Babraham Bioinformatics

# Exercises:
# Differential Methylation

*Version 2021-04*

# 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 (https://www.r-project.org/)
- EdgeR (https://bioconductor.org/packages/release/bioc/html/edgeR.html)

## *Data*

The data in this practical are the Bisulphite Sequence data from GEO accession GSE30199. This study looked at allele specific methylation to identify imprinted genes in the mouse genome.

The data used here was processed using Bismark initially and was then run through SNPSplit (https://www.bioinformatics.babraham.ac.uk/projects/SNPsplit/), to separate out the reads coming from the two different parental genotypes. The final data still looks the same as for a normal bismark bisulphite run, just separated into two groups.

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 `allele_specific_methylation.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 coverage files were imported into SeqMonk
- The names of the data sets were shortened
- Replicate sets were created from the replicates from the two different parental strains (called allele1 and allele2)

Spend a few minutes looking at the raw sets of calls. Think about the following questions:

1. What is the overall coverage like? Is most of the genome covered, or is it patchy? This is supposed to be an unbiased library, do the reads look like they match this expectation?

2. How deeply sequenced are the samples? We need to think about the resolution we have so look at individual bases and see what sort of fold coverage of the genome we are going to have.

3. Have a look at the coverage on the X and Y chromosomes. Does this look the same as the rest of the genome? If not, why not?

4. Do all of the individual replicates have comparable coverage and methylation (to the degree you can see from the raw calls)? Can you see obvious differences between them?

## Exercise 2 – Quantitation

We next want to have a more quantitative look at the data. We're going to take an unbiased view of the genome to start with and see what seems to be happening overall in these samples.

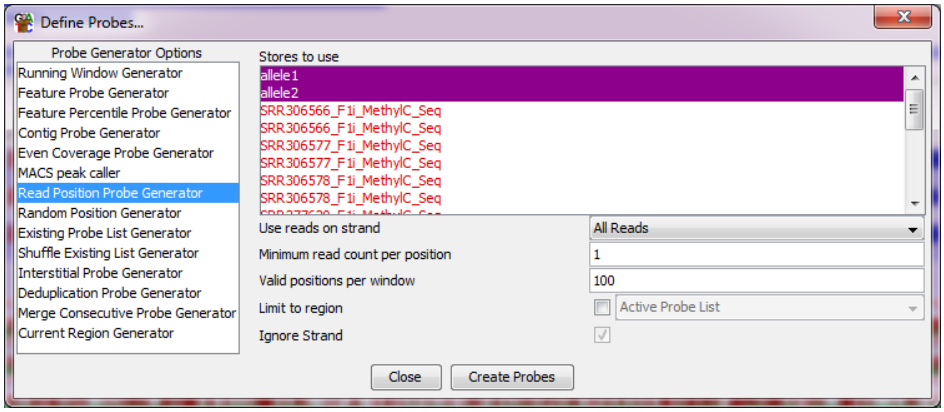
From the initial assessment of the coverage we should see that we really only have about a 1x coverage of the genome. The changes we're looking for are likely to be quite high (potentially 50% or more), so looking back at our power analysis table we can see that we would be justified in using an unbiased window size of 100 CpGs.

Window Size (# CpG cytosines)

| Absolute methylation change (from 80%) | 1 | 10 | 25 | 50 | 100 | 200 | 500 |
|---|---|---|---|---|---|---|---|
| 1 | 158805 | 14212 | 5419 | 2609 | 1254 | 602 | 228 |
| 5 | 6794 | 608 | 232 | 112 | 54 | 26 | 10 |
| 10 | 1825 | 164 | 63 | 30 | 15 | 7 | 3 |
| 20 | 509 | 46 | 18 | 9 | 5 | 2 | 1 |
| 50 | 94 | 9 | 4 | 2 | 1 | 1 | 1 |

Required Fold Genome Coverage

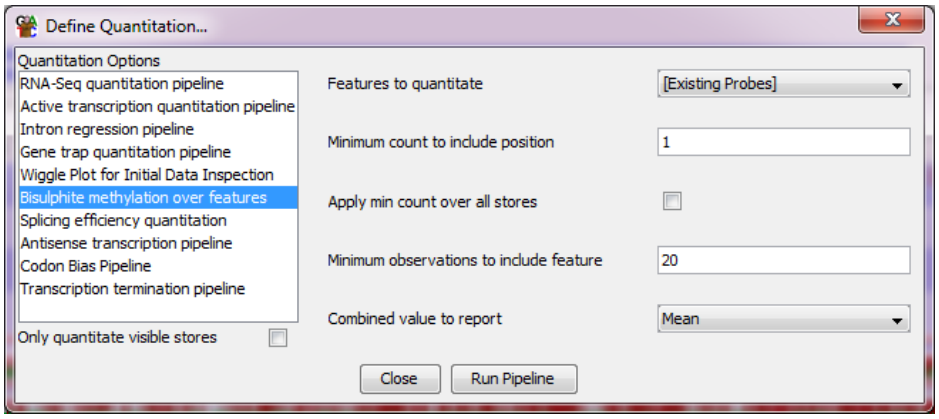
Create 100 CpG probes in the data. To do this go to Data > Define Probes > Read Position Probe Generator. You need to select the two replicate sets to define the data then increase the "Valid positions per window" up to 100.




Once you've created the probes you can dismiss the default quantitation options (press cancel) since we're going to use the bisulphite pipeline to quantitate the data.

Select Data > Quantitation Pipelines then select the "Bisulphite Methylation over Features" pipeline.

