Babraham Bioinformatics

Exercises:
Analysing ChIP-Seq data

Version 2018-03-2

Licence

Introduction

In this session we will go through the differential enrichment analysis of a ChIP-Seq experiment. This will include:

- Quality control
- Peak Calling
- Quantitation and Normalisation
- Differential enrichment analysis and validation of results.

Software

The software which will be used in this session is listed below. Software which requires a linux environment is indicated by an asterisk*:

- SeqMonk (http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/)
- R (https://cran.r-project.org/)
- LIMMA (https://www.bioconductor.org/packages/release/bioc/html/limma.html)

Data

The data in this practical comes from GEO accession GSE69646 and represents CTCF ChIP from Naïve and Primed Embryonic Stem cells and their accompanying inputs.

The data has already been mapped to the GRCh38 genome using bowtie2 and was imported into seqmonk using the default import parameters (a MAPQ filter for >=20). All reads were extended by 250bp to more closely reflect the true insert size of the data.

Exercise 1: Quality Control

Step 1.1

Using what you learned in the previous exploration exercise, look at the data presented here. The project has already had 500bp probes tiled over the genome and a linear read count quantitation has been performed (you can repeat this part if you like). Try to answer the following questions.

1. Does the data show obvious enrichment?
2. Do the enriched regions show a characteristic pattern of strand bias in the reads? If not, why not?
3. Is there any evidence for PCR duplication in the samples?
4. Do the inputs look similar enough to each other that we can consider them together rather than separately?
5. Are there regions which appear to be enriched in the input sample? If so could you identify them?
6. In the ChIP samples are the enriched regions similar in both samples? Is the level of enrichment similar across all samples? If not, does it group by condition?
7. Are the peaks narrow or broad in nature?
8. Are the peaks associated with any particular feature? Could we use features to quantitate the data or should we peak call?

Exercise 2: Peak Calling

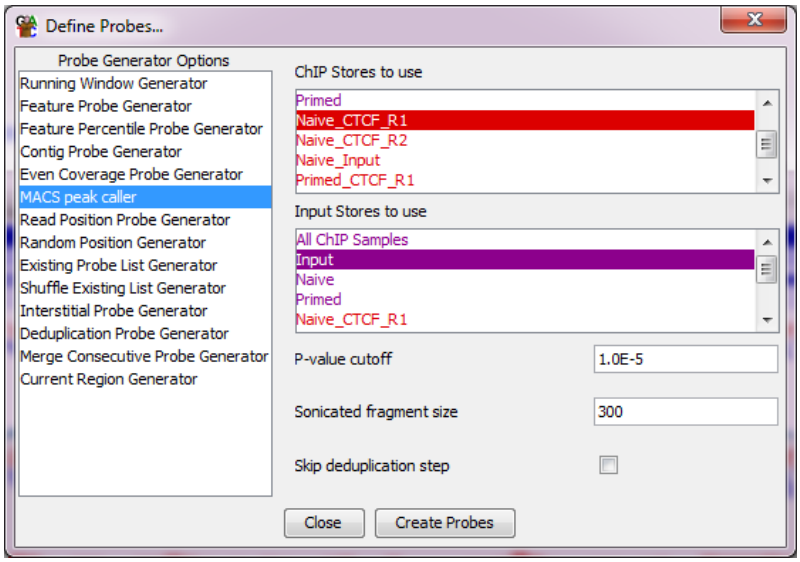
Step 2.1

In this case peak calling is appropriate. We are going to call peaks in each of the 4 ChIP samples separately and then combine them to get the full set of regions to analyse. We will be using the MACS implementation built into SeqMonk for simplicity, but we could equally have done this using the command line version of MACS and then imported the peak lists afterwards.

The steps below will need to be performed 4 times, once for each of the ChIP samples.

To generate the peak lists you go to **Data > Define Probes > MACS Peak Caller**.

In the options, the ChIP store will be the single data set (coloured red) which you want to analyse. The Input Store will be the Input Replicate Set (coloured purple) which will contain the reads from both input samples. You can leave the p-value and fragment size options set to their defaults.



After creating the probes you can use a linear read count quantitation to quantitate them. After each call have a look at the positions which have been called and see if you agree with the decisions which were made by the caller.

In order to retain the peaks from the previous sample when calling the next sample we need to turn them into an annotation track. To do this right-click on the "All Probes" Probe List and select "**Convert to Annotation Track**". Name the track after the sample you used to create the peaks.

Repeat this until you have 4 annotation tracks with the peaks from each of your ChIP samples.

Have a look at the number of peaks found in each sample. Are there any differences? What might be the cause of these?

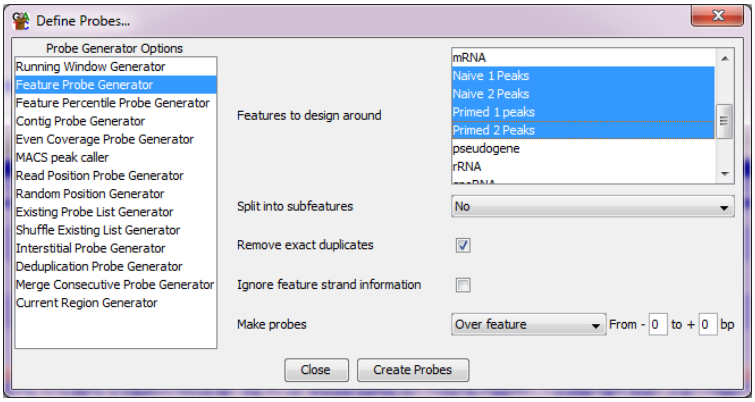
Find some locations where you don't see consistent peak calls across all 4 samples. Is there no enrichment at all in the samples where a peak wasn't called?

Step 2.2 Making probes

We now want to make measurement probes over any region which was identified as potentially interesting (ie a peak) in any of the samples. To do this we're going to use the Feature Probe Generator on the set of annotation tracks we made in Step 2.1.

Select **Data > Define Probes > Feature Probe Generator**

Select all of the peak tracks you made in Step 2.1 and make probes over those regions.

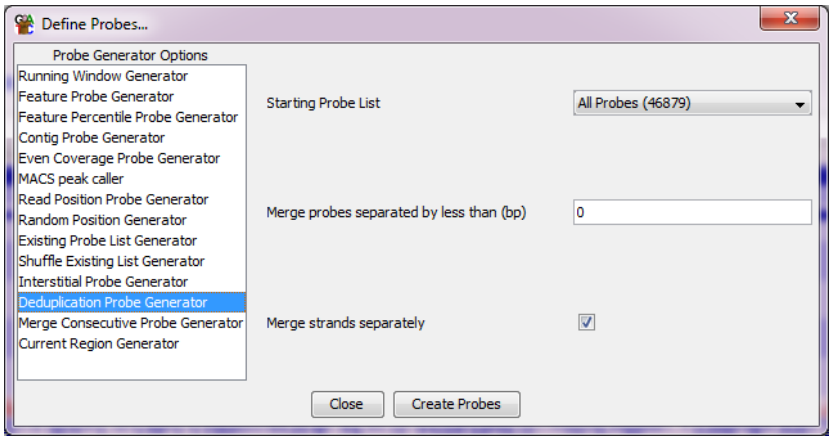


Again, you can use a read count quantitation.

