Babraham Bioinformatics

Exercises: RNA-Seq Splicing Analysis

Version 2018-04

Licence

Introduction

In this follow-on exercise to the basic RNA-Seq exercises we will look at the same data, but this time trying to find evidence for differential splicing between the two conditions. This might be to extend our list of candidate differentially expressed genes, or it might be to specifically look for differential splicing regulation.

Software

The software which will be used in this session is listed below. In this case we are starting with data which has already been mapped and loaded into a seqmonk project so we're only looking at the software we're using for visualisation and statistical analysis:

- SeqMonk (http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/)
- R (http://www.r-project.org/)
- DESeq2 – part of bioconductor (http://www.bioconductor.org/)

Data

This data is the exact same BAM files used for the original differential expression exercise. The only difference is that this time the import options used were different, and the "Import introns rather than exons" option was selected so that we are looking at the set of observed splice junctions.

The data should be provided to you in a SeqMonk project file called "mouse_mapped_data_splicing.smk" which you can open with `File > Open Project`.

Exercise 1: Initial Exploration

We've already looked at this dataset so we shouldn't need to do too much more with it at this stage. What you can do though is to spend a few minutes looking through the data you are going to work with. You should be able to see that the 'reads' you get in this dataset are actually the intron positions from reads which spanned a splice junction.

You should be able to see that the vast majority of the introns match with the known, expected structure in the reference genome we're working with. You will also see that in most genes there will be a single splicing structure for each gene which will dominate in each sample.

Exercise 2: Probe Generation and Quantitation

There are two possible approaches to analysing this data. We could quantitate all of the splice junctions we observed, or we could limit ourselves to those with match exactly with what we expect from the annotation we have in the genome.

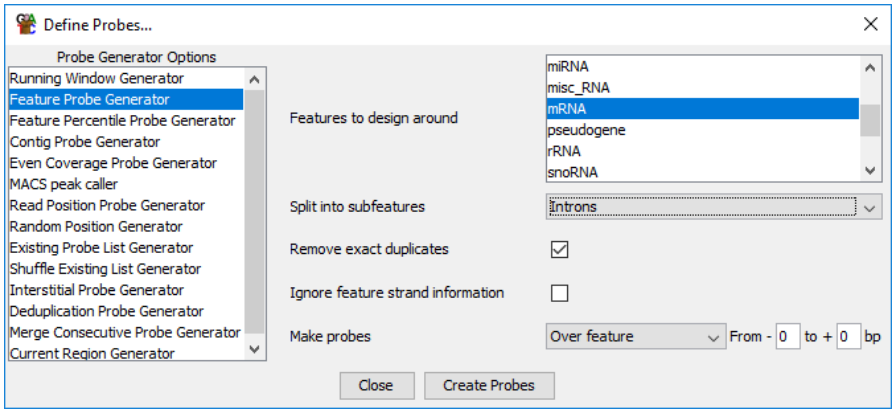
In this exercise we'll use the second approach, which could mean that we might potentially miss any completely novel splice sites, but it will make the overall analysis a bit cleaner, and would be fine if we're only really interested in looking at known genes and transcripts.

Step 2.1 – Generating Intron Probe and Quantitating

We will therefore measure specifically over annotated introns. To do this we're going to use the Feature Probe Generator and the mRNA track.

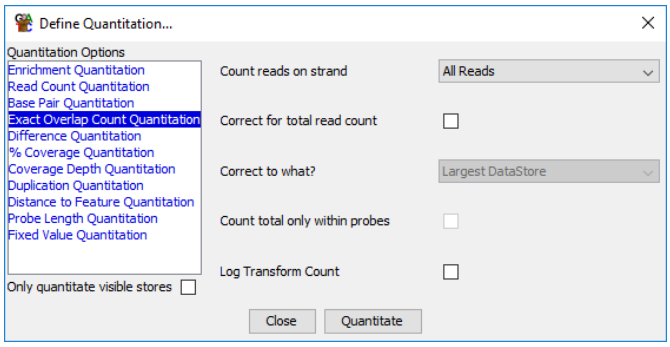
`Data > Define Probes > Feature Probe Generator`

We will put probes over all introns and then quantitate the reads which match with them. To make intronic probes we use the mRNA track, and then opt to use the Split into SubFeatures option to generate intron probes. We are going to remove duplicate probes, since the same intron will often appear in several different transcripts.