For the pipeline we'll set a minimum number of observations to 20. We know that there are 100 CpGs in each window, so only 1 in 5 need to be observed in each sample for it to be quantitated.




After the quantitation has finished then run Reports > DataStore Summary Report and look at the last column (Valid Quantitations) to see how many probes in each sample you were able to successfully quantitate. Compare this number to the total number of probes in your project. You want to see that the vast majority of probes in each sample were able to be quantitated.

## Exercise 3 – Data Inspection

Now that we have some quantitative values for the data we can start to do a more systematic profiling of the distributions of methylation values we see, and the differences between samples. The sorts of questions we are trying to address are:

1. What is the overall distribution of methylation values and levels in the samples. How methylated are they overall and do the methylation values form a single distribution, or are there obviously separate high and low methylated regions.

2. Do the distributions of methylation values between the different samples look very similar overall? If not then do we need to consider some additional normalisation, or are the differences likely to reflect interesting biology meaning we should characterise them directly.

3. Do we see clear methylation differences between samples or are there not obvious candidates for differential methylation? If there are differences then do they cluster by the replicates for a given allele, or are they more randomly patterned over the samples?

To try to answer these questions use the visualisation tools in seqmonk to explore the quantitated values you have created. Some suggestions for plots which will be informative;

1. To look at the distributions you can use:
    a. Plots > BeanPlot
    b. Plots > Cumulative Distribution Plot

2. To look at the comparison of methylation values you can use
    a. Plots > Scatterplot – double click on points to see the corresponding region in the main chromosome view
    b. The various tools under Plots > Data Store Similarity to look at the overall similarity of all of the samples together.

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

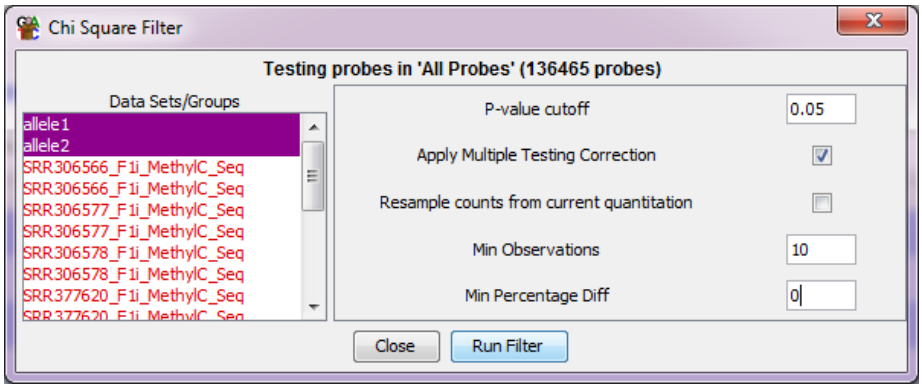
We are going to get you to run a few different statistical tests to look for differential methylation. The first is a simple Chi-Square test. This isn't an ideal test to run on this data because it takes no account of the variance we see between the replicates which exist in this experimental design, but it could be useful if we had very low coverage and needed to merge raw data to get sufficient observations to achieve a significant result.

To run the test use:

Filtering > Filter by Statistical Test > Proportion Based Statistics > Unreplicated Data > Chi-Square (for/rev)

We'll remove the default constraint of requiring a 10% absolute difference, so we just filter on the basis of significance.

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. We need to select our two replicate sets, which will cause the test to analyse the merged raw data contained in each of these.




Run the test and save the results. The filtered list will become a sub-list of "All Probes" in your Data View. You will need to expand the All Probes part of the tree to see the new list which was created.

How many hits did you get?

After running each filter we want to do some sanity checking to make sure the results we see are sensible. The easiest way to do this is to draw a scatterplot of the two alleles against each other using All Probes (Plots > Scatterplot) and then use the "Highlight Sublists" option to highlight on the plot the hits from the statistical test. You should be able to see that the points which were selected fell on the outside of the main cloud of points.

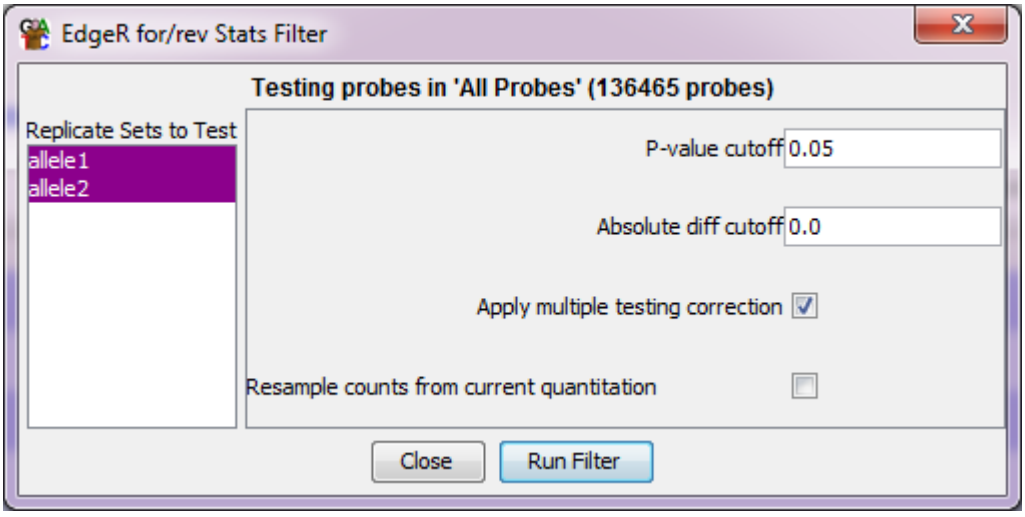
## Exercise 5: Replicated differential methylation using EdgeR

EdgeR has an alternative method of analysing replicated methylation data based on the binomial distribution of linked pairs of counts. It is applied in the same way as logistic regression and should achieve roughly similar results. We will run this version of the test so we can compare the hits we get.