The problem at this stage is that in most cases a pretty much identical peak was called in multiple samples, so we have a lot of redundancy in our peak set. To reduce this we need to deduplicate the peaks we have. To do this we can select:

Data > Define Probes > Deduplication Probe Generator

You can run the generator on its default settings which will just merge together any overlapping probes. Since our probes do not have any directionality (which is why they're all grey in the display) the option to merge strands separately doesn't do anything.



Again, you can quantitate with a read count quantitation.

Look carefully through your final set of peaks. Compare it to the individual peak tracks you have for each sample, and the data you can see and check that it looks like you have captured all of the potentially interesting places in the genome.

Exercise 3: Filtering, Quantitation and Normalisation

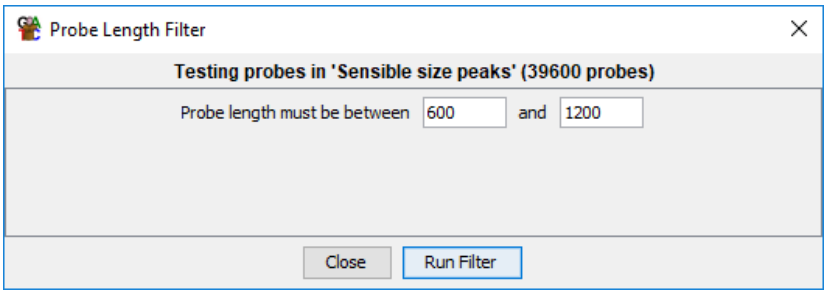
Step 3.1

Before we go too far into quantitation and normalisation there are a few things we can clean up. Specifically, there are two types of filter we can apply to give us a cleaner set of peaks to work with.

Firstly we can remove peaks which fall outside the expected size range. For CTCF we expect the binding site to be relatively small, so the total area of enrichment will be about twice the insert size, so somewhere from 600-1200bp. We can check the range of sizes we actually see using **Plots > Probe Length Histogram**.

We can remove very small peaks which could be artefacts, or very low enrichment, or large peaks which might well be cases where multiple closely spaced peaks were merged together generating a mixed signal. In a more detailed analysis we could split these apart, but here we will just discard them.

To filter on probe length we do **Filtering > Filter by Probe Length** and then set the size range we want to keep.



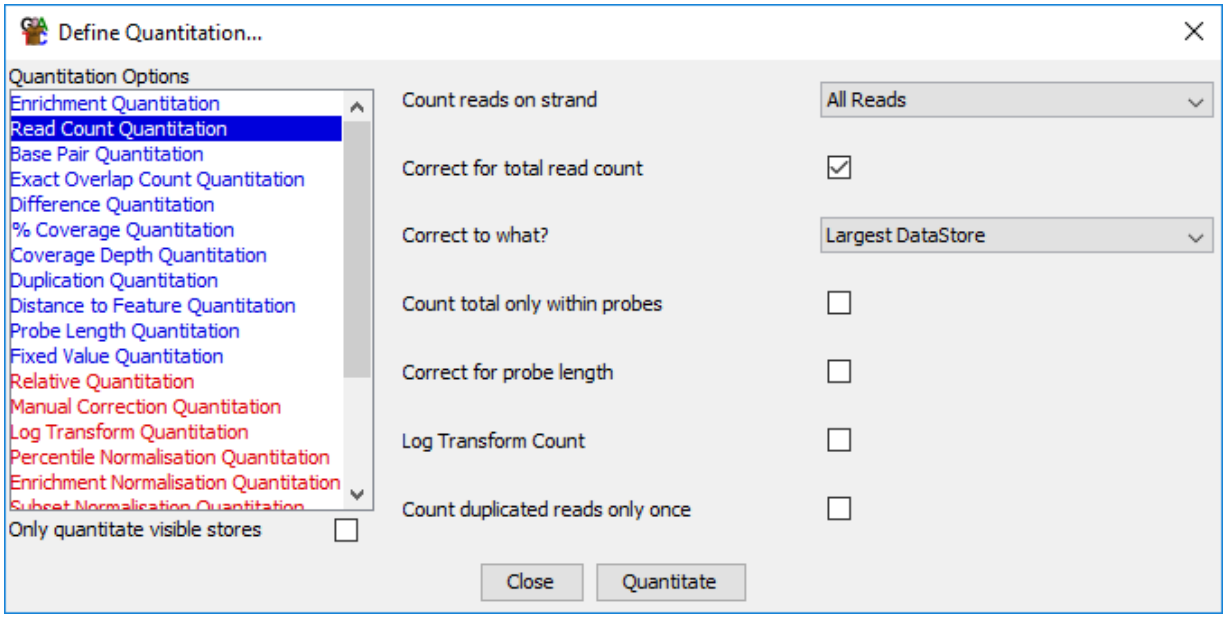
We can save the list to “Sensible size peaks”

The other things we can remove are peaks where there is an unusually high amount of signal in the input. We can inspect this by finding the input replicate set in the data view, right clicking on it and selecting “**Probe Value Histogram**”. This should show us the range of observed values. Hopefully none of them should be too extreme, but you can use a Probe Value Filter to remove peaks with unusually high input signal (you can decide on a sensible cutoff for this).

Step 3.2 Quantitation Consistency Assessment

Now we want to quantitate and normalise the values for each peak so that they are directly comparable. This will also allow us to assess whether there is a biologically relevant difference in overall enrichment between the samples.

You should already have your peaks quantitated by a linear read count quantitation, but if not you can do that now under **Data > Quantitate Existing Probes > Read Count Quantitation**.



To look at the overall enrichment you can draw a Cumulative Distribution Plot (**Plots > Cumulative Distribution Plot**). You should expect to see a large difference between the inputs and the ChIPs, but do all of the ChIPs look equally enriched? If not then is the enrichment linked to the condition?

You can also use a scatterplot (**Plots > Scatterplot**) to look at the relationship between the different ChIP samples to try to confirm what you see in the distribution summary plots.