Once we've generated the probes we need to quantitate them. The appropriate quantitation here is the Exact Overlap Quantitation is the appropriate one to use since the reads should exactly mirror the start and end positions of the corresponding intron probe.

For an initial analysis we will just use the default options which will generate raw, uncorrected counts.



We could limit by strand since this is a directional library, but we're not going to get any places where introns from both strands will align exactly so it doesn't really matter.

Once we have an initial quantitation you can look at the quantitated data in the chromosome view. You should be able to see a roughly even representation of introns across the length of larger transcripts. You will also see more clearly where there are cases where more than one structure is present, since you should see a local change in quantitated values at the point where the two splice variants diverge.

After you've looked around the quantitation we can also do a QC check where we find out how much of the data is captured by using the annotated introns. The easiest way to do this is to run:

`Reports > Data Store Summary Report`

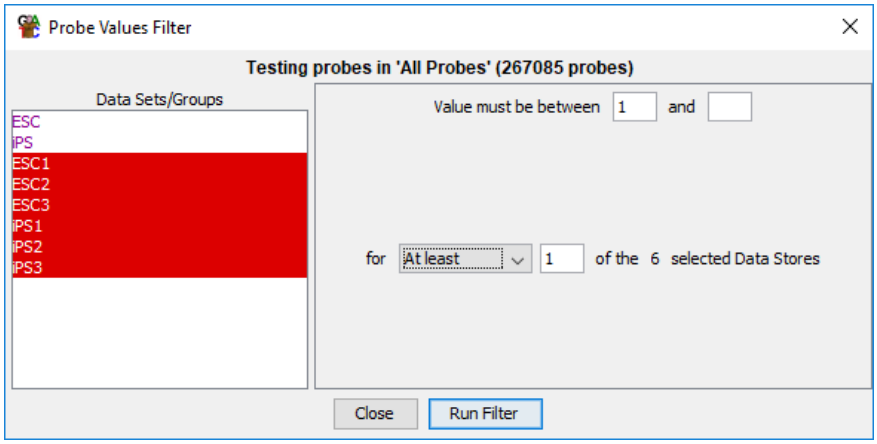
This will provide a lot of potentially useful metrics, but the relevant ones for us are the "Total Read Count" (the first quantitative column) and the "Total Quantitation" which is the sum of the intron counts. The difference between these two represents the number of non-canonical splice sites which were seen somewhere in the data. You can either just look at these to determine if the quantitation strategy we have used has been effective, or you could even export the table and then plot the two columns as a figure in R if you wanted a more formal answer.

Step 2.2 – Removing unmeasured introns

To save testing introns which are completely unmeasured within our data we can use a simple values filter to remove those which were never observed anywhere in any of the samples. As we're still quantitated by raw counts we can simply select only probes which have a value of 1 or more in any sample. To do this select:

`Filtering > Filter on Values > Individual Probes`

We'll filter on the individual data sets rather than the replicate sets, which is the most lax filter we can apply.



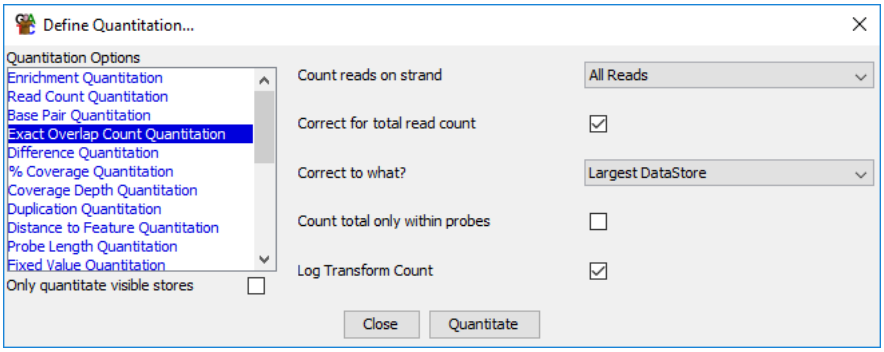
Save the list to use later, and see how many of the original introns remain after this filter.