We're again going to start from the All Probes list, but this time we're going to select:

Filtering > Filter by Statistical Test > Proportion Based Statistics > Replicated Data > [R] EdgeR (for/rev)

As before we are going to select the two replicate sets (allele1 and allele2), and require a p<0.05 difference after multiple testing correction.



You will again see the R script running. Again it will take a few minutes to complete so please be patient. After you have saved the hits then review them in a scatterplot as you did for the chi-square test.

## Exercise 6: Collating the hit lists

You should now have sets of hits from two different statistical methods. We now want to see how similar these lists are. We can look at the overall statistics for the degree of overlap between the lists using:

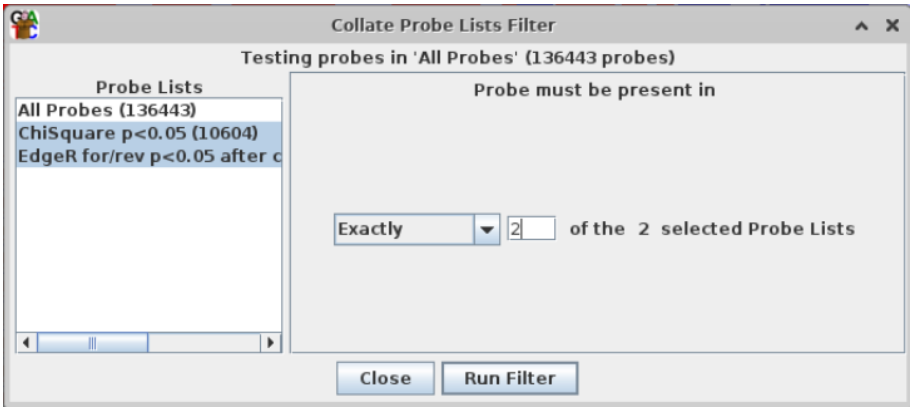
Plots > Probe List Overlap > Probe List Overlap Matrix

Are the lists largely similar?

For the final set of hits we're going to take the subset of hits which were detected by both of the methods we used. To do this we're going to use the Collate Lists Filter.

Filtering > Combine Existing Lists > Collate Multiple Lists

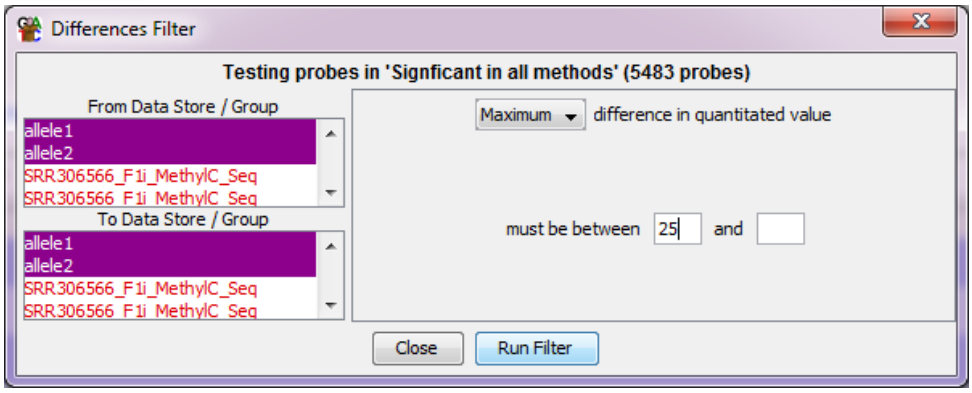
We're going to keep probes which were present in both of the hit lists we generated.




Finally, we are going to add a constraint for a certain degree of absolute change between the two conditions. Imprinted genes should show fairly large changes, so we're going to require at least a 25% absolute difference in methylation to make it through to our final hit list.

We are therefore going to start by selecting the collated hit list we've just created and then applying a differences filter.

Filtering > Filter on Value Differences > Individual Probes




By selecting both replicate sets in the From and To groups we do the subtraction both ways round, so filtering on the maximum differences will find changes which went up in either allele.
How much did this filter reduce our hit list by?

## Exercise 7: Reporting and linking to imprinted genes

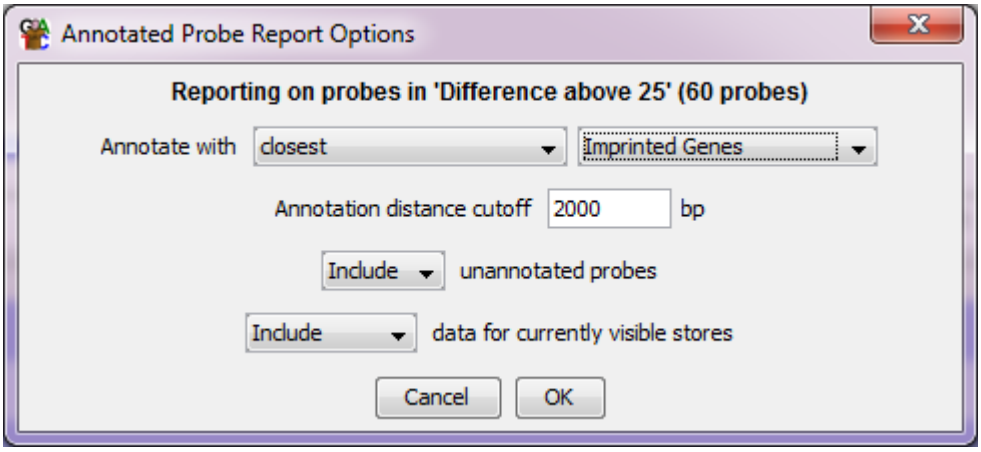
The final part of the exercise is to create a report on the hits we made, and to link these to known imprinted genes.

