate the predictions which are made.

We're going to be running the statistical analysis in SeqMonk, but using methods which are present in most of the available methylation analysis packages, so the same tests could be performed in a non-interactive way by using those packages.

Software

The software packages used in this practical are:

- SeqMonk (http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/)
- R (core functions only) (https://www.r-project.org/)

Data

The data in this practical are reduced coverage resampled subsets 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 – Loading Data

To save time the data for this practical have already been imported and grouped in a SeqMonk project, so you can just load this.

The project file is called `differential_methylation_data.smk` and is located in the Differential_Methylation sub-folder of the course data. You can load this by selecting:
File > Open Project

Some work has already been performed on this data, namely:

- The Bismark methylation extractor data was imported using the generic text import
- The replicates for the two conditions were combined into replicate sets
- The display was changed to show only the replicate sets
- The project was quantitated using the bisulphite pipeline over sets of 50 CpGs

Exercise 2 – Data Inspection

Draw a scatterplot of the methylation levels in the Oocyte and ES cells (Plots > Scatterplot). What is the overall structure of this data. We're going to go ahead and do a simple differential methylation analysis, but can you think of a better way to approach the data given the overall structure you can see?

Exercise 3 - Viewing the methylation consistency

Since the view is showing just the replicate sets what you are seeing is the average methylation value for the different replicates. There are a couple of things you can do to look at the variability.

Firstly you can put error bars onto the measures which are there. To do this select:
View > Data Track Display > Replicate Set Variability
...and then try the different options to see how their view of variability changes.

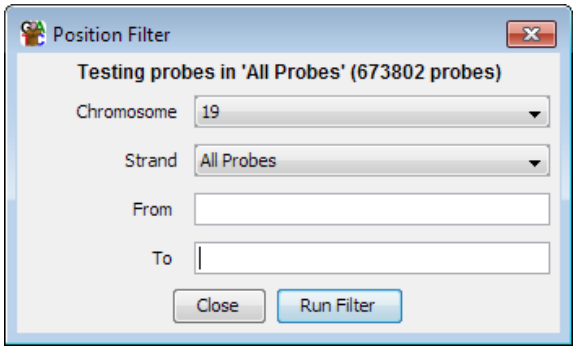
Alternatively you can split the replicate sets into their individual data sets. To do this select:
View > Data Track Display > Replicate Set Display > Expanded

Exercise 4: Unreplicated differential methylation using a Chi-Square test

We will analyse these samples using the merged data in the replicate sets. This won't take any account of how consistent the methylation differences are, but groups all of the data together to allow us to perform a simple test.

To speed things up we're only going to analyse a single chromosome, so start by selecting
Filtering > Filter by position

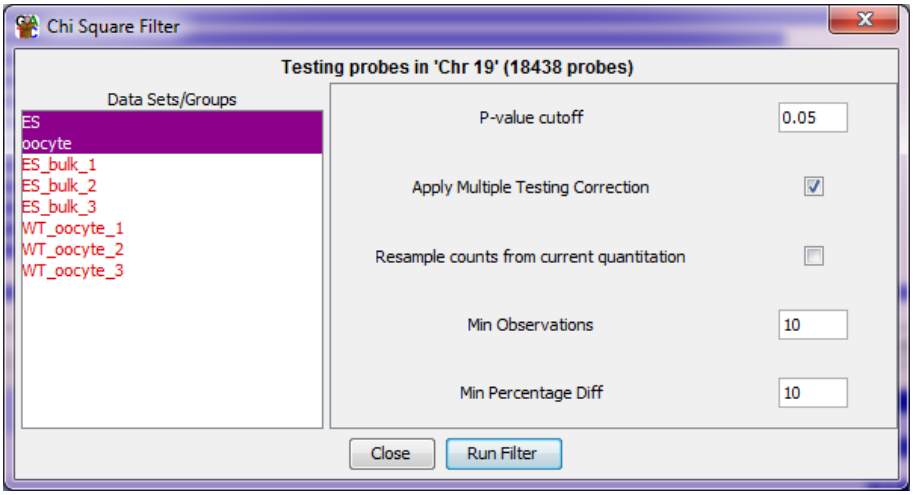
And then select chr19 and put no positions, so you select the whole chromosome.
.