Step 3.2 Quantitation Normalisation

You should have seen that there is a substantial difference in overall enrichment between the naïve and primed samples. This is probably a real biological difference and with more replicates we could build a very convincing case that the overall enrichment changes. Since we only have 2 replicates then this result may just be indicative, but this is likely the largest biologically relevant difference in these samples.

However, we also want to know if there are particular peaks which change in an unusual way given the overall differences between the samples. We therefore want to normalise away the global changes we see so that we can more directly do a positional analysis.

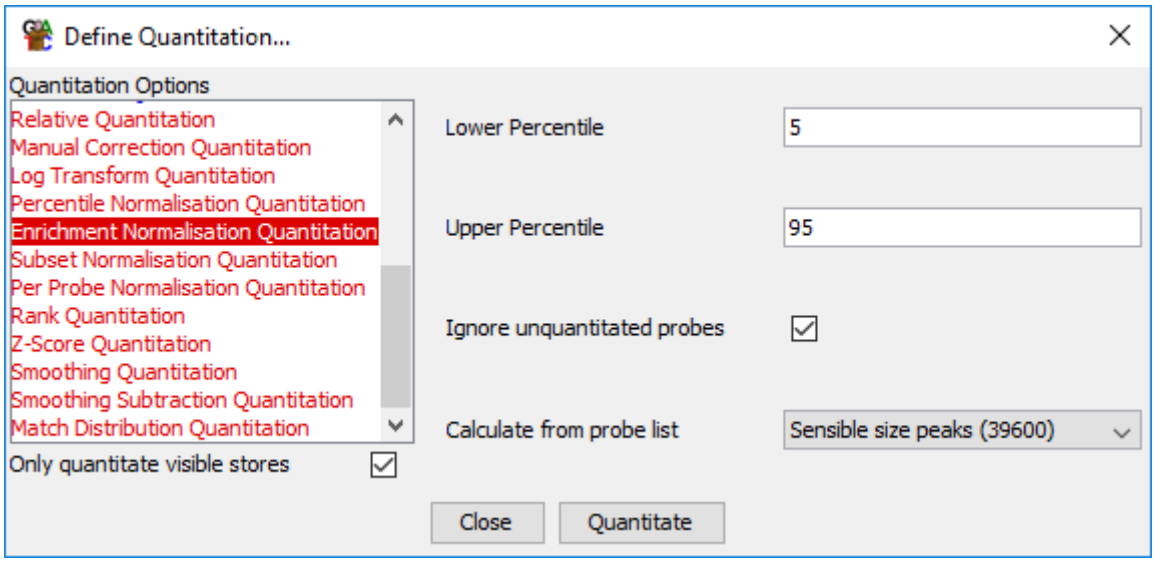
To do this we are going to use the enrichment normalisation quantitation. This sets two points of reference – a lower one for a background level, and an upper one for a highly enriched region. The tool then scales all samples between these two points.

The first thing we need to do is get rid of the input samples, since we don't want to normalise those, so use **View > Data Track Display** to remove them from the current view.

We can then normalise the remaining 4 ChIP samples. Select

Data > Quantitate Existing Probes > Enrichment Quantitation.

We are going to normalise between the 5th and 95th percentiles. Have a look back at the cumulative distribution plot and make sure you understand why these particular values were chosen.



Note that we tick the box which says "Only quantitate visible stores" so that we don't change the input samples (which you should have removed from the view earlier).

Once the data has been normalised plot the Cumulative Distribution Plot again and see if the data looks better matched now. You can also do a scatterplot between the naïve and primed replicate sets to check that the values are now directly comparable. When looking at the plot see if you think there is an obviously separate group of points which are behaving differently between the samples, or whether you're just looking at degrees of enrichment difference.

Step 3.3 Sample Clustering

Now that we have values which are more directly comparable it would be interesting to see whether we can separate our two sample groups based on the normalised data. We have a few different ways to represent the overall releationship between the samples and you can try all of:

- Plots > Data Store Similarity > Data Store Tree
- Plots > Data Store Similarity > PCA Plot
- Plots > Data Store Similarity > tSNE Plot

Do these results suggest that we are likely to find a large or small number of places which are consistently different between the two samples?

Exercise 4: Differential Enrichment Analysis

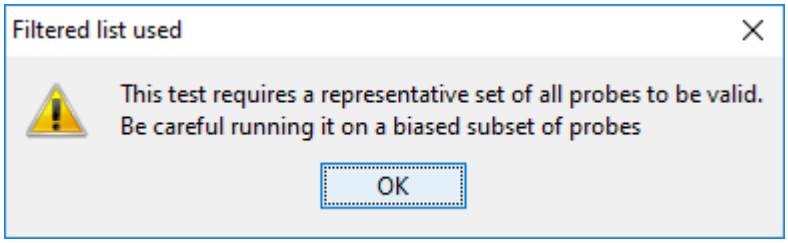
Step 4.1 LIMMA statistical testing

Since we are no longer working with raw counts but have applied a more complex normalisation to the data we don't have the option of using count based statistics such as DESeq or EdgeR. If our samples had been better matched we could have done a raw, uncorrected count and then used these tools.

As it is since we are effectively working with continuous data we can use the LIMMA tool to find differentially enriched regions.

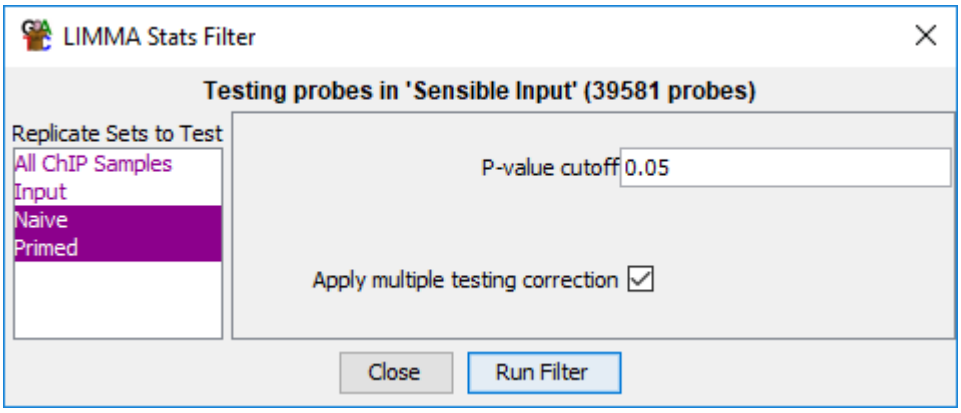
To do this select:

Filtering > Filter by Statistical Test > Continuous value statistics > Replicated Data > LIMMA. If you are using your filtered set of peaks (which you should) then you will get a warning saying:



This would be a problem if the filtering we'd done was related to the difference between the samples, but in our case the filtering is on an independent factor so we can ignore the warning this time around.