Exercise 3: Exploration

Before we do any statistical analysis we should do a bit more exploration. As we're going to look at the quantitations we need to re-quantitate the data using log transformed, normalised measures. To do this simply go back to:

`Data > Quantitate Existing Probes > Exact Overlap Count Quantitation`

..and turn on the normalisation and log transformation options.



You can now perform some of the same exploration steps you used before. Comparing the distributions of the values using the `Cumulative Distribution Plot` and the relatedness of the samples using a `Scatterplot`. When performing these parts use your filtered list of measured introns to get rid of the large number of zeros within the data.

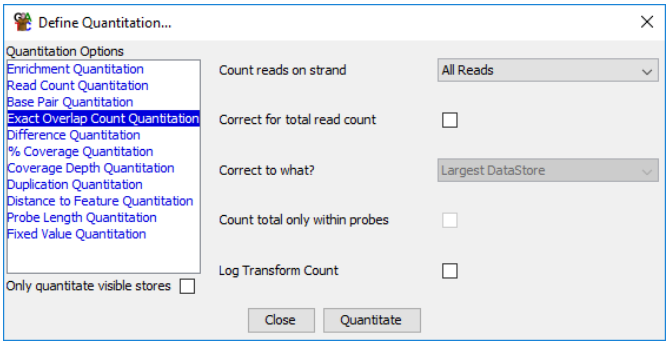
When looking at the data in a scatterplot look at some of the more extreme changes and see if they are localised to a specific intron, or whether there is a wider change in the gene.

Exercise 4: Statistical Analysis for Differential Splicing

In this part we're going to use two different statistical tests to identify potentially interesting introns. We will use a standard DESeq2 workflow to identify introns with differential abundance (either because of expression or splicing changes) and we'll use logistic regression to identify pairs of introns with differing ratios between the two conditions.

Step 4.1 - Requantitation

As we're going to be running count based statistics we need to go back and re-quantitate the data as completely raw counts.



Step 4.2 – Run DESeq

As before, run:

`Filtering > Filter by Statistical Test > Count Based Statistics > Replicated Data > DESeq2`

Select the two replicate sets and run the test. Save the list of hits.

Step 4.3 – Run Logistic Regression

The splicing logistic regression filter can be found at:

`Filtering > Filter by Statistical Test > Proportion based statistics > Replicated Data > Logistic Regression Splicing`

Make sure you definitely use the Splicing form of the logistic regression filter. Select the two replicate sets and run the test.

Step 4.4 – Review the DESeq2 hits

Having run the tests we want to be able to review the hits. This means we're going to need to move back to normalised log transformed counts. Go back to the Exact Overlap Count Quantiation and turn the normalisation and log transformation back on.