From the data view at the top left, select the chr19 list.

Now we can run the statistical filter. Select:
Filtering > Filter by statistical test > Chi-Square > For/Rev

In the filter options make sure the two purple replicate sets are selected and press "Run filter". Save the probe list which is produced.



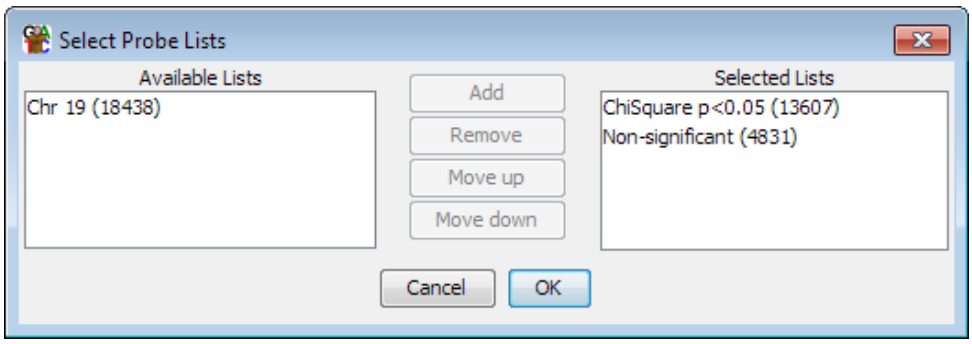
As a comparison we're then going to make a list of the probes which didn't change. To generate this select:
Filtering > Logically Combine existing lists

Use the options below to isolate the non-significant Chr19 probes.
Chr 19 BUTNOT ChiSquare p<0.05

Call this list "Non-significant"

Now you can view the effects of the filter. To do this select the chr19 list in the data view, then select Plots > Scatter plot.

From the drop down boxes at the top select the two replicate sets, then press the "Highlight sublists" button at the top (make the plot a bit bigger if you can't see this). In the highlight options add in the significant and non-significant lists and update the plot. Have a look to see if you can see an obvious difference between the significant and non-significant probe sets.



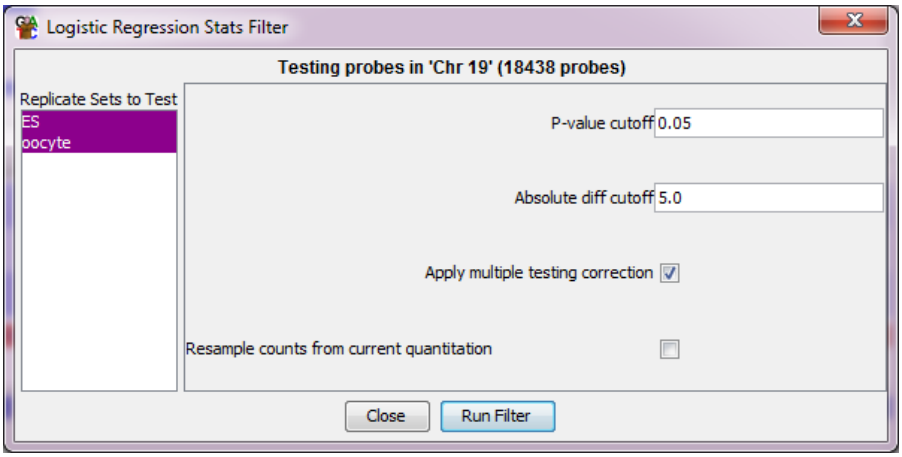
Exercise 5: Replicated differential methylation using Logistic Regression

The Chi-Square statistics don't take any account of the variability between biological replicates so a better option if you have replicates is to use a logistic regression test. This also uses absolute counts between your samples but also allows for replication. Logistic regression is present in many of the BS-Seq R packages, but we're going to use it using the R bridge within SeqMonk.