To run the filter you simply need to select the two replicate sets (Naïve and Primed) and then press "Run Filter".



Save the list of hits you get.

Exercise 5: Validating and exporting peak locations

Step 5.1

The first thing to check is that the peaks which were selected by the statistics make sense. We can do a number of sanity checks on them.

Firstly we can highlight the LIMMA hits on a scatterplot of Naïve vs Primed. We should see that the hits fall on the outside of the cloud of points. You can double click on some of the more extreme points on the outsides of the distribution to see what the differences look like in terms of the raw data. Do you see a similar size of effect for peaks which increased and decreased? If not, why might this be?

To get a different view of the comparison you can try log transforming the normalised data. You can do this with **Data > Quantitate existing probes > Log transform normalisation** (take the default values) then re-plot your scatterplot and see whether the linear or log view looks more intuitive to you.

To give you some context for the LIMMA hits you got, in the example plots at the end there is a graph which shows the hits generated by LIMMA, EdgeR and DESeq. You can see that there is a large overlap between the different tools, but that LIMMA gives the most hits, probably due to being able to use a more appropriate normalisation.

Step 5.2 Checking strand and position bias

Since the selection of fragments in ChIP seq usually occurs on double stranded DNA there should be no reason for there to be regions of the genome where either forward or reverse reads dominate. If a peak region seems to be heavily strand biased then it is usually a sign that there is a technical (probably mapping) problem in that region.

You can check the strand bias in your sensible list of peaks by running **Plots > Strand Bias Plot**. You can plot the bias across all ChIP Samples (feel free to have a look at the bias in the naïve and primed separately if you like). You should see that the accuracy of strand independence is a function of the number of reads and gets more stable as the read count increases. You should also see that the vast majority of peaks have sensible strand bias. You will also see a small number of regions where the bias seems usually high. Have a look at some of these and see if the data there looks odd at all.

You can also highlight the LIMMA hits and see if they have a tendency to be more strand biased that all peaks.

We can also look at the relative enrichment across the peaks in the different samples. To do this we can use an aligned probes plot under **Plots > Aligned Probes Plot**. We will do this only on the LIMMA hits, so select this list before drawing the plot. We can order the plot by the first Naïve ChIP sample.

Have a look at the results. Do you see clear enrichment around the centre of each peak. Does the strength of enrichment match between the different samples and do you understand any differences?

Step 5.3 Exporting Peaks

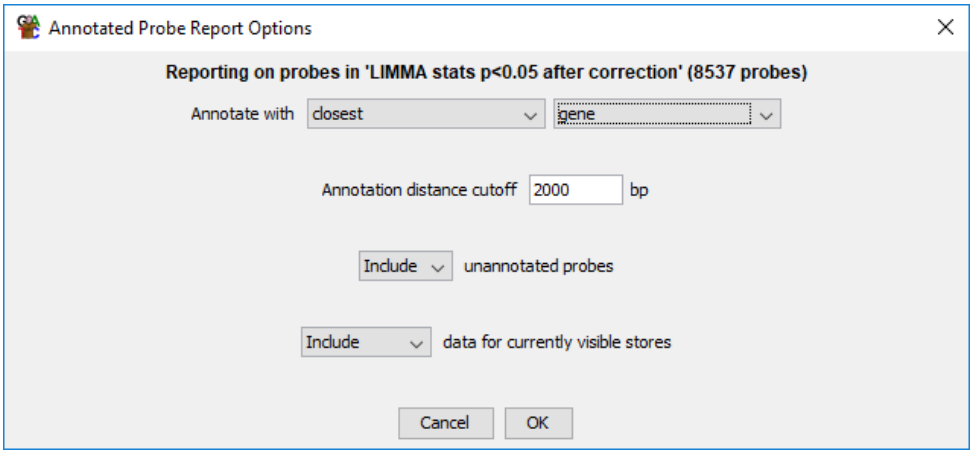
Finally, we are going to export a table of the hits we selected which we could take off for further analysis. We're going to annotate them with the name of the closest gene, but you should be very careful in using such gene lists in any gene set analysis since our preliminary evaluation showed there wasn't a strong linkage between the position of peaks and genes. We could however look for

conserved sequence motifs in the selected regions to see if a sequence based rationale for the selection of the regions could be found.

To generate a report of the hit locations first select the LIMMA hit list in the Data view. Then select:

Reports > Annotated Probe Report.

We are going to annotate with the closest gene to each hit (up to 2kb away).