Your project file already has an annotation track of known imprinted genes. This was taken from http://www.geneimprint.com/site/genes-by-species and we are going to use this to annotate our hits.

Select your 25% filtered list of consistent hits, and then select:

Reports > Annotated Probe Report

We're going to annotate the hits with the closest imprinted gene (up to 2kb away)



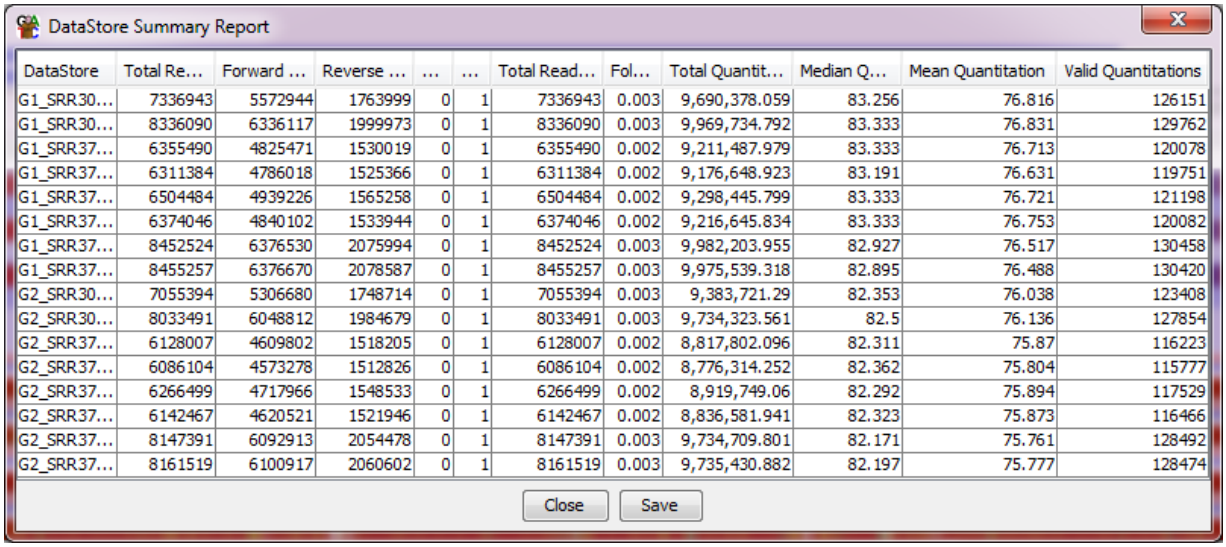

Once you've got the report window up you can sort the hits by their difference (since this was the value annotated by the differences filter).

If you look at the most different hits, how many of them match a known imprinted gene? Are there any strong hits which don't match the known imprinted genes? If so, is this just an annotation problem, or are there really potentially new imprinted genes to discover?

## Example Figures

The figures below show some of the views you should have created when you run this exercise. You can check that yours look like these.