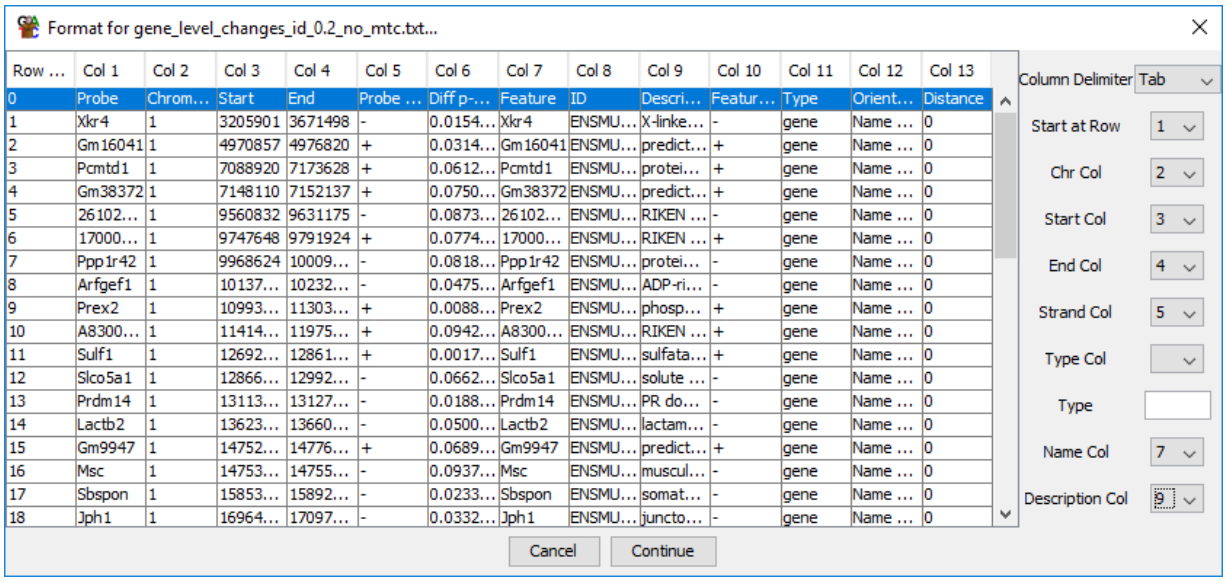
Start by reviewing the DESeq2 hits. Draw a scatterplot of the two replicate sets against each other and highlight the DESeq hits onto it and check that they make sense.

Since our DESeq2 hits could come from either expression or splicing changes we want to see if any of the hits look like they are specifically related to splicing changes. To determine this you have been provided with a list of genes which changed in the gene level analysis. This is a very lax filtered list which should capture any gene with any appreciable level of overall change. For a specific splicing hit we would want one of our DESeq hits to **not** overlap with the genes in the global change list.

To import the new annotations into our project we use:

`File > Import Annotation > Text (Generic)`

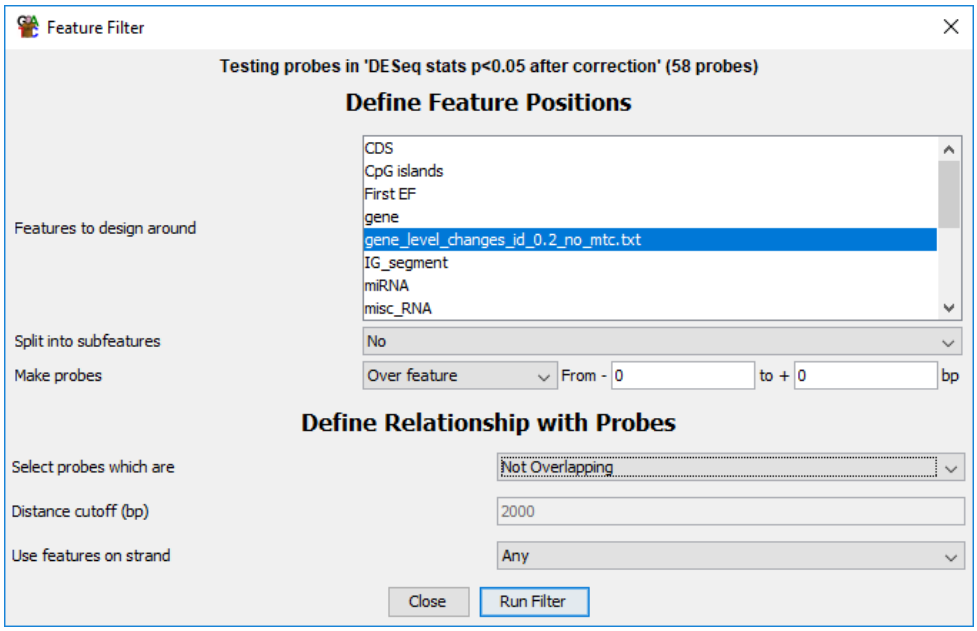
.and then select the gene_level_changes_id_0.2_no_mtc.txt file you were given. You will need to tell SeqMonk which column is which in the text file.