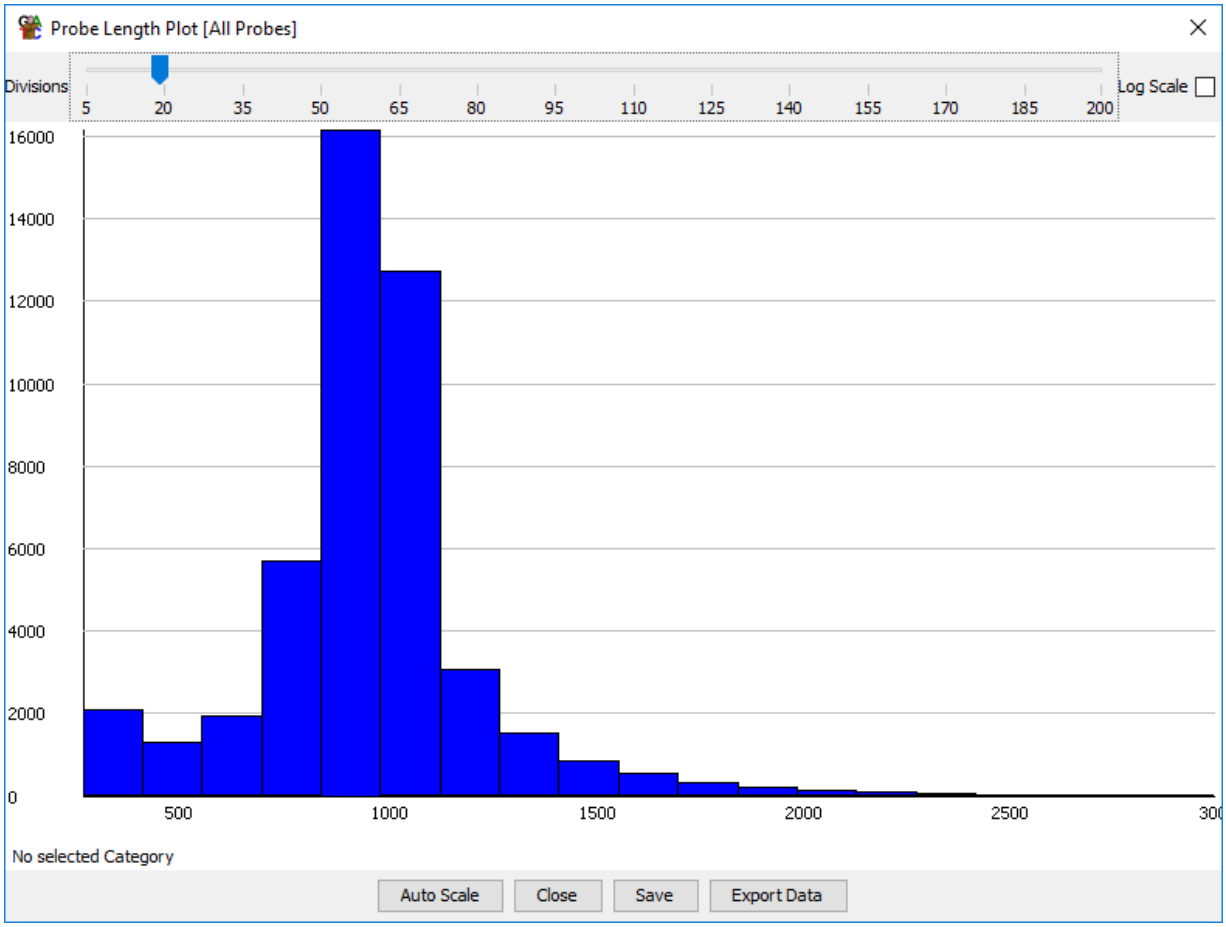


As well as being able to save the final list of interesting peaks to a text file, you can also use this display to order the peaks by their FDR value, so you can look at the most significant peaks.