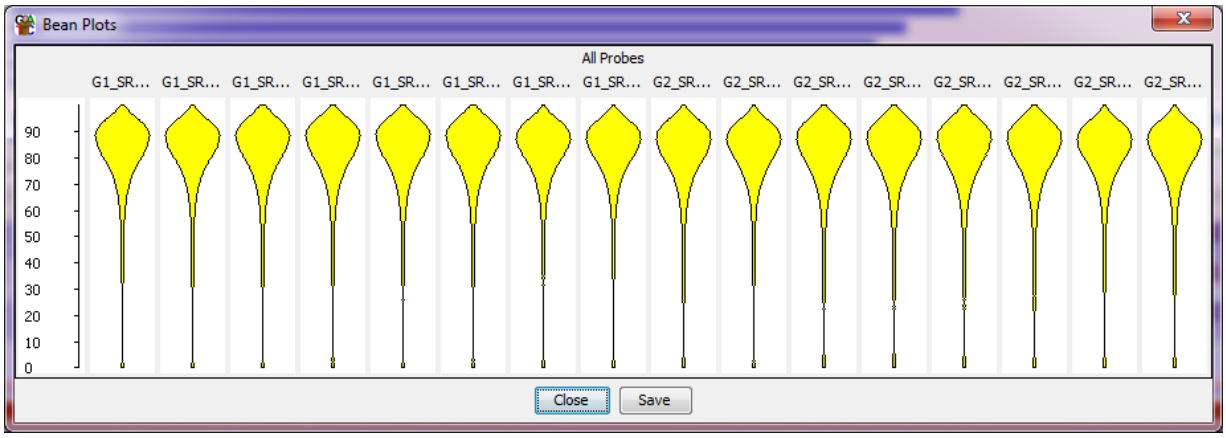
### Exercise 2: Quantitation

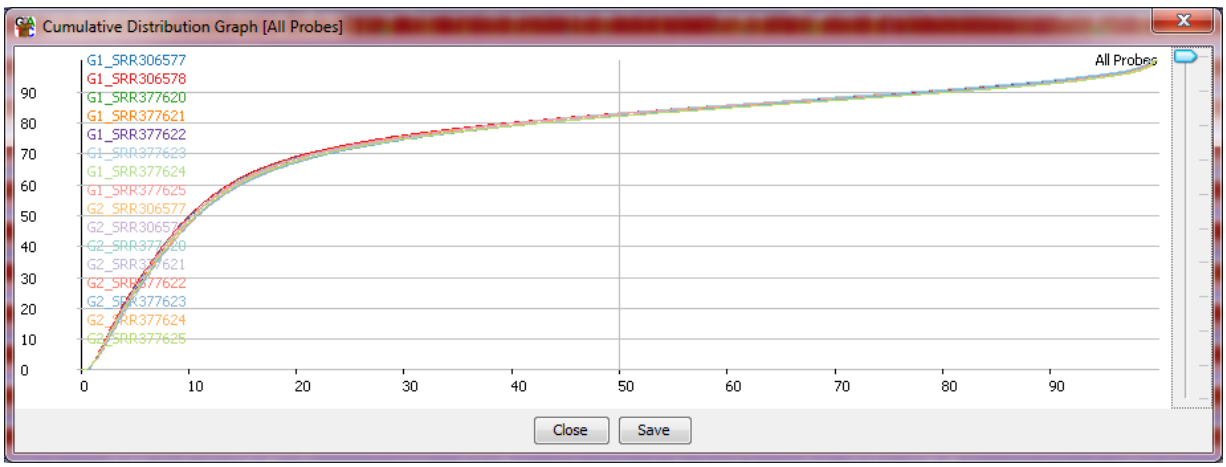



| DataStore | Total Re... | Forward ... | Reverse ... | ... | ... | Total Read... | Fol... | Total Quantit... | Median Q... | Mean Quantitation | Valid Quantitations |
|---|---|---|---|---|---|---|---|---|---|---|---|
| G1_SRR30... | 7336943 | 5572944 | 1763999 | 0 | 1 | 7336943 | 0.003 | 9,690,378.059 | 83.256 | 76.816 | 126151 |
| G1_SRR30... | 8336090 | 6336117 | 1999973 | 0 | 1 | 8336090 | 0.003 | 9,969,734.792 | 83.333 | 76.831 | 129762 |
| G1_SRR37... | 6355490 | 4825471 | 1530019 | 0 | 1 | 6355490 | 0.002 | 9,211,487.979 | 83.333 | 76.713 | 120078 |
| G1_SRR37... | 6311384 | 4786018 | 1525366 | 0 | 1 | 6311384 | 0.002 | 9,176,648.923 | 83.191 | 76.631 | 119751 |
| G1_SRR37... | 6504484 | 4939226 | 1565258 | 0 | 1 | 6504484 | 0.002 | 9,298,445.799 | 83.333 | 76.721 | 121198 |
| G1_SRR37... | 6374046 | 4840102 | 1533944 | 0 | 1 | 6374046 | 0.002 | 9,216,645.834 | 83.333 | 76.753 | 120082 |
| G1_SRR37... | 8452524 | 6376530 | 2075994 | 0 | 1 | 8452524 | 0.003 | 9,982,203.955 | 82.927 | 76.517 | 130458 |
| G1_SRR37... | 8455257 | 6376670 | 2078587 | 0 | 1 | 8455257 | 0.003 | 9,975,539.318 | 82.895 | 76.488 | 130420 |
| G2_SRR30... | 7055394 | 5306680 | 1748714 | 0 | 1 | 7055394 | 0.003 | 9,383,721.29 | 82.353 | 76.038 | 123408 |
| G2_SRR30... | 8033491 | 6048812 | 1984679 | 0 | 1 | 8033491 | 0.003 | 9,734,323.561 | 82.5 | 76.136 | 127854 |
| G2_SRR37... | 6128007 | 4609802 | 1518205 | 0 | 1 | 6128007 | 0.002 | 8,817,802.096 | 82.311 | 75.87 | 116223 |
| G2_SRR37... | 6086104 | 4573278 | 1512826 | 0 | 1 | 6086104 | 0.002 | 8,776,314.252 | 82.362 | 75.804 | 115777 |
| G2_SRR37... | 6266499 | 4717966 | 1548533 | 0 | 1 | 6266499 | 0.002 | 8,919,749.06 | 82.292 | 75.894 | 117529 |
| G2_SRR37... | 6142467 | 4620521 | 1521946 | 0 | 1 | 6142467 | 0.002 | 8,836,581.941 | 82.323 | 75.873 | 116466 |
| G2_SRR37... | 8147391 | 6092913 | 2054478 | 0 | 1 | 8147391 | 0.003 | 9,734,709.801 | 82.171 | 75.761 | 128492 |
| G2_SRR37... | 8161519 | 6100917 | 2060602 | 0 | 1 | 8161519 | 0.003 | 9,735,430.882 | 82.197 | 75.777 | 128474 |