Once we've imported this we should see a new annotation track with these gene hits in it.

To find out whether we have any novel DESeq hits we can use a feature filter on our DESeq hit probe list and we're going to look for probes which don't overlap the genes we just imported.

`Filtering > Filter by Features`

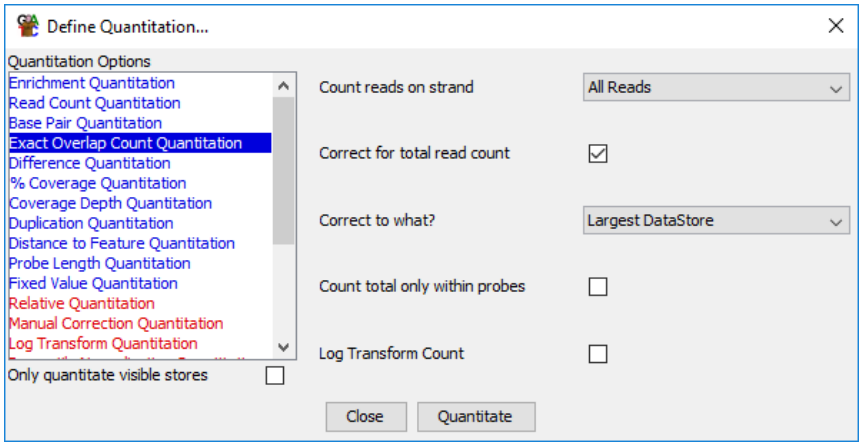


Do any of your hits pass this filter?

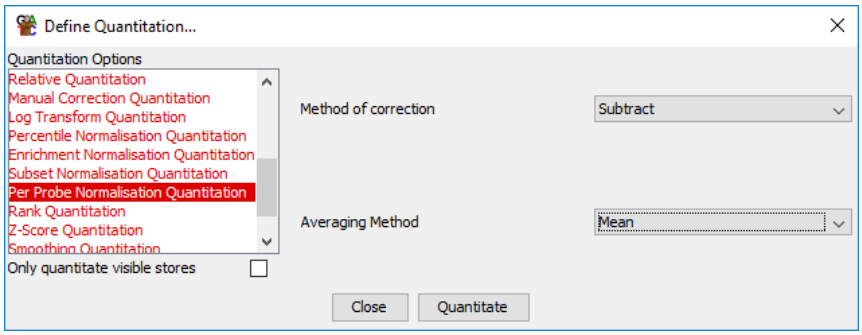
Step 4.5 Reviewing Logistic Regression Hits

Since we're looking at a change in ratios we can adjust our quantitation to make these as easy as possible to see. What we're going to do is to do a normalised linear quantitation, followed by a per-probe normalisation so what we're looking at is the difference in quantiation in each sample from the mean values across all samples.

Start by rerunning the exact overlap quantitation using linear normalised counts.



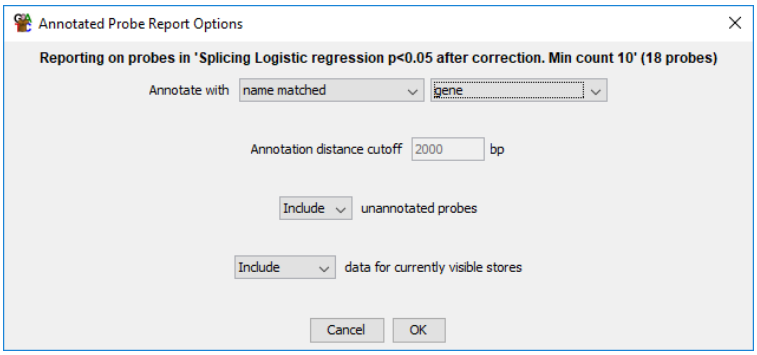
After that, go back to the quantitation options and use the per-probe normalisation quantitation. We will subtract the mean value across all 6 samples from each individual measurement.