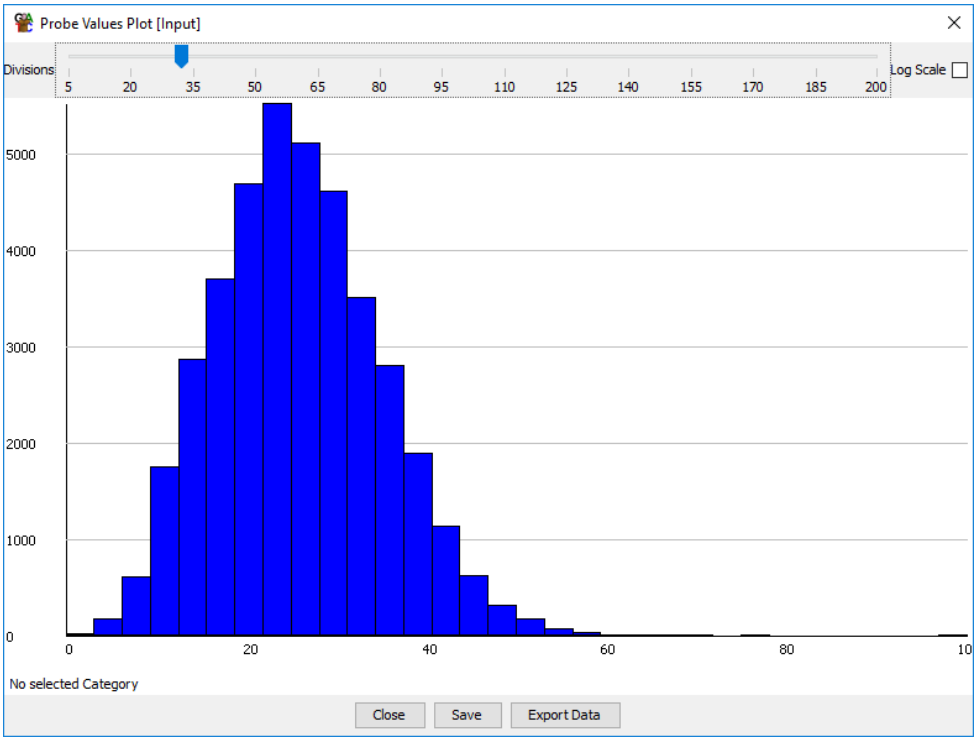
Example Plots

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

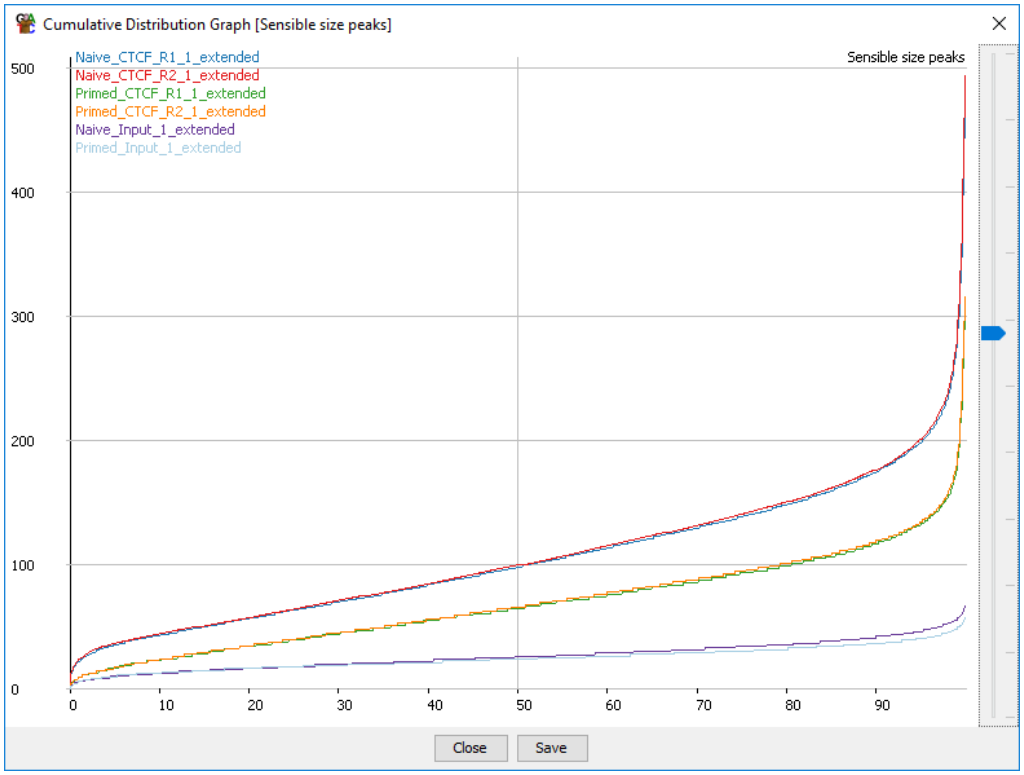
Step 3.1 Peak length distribution



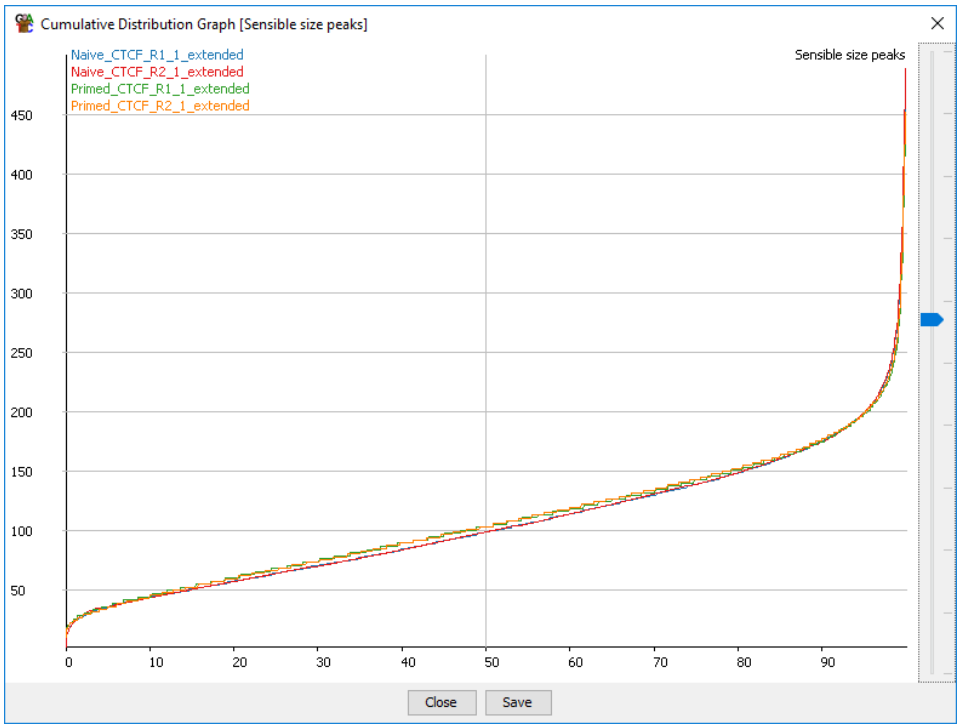
Step 3.1 Input signal distribution



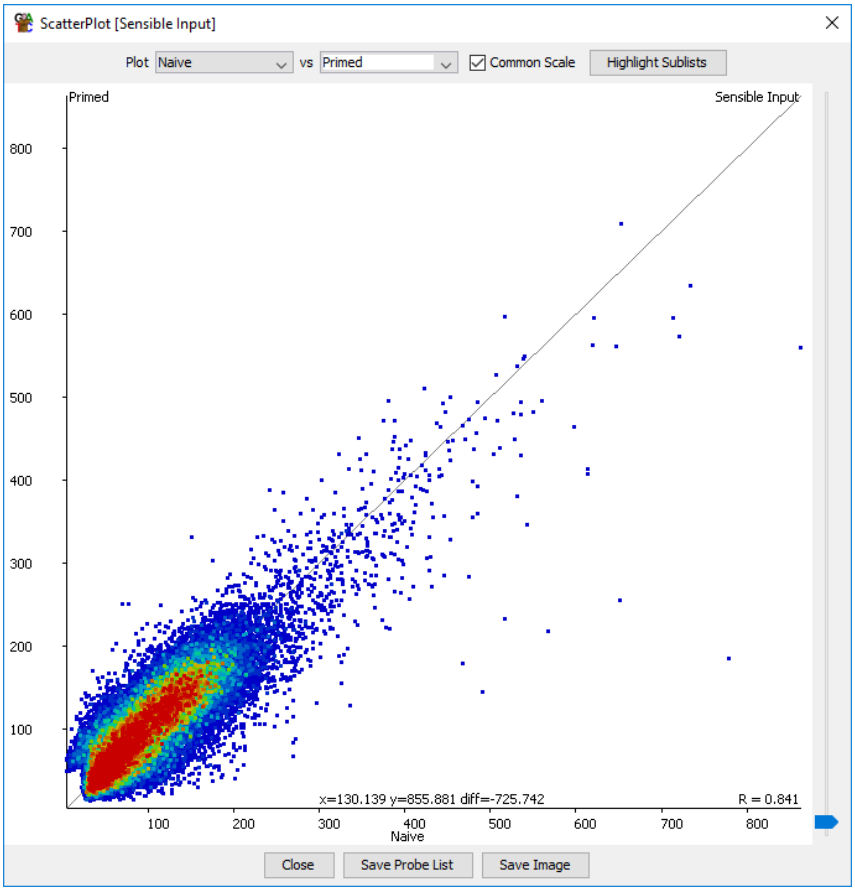
Step 3.2 Cumulative distribution plot before normalisation



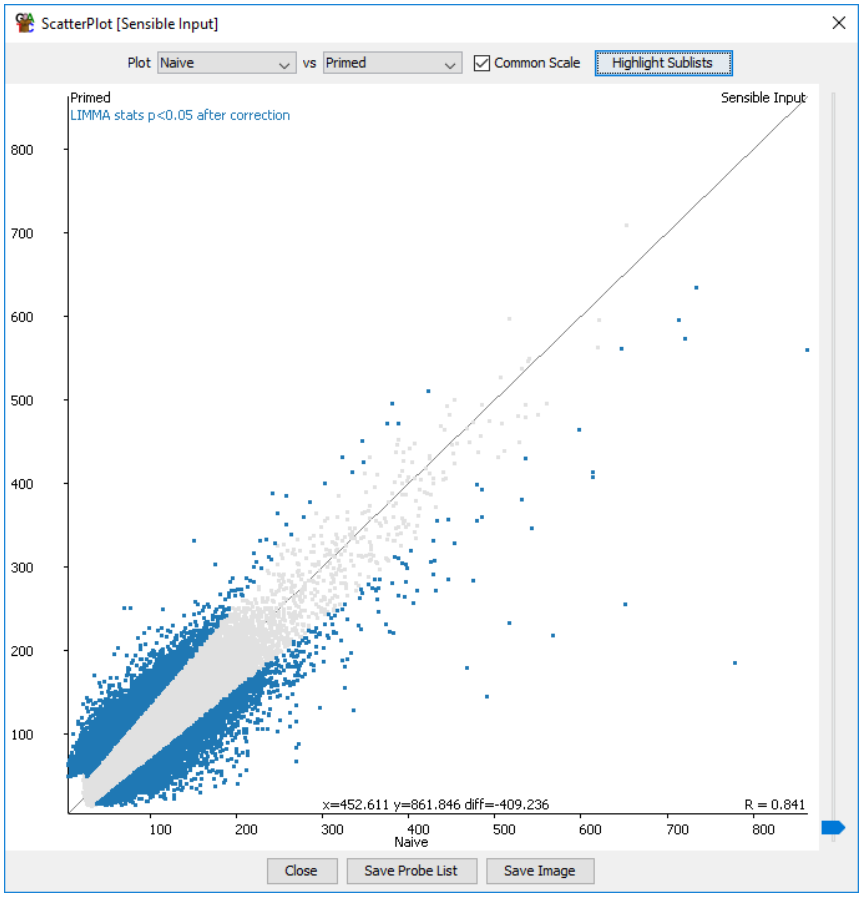
Step 3.2 Cumulative distribution plot after enrichment normalisation