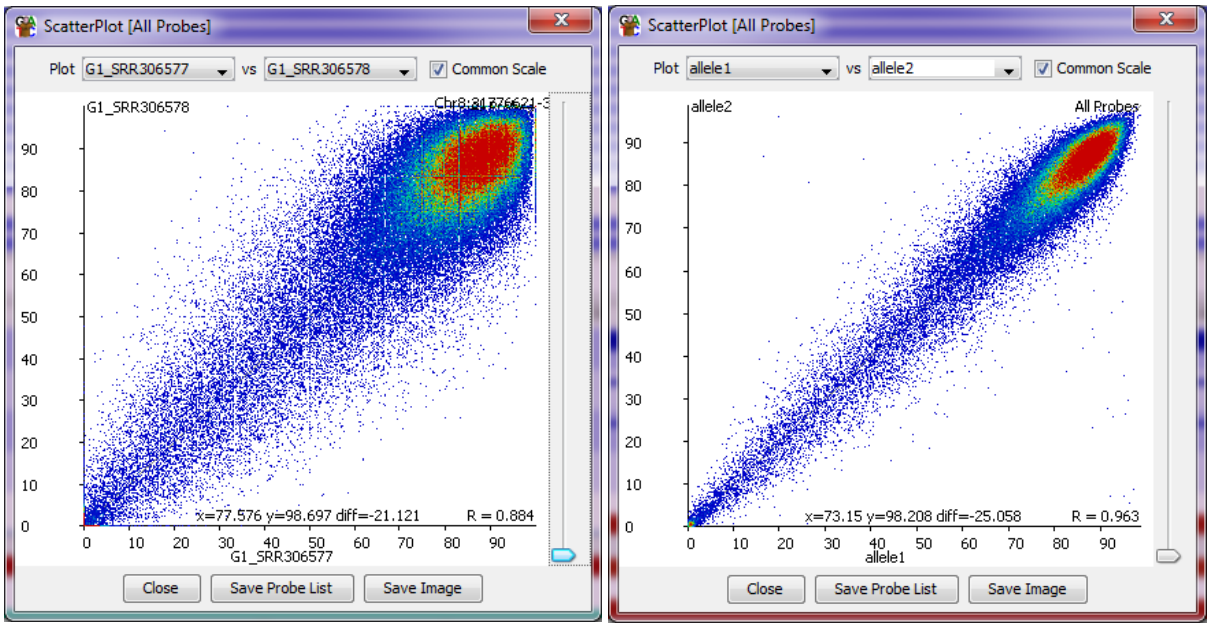
The datastore summary report shows that over 90% of all (136,465) of the probes were successfully quantitated, so the quantitation stringency used was appropriate for this data.

### Exercise 3: Data Inspection



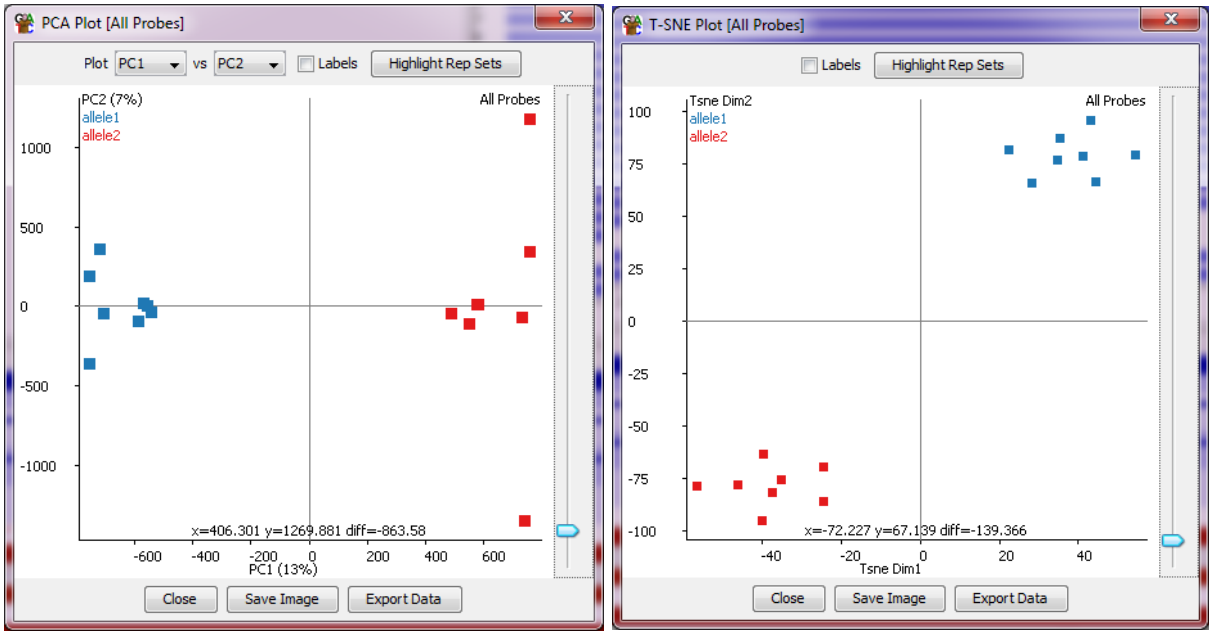



The distributions of methylation values show that the vast majority of the genome is highly (>=80%) methylated, with only a small subset staying unmethylated. The Methylation profiles of all of the samples are extremely similar to each other, showing there are no global differences to explain or normalise and showing that the analysis can proceed using the current quantitations.




The scatterplots show some noise when comparing replicates within the same sample (left), but the differences seen on an individual comparison are mostly not widespread across other replicates.

When comparing the two replicate sets we can see a clearer picture, but with some large changes in a small number of points. When looking at the underlying quantitations we generally see very similar changes in all of the replicates suggesting that these differences should reflect real biology.