We're going to start from the same Chr19 probes we used in the previous exercise, but this time we're going to select:

Filtering > Filter by statistical test > R-Filters > Logistic Regression For/Rev

As before we are going to select the two replicate sets (ES and oocyte), and require a p<0.05 difference after multiple testing correction.



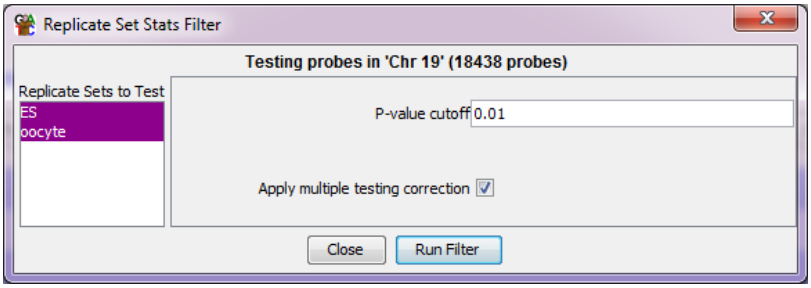
You should see the R script running to generate the list of hits then these will automatically be re-imported to your project.

Use the same procedure as for Exercise 3 to view the hits in context on a scatterplot.

Exercise 6: Normalised differential methylation using a t-test

The quantitation we are using is not a simple percent methylation calculation (ie meth calls / total calls *100) but is a more complex quantitation which aims to remove biases due to coverage differences. If we want to use this quantitation directly, or if we wanted to apply any correction or normalisation to the methylated values then we can no longer use the count based tests we used above.

As a final test we can run a t-test one the percentage methylation values to look for significance. To do this select your chr19 probes and then select Filtering > Filter by Statistical Test > Replicate Set Stats.



If we select our two replicate sets here we will do a t-test between the two sets of methylation values. Run the test and visualise the results on a scatterplot as before. How do the results compare to the logistic regression? Are there more or fewer hits? Why might this be? Are the results cleaner than before?

To generate a more stringent list of hits we could logically combine the results of the logistic regression and the t-test. Try this and see if the results are cleaner.

Exercise 6 [Optional]: Division based on structure

At the start of the exercise you were asked to look at the overall structure of the data where you should have seen that there were 3 obviously separate groups of probes. Use the values filter (Filtering > Filter on Values > Individual Probes) along with the logically combine filter to separate these 3 separate groups of probes. For each of the separated groups use the features filter (Filter > Filter on Features) to find out what proportion of the probes in each group overlap with a gene.

Exercise 7 [Optional]: Independent analysis of a more challenging dataset

Now that you have had experience of visualising and analysing a relatively straight forward dataset we will let you have a go at a more difficult one. You should look at the file “difficult_bs_seq.smk”.

This data is whole genome BS-Seq from pig embryos treated in 3 different ways (Control, Fluids, In Vivo), with 3 replicates for each condition. The coverage is not hugely high and you will see that the samples are taken during a period of reprogramming, so there are some differences in the overall methylation levels between the samples.

Use the tools and techniques you’ve been shown to try to understand and analyse this data. Steps to perform will be:

1. Perform an unbiased quantitation of the data using fixed CpG windows (100 CpGs per window should be about right for this data). See if you have sufficient coverage to insist on using matched CpGs in all samples for quantitation, and revert to a non-matched quantitation if not.

2. Look at the global distributions of methylation values in the samples and see what you can learn from them. If they differ (which they do!) then see if this difference is related to the experimental conditions. See if you can normalise the global differences away to make a cleaner comparison of individual regions. To normalise the methylation values select Data > Quantitate Existing Probes > Match Distribution Quantitation.

3. Look at the comparison of the Control vs Fluids samples to start with. Look at the overall relationship between them and then use the statistical tools (logistic regression) to find significant differences. If you have normalised the values then you’ll need to use the option to “Resample Counts from Current Quantitation”. Visualise and review the hits and see if you think any of them would be worth following up.