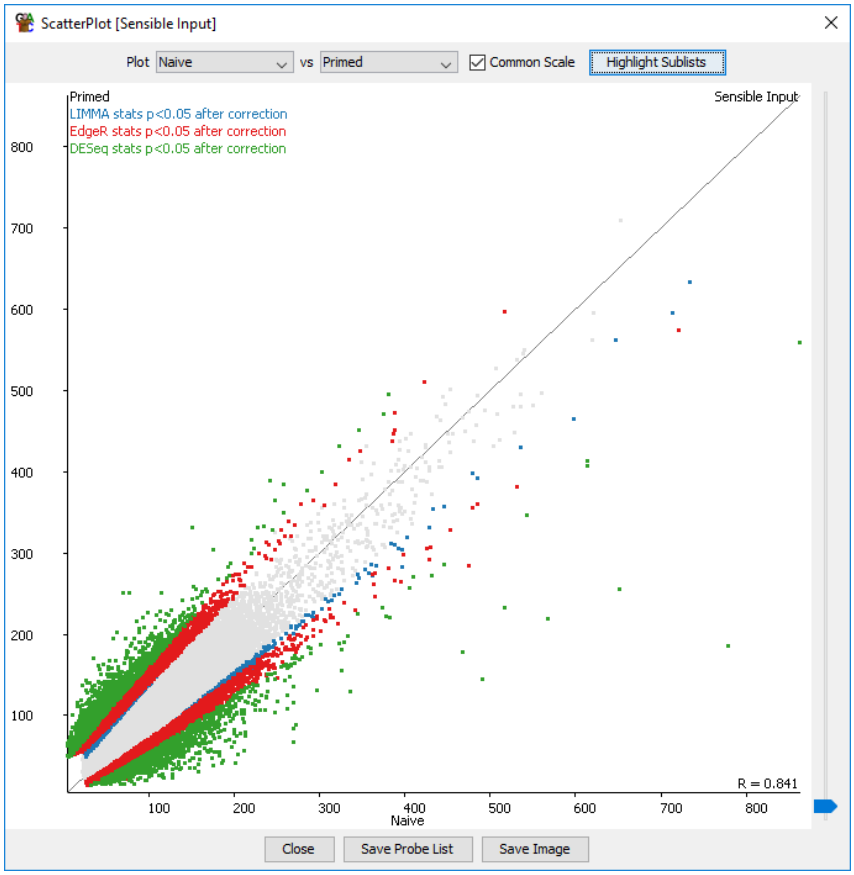
Step 3.2 Scatterplot after enrichment normalisation



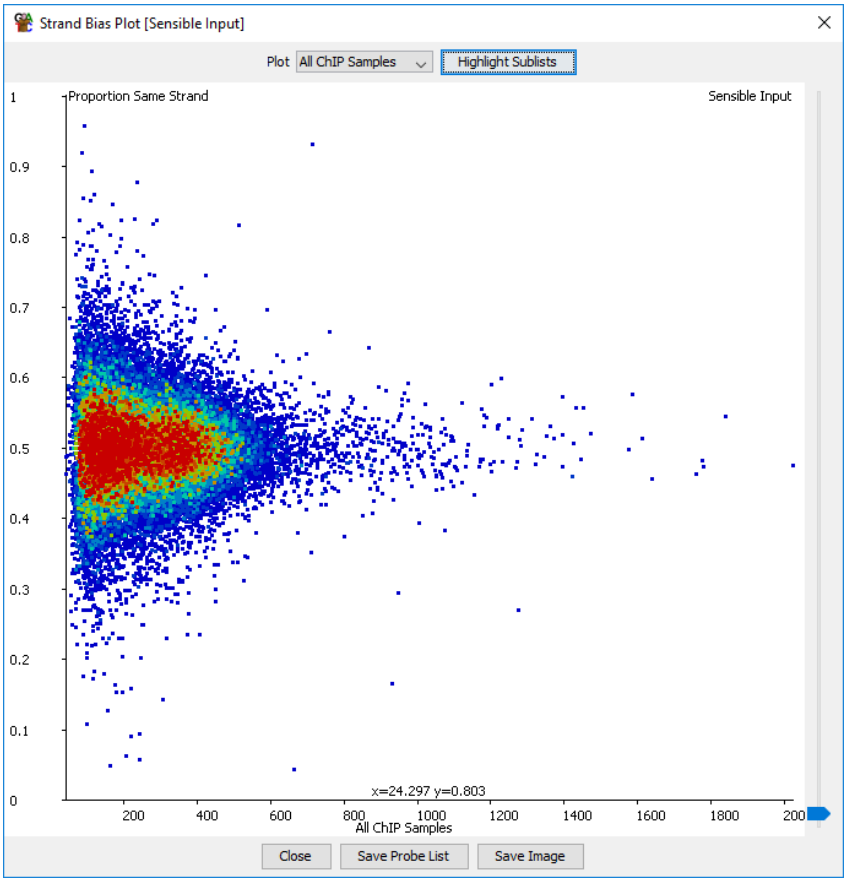
Step 5.1 LIMMA hits mapped onto a scatterplot