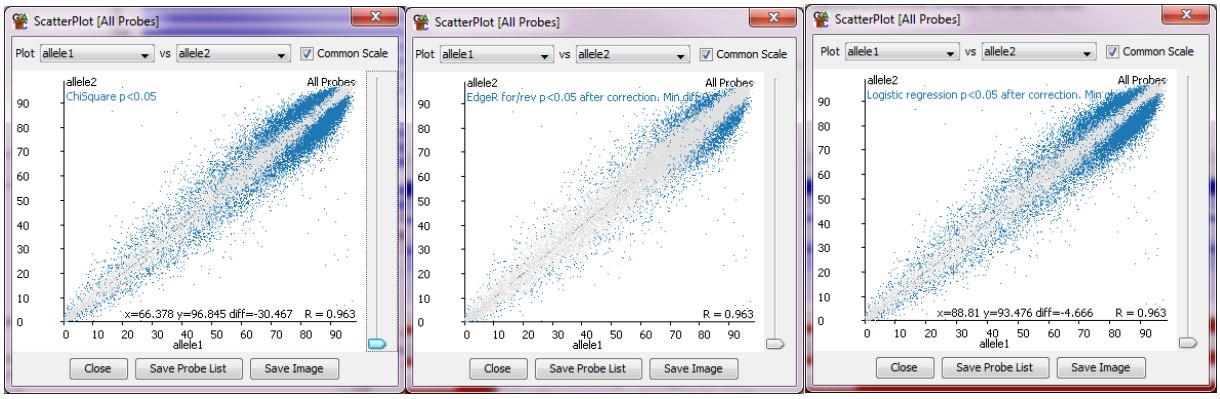



Both PCA and tSNE plots show a clear and consistent difference between the replicates from the two alleles suggesting that the differences in methylation overall divide nicely into the experimental groups we expect.

## Exercises 4,5 and 6: Differential Expression

The number of hits found using the different statistical tests should be:

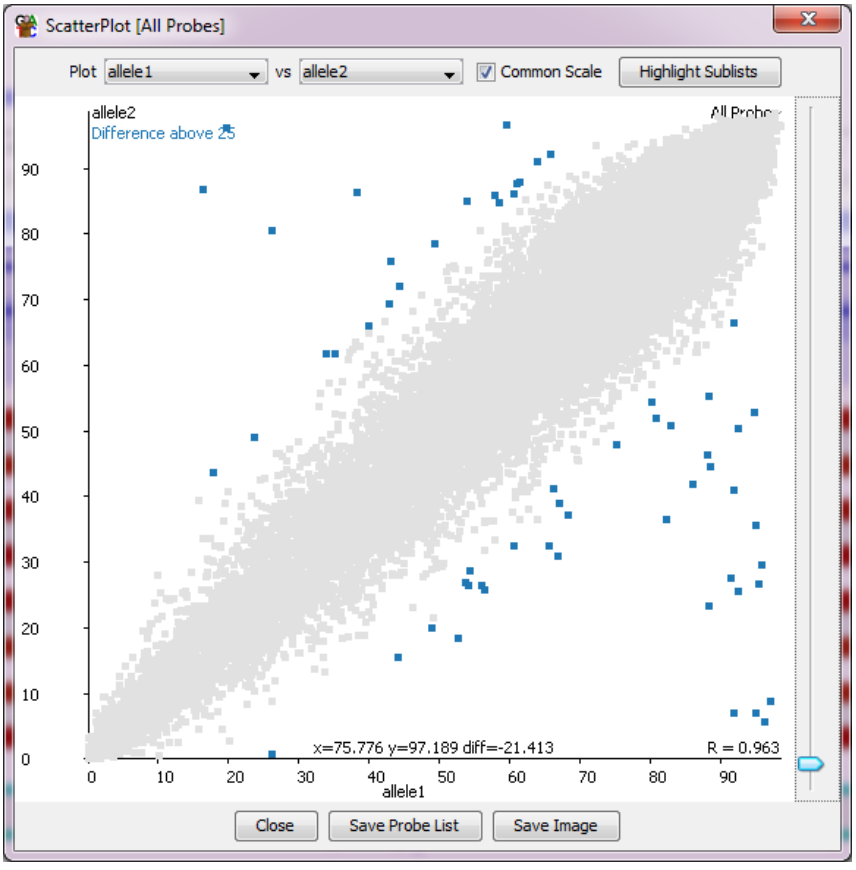
- Chi-Square: 19496
- Logistic Regression: 17485
- EdgeR: 6339




All of the sets of hits look sensible when highlighted on a scatterplot of the entire data set.

When collating the hits 5483 were present in the hits from all techniques.

After applying the 25% absolute difference cutoff then only 60 probes should remain.




## Exercise 8: Reporting

When the report is made, most of the top hits should match to known imprinted genes.