Once we've done this we can create a report from our Logistic Regression hits and then use that to help us review them. Firstly, use

`Reports > Annotated Probe Report`

..on your logistic regression hits. Annotate them with the names of the genes from which they originally came.



If you sort the report by FDR you should see the top hits. Have a look at the best hits and see if you can see why they would have been selected, and whether you think the changes you're seeing would be worthwhile pursuing.

Example Plots

So you know what you should be seeing here are copies of the plots you should generate in this practical:

Exercise 2

Step 2.1 – Efficiency of intron measurement

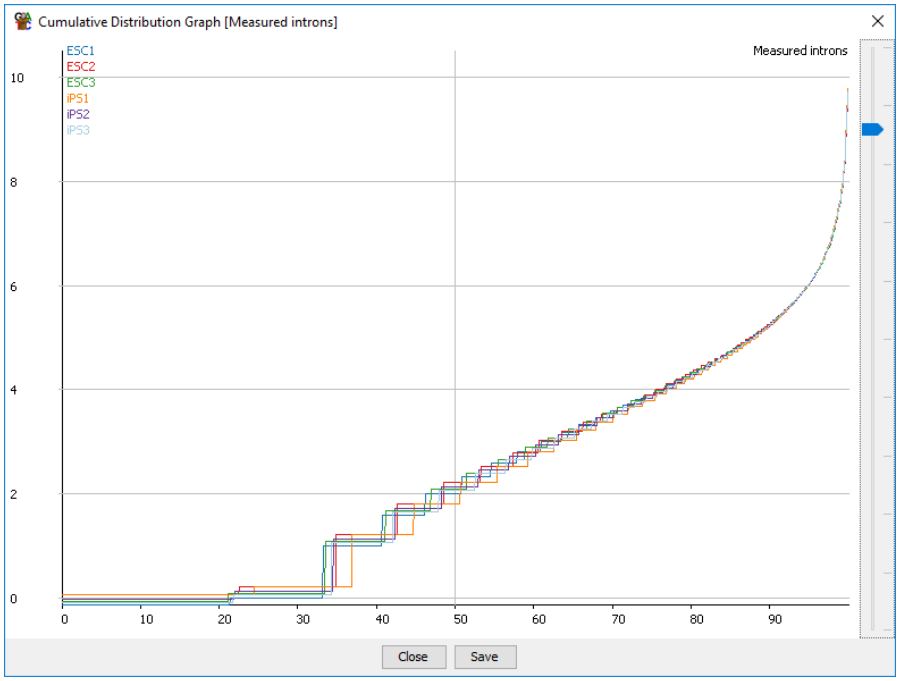
DataStore Summary Report

DataStore	Total Read C...	Forward Re...	Reverse Re...	...	Mean Read ...	Total Read Length	Fold Coverage	Total Quantitation	Me...	Mean Qua...	Valid Qua...
ESC1	2742270	1363954	1378316	0	1,990	5457855390	2.002	2,731,271	0	10.226	267085
ESC2	2378240	1179565	1198675	0	1,989	4732473974	1.736	2,367,257	0	8.863	267085
ESC3	2599285	1289250	1310035	0	2,009	5222742568	1.916	2,588,038	0	9.69	267085
iPS1	2371694	1170070	1201624	0	1,881	4463278623	1.638	2,362,122	0	8.844	267085
iPS2	2508103	1240556	1267547	0	1,978	4961540505	1.82	2,497,589	0	9.351	267085
iPS3	2626057	1298310	1327747	0	1,900	4990764622	1.831	2,613,454	0	9.785	267085

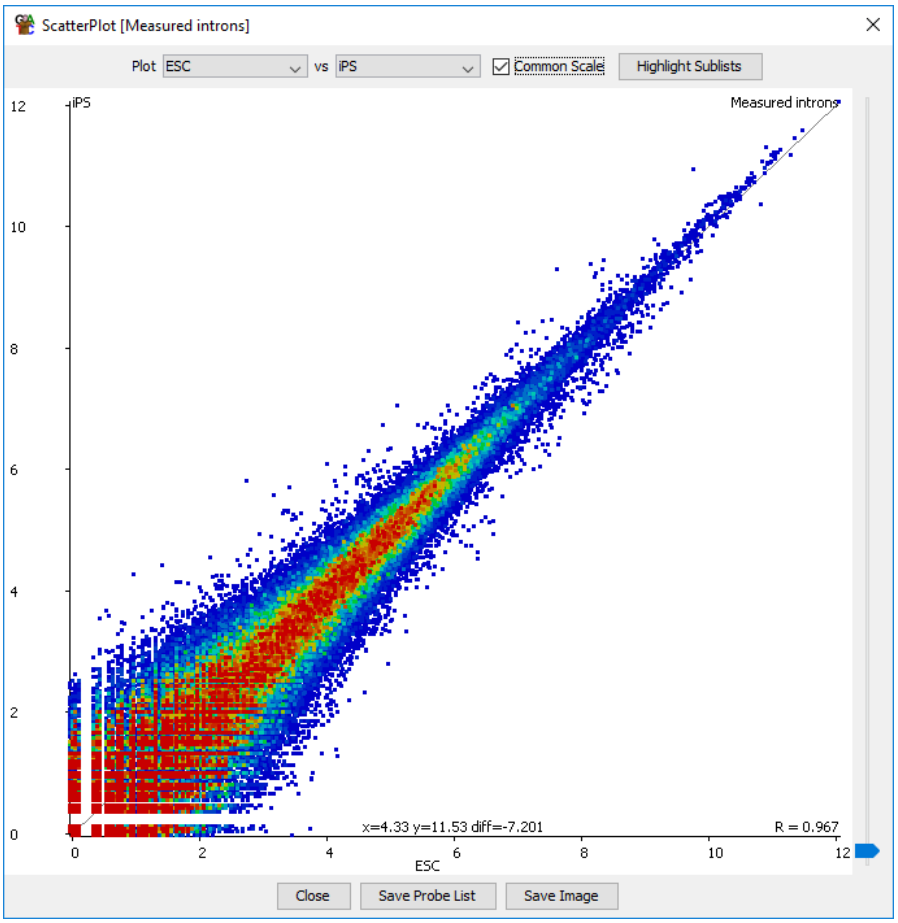
Close Save

You should see that the vast majority of the data (>99%) matches exactly against an already annotated intron. Given this, it doesn't seem that there would be much to gain from analysing the non-standard introns.

Exercise 3: Exploration



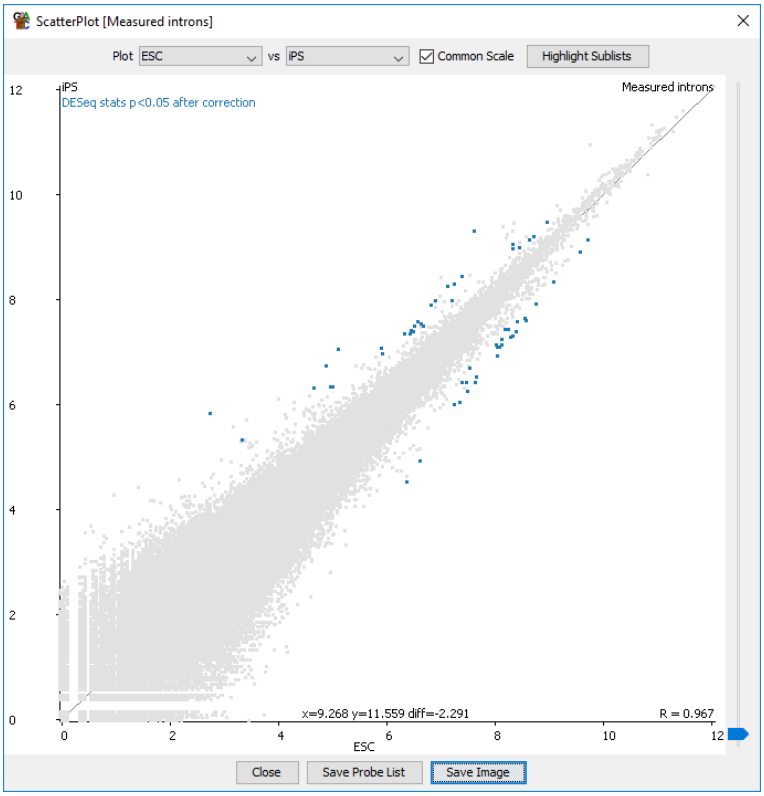
The cumulative distribution plot shows that the data are very well normalised.