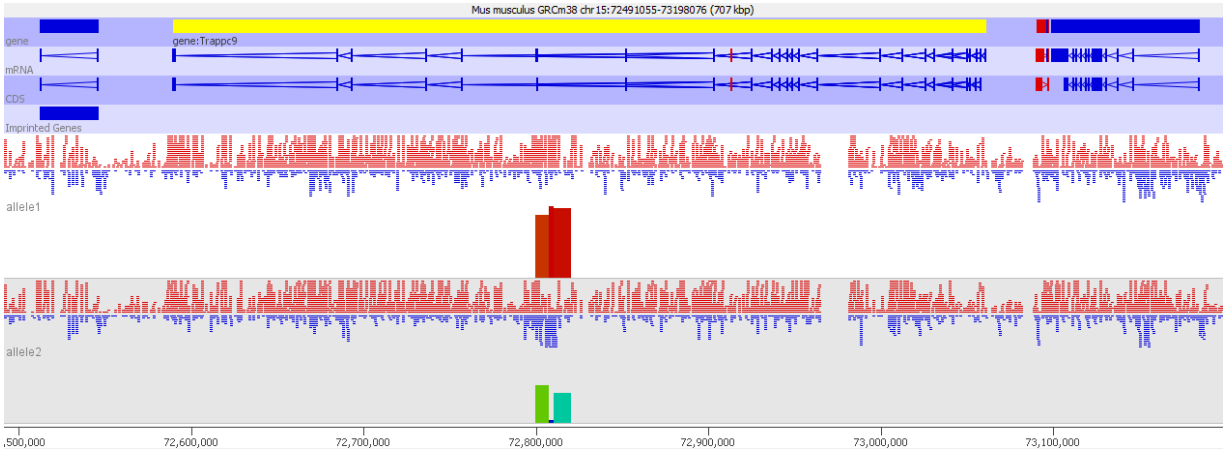
Annotated Probe Report

| Probe | Chromo... | Start | End | Probe ... | Diff... | Feature | ID | Descrip... | Featur... | Type | Orienta... | Distance | allele1 | allele2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr15:7... | 15 | 72807234 | 72810276 | | 90.735 | | | | | | Not found | 0 | 96.344 | 5.609 |
| Chr17:1... | 17 | 12741767 | 12744741 | | 88.447 | Igf2r | ENSMUS... | insulin-lik... | - | Imprinte... | overlapping | 0 | 97.22 | 8.773 |
| Chr2:17... | 2 | 174295753 | 174299082 | | 87.929 | Gnas | ENSMUS... | GNAS (g... | + | Imprinte... | overlapping | 0 | 95.005 | 7.076 |
| Chr6:30... | 6 | 30734310 | 30738193 | | 84.949 | Mest | ENSMUS... | mesoder... | + | Imprinte... | overlapping | 0 | 92.024 | 7.074 |
| Chr2:17... | 2 | 174278824 | 174286133 | | 76.379 | Nespas | ENSMUS... | neuroen... | - | Imprinte... | overlapping | 0 | 19.718 | 96.097 |
| Chr7:14... | 7 | 142577609 | 142584797 | | 70.227 | H19 | ENSMUS... | H19 feta... | - | Imprinte... | overlapping | 0 | 16.48 | 86.707 |
| Chr2:17... | 2 | 174316815 | 174328042 | | 68.873 | Gnas | ENSMUS... | GNAS (g... | + | Imprinte... | overlapping | 0 | 95.504 | 26.63 |
| Chr7:14... | 7 | 143295469 | 143303701 | | 66.881 | Kcnq1 | ENSMUS... | potassiu... | + | Imprinte... | overlapping | 0 | 92.559 | 25.678 |
| Chr7:67... | 7 | 6729472 | 6743396 | | 66.247 | Peg3 | ENSMUS... | paternall... | - | Imprinte... | overlapping | 0 | 95.844 | 29.598 |
| Chr6:47... | 6 | 4738481 | 4748382 | | 65.079 | Sgce | ENSMUS... | sarcogly... | - | Imprinte... | overlapping | 0 | 88.321 | 23.242 |
| Chr7:67... | 7 | 6711416 | 6729471 | | 63.986 | Peg3 | ENSMUS... | paternall... | - | Imprinte... | overlapping | 0 | 91.602 | 27.616 |
| Chr6:30... | 6 | 30738194 | 30803026 | | 59.487 | Mest | ENSMUS... | mesoder... | + | Imprinte... | overlapping | 0 | 95.02 | 35.534 |
| Chr9:89... | 9 | 89868739 | 89885125 | | 54.339 | | | | | | Not found | 0 | 26.283 | 80.621 |
| Chr15:7... | 15 | 72810284 | 72820422 | | 50.925 | | | | | | Not found | 0 | 91.901 | 40.976 |
| Chr12:1... | 12 | 109541168 | 109551041 | | 47.991 | Meg3 | ENSMUS... | maternal... | + | Imprinte... | overlapping | 0 | 38.285 | 86.276 |
| Chr11:1... | 11 | 12014455 | 12036723 | | 46.007 | Grb10 | ENSMUS... | growth f... | - | Imprinte... | overlapping | 0 | 82.444 | 36.437 |
| Chr7:12... | 7 | 128684615 | 128694573 | | 44.141 | | | | | | Not found | 0 | 86.06 | 41.919 |
| Chr2:15... | 2 | 152683221 | 152690703 | | 44 | H13 | ENSMUS... | histocom... | + | Imprinte... | overlapping | 0 | 88.538 | 44.538 |
| Chr18:3... | 18 | 36978984 | 36981518 | | 42.158 | | | | | | Not found | 0 | 92.509 | 50.351 |
| Chr10:1... | 10 | 13090449 | 13101377 | | 41.971 | Plagl1 | ENSMUS... | pleiomor... | + | Imprinte... | overlapping | 0 | 94.779 | 52.808 |

Close Save

In this version of the report I collapsed the replicate sets in the Data Track Display options so that we only see the mean values for allele1 and allele2 rather than the values for all of the individual replicates.

Of the top hits which aren't annotated to an Imprinted gene, there are a cluster of hits to the Trappc9 gene, which look like genuine methylation differences.




There is also a hit to the Rasgrf1 gene, which is a known imprinted gene, but where the DMR was slightly too far away from the gene to be picked up by the annotation process.