The scatterplot shows the two sample groups are very similar overall, but with some individual changes.

Exercise 4: Statistical Analysis

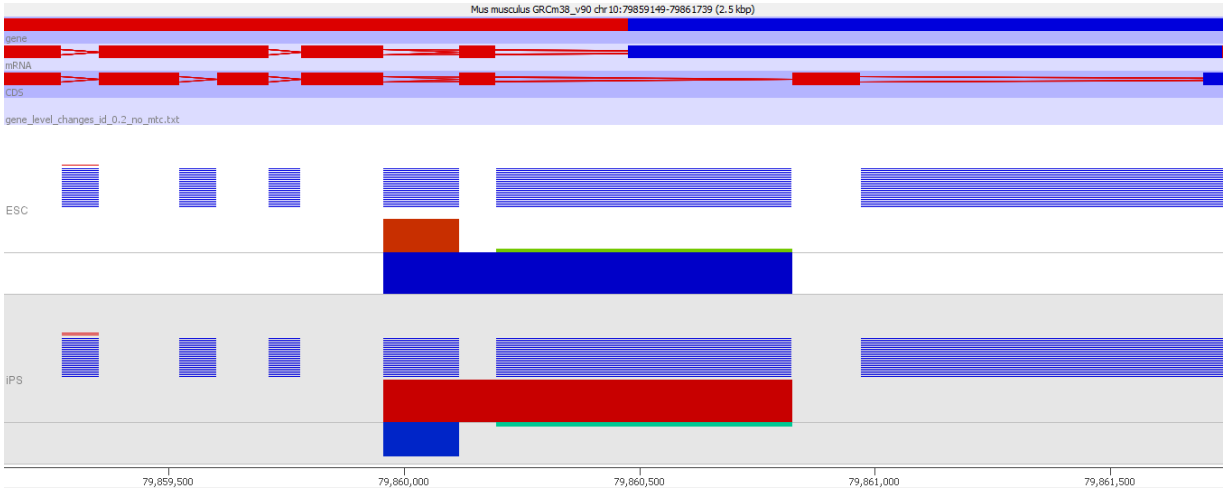
Step 4.4 Reviewing DESeq hits



There aren't many DESeq hits, but the ones which have been found seem to make sense.

None of the hits fall outside genes which changed overall, so we can't see any great splicing specific candidates to pursue.

Step 4.5 Reviewing Logistic Regression Hits



This is the view of the top hit in the logistic regression. I've collapsed the replicate sets down to a single value to make the overall change clearer.

What you can hopefully see is that in the ESC samples the tendency is for the splicing to occur twice, leaving an intermediate exon. In the iPS samples the single splice is more frequent, meaning that the intermediate exon is excluded more frequently than in the ESC.

You will also see that the consistency between the replicates isn't great for this change, and none of the hits here overlap with the DESeq hits. These results are therefore not very convincing so it would be worth doing some more specific validation on them before committing to any extensive work on the back of these results.