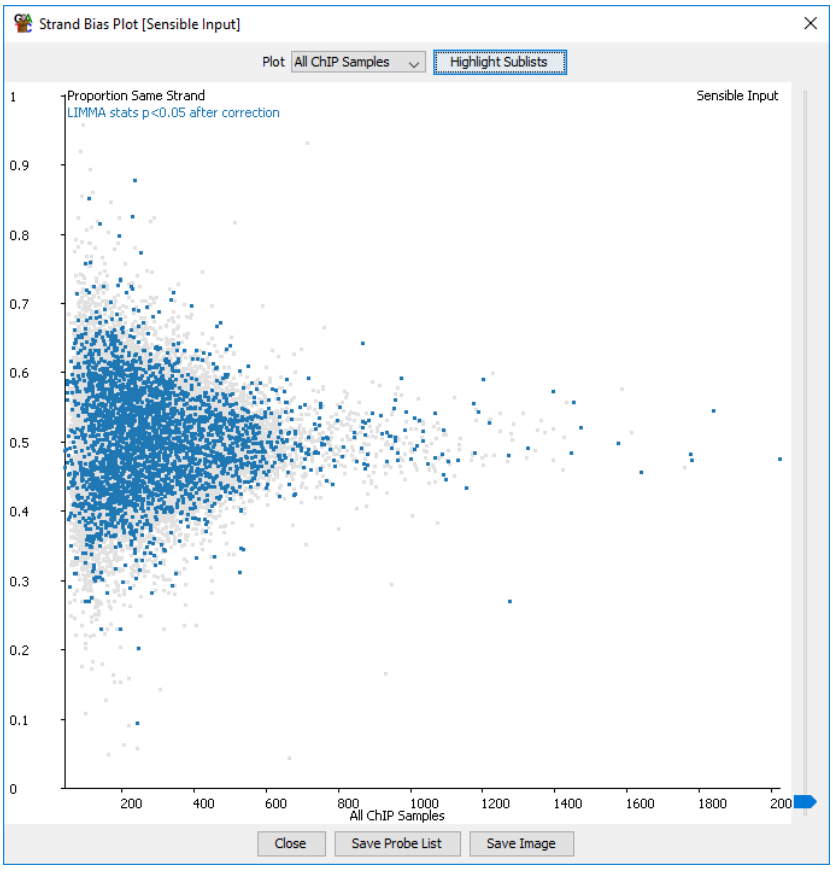


Step 5.1 comparison of the hits you get from LIMMA, EdgeR and DESeq

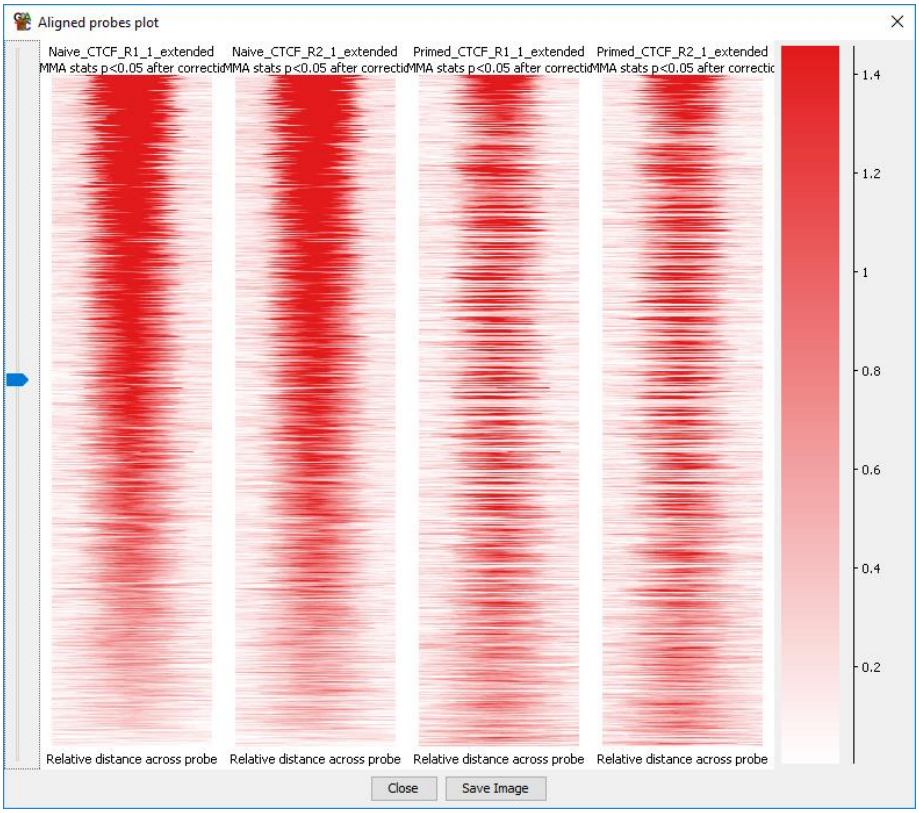


Step 5.2 Strand Bias Plot





Step 5.2 Aligned Probes Plot



Step 5.3 Exported table of hits

Annotated Probe Report

Probe	Chr...	Start	End	...	FDR	Feature	ID	Description	...	Type	Orientation	Dist...	Naive	Primed
Chr3:1949...	3	194928901	194929950		0						Not found	0	778.99	185.713
Chr5:3329...	5	33297451	33298650		0	AC010343.3	ENSG...	No description	-	gene	overlapping	0	490.758	145.099
Chr6:3327...	6	33279601	33280800		0	B3GALT4	ENSG...	beta-1,3-galactosyltransferase 4 [Source:HG...	+	gene	overlapping	0	651.497	256.064
Chr14:106...	14	106882801	106883850		0						Not found	0	270.063	67.432
Chr2:1139...	2	113979151	113980350		0	LINC01191	ENSG...	long intergenic non-protein coding RNA 1191 ...	+	gene	overlapping	0	568.108	218.384
Chr19:586...	19	58606051	58606950		0						Not found	0	202.237	40.972
Chr4:9616...	4	96169801	96170850		0						Not found	0	175.806	37.829
Chr4:1890...	4	189096601	189097500		0						Not found	0	186.541	44.801
Chr10:100...	10	100569451	100570350		0						Not found	0	35.939	170.397
Chr11:135...	11	135075301	135076500		0	AP005135.1	ENSG...	No description	-	gene	overlapping	0	467.116	178.836
Chr17:763...	17	76370101	76371150		0	PRPSAP1	ENSG...	phosphoribosyl pyrophosphate synthetase as...	-	gene	overlapping	0	270.063	84.253
Chr22:502...	22	50275801	50276850		0	PLXNB2	ENSG...	plexin B2 [Source:HGNC Symbol;Acc:HGNC:9...	-	gene	overlapping	0	517.333	233.262
ChrX:1317...	X	131767801	131769000		0	FIRRE	ENSG...	firre intergenic repeating RNA element [Sourc...	-	gene	overlapping	0	205.25	51.469
Chr2:4447...	2	44479051	44480250		0	CAMKMT	ENSG...	calmodulin-lysine N-methyltransferase [Sourc...	+	gene	overlapping	0	861.846	559.529
Chr13:114...	13	114352351	114353550		0						Not found	0	211.947	56.764
ChrX:4057...	X	40573501	40574400		0						Not found	0	272.863	89.244
Chr20:303...	20	30376651	30377550		0	FRG1BP	ENSG...	FSHD region gene 1 family member B, pseudo...	+	gene	overlapping	0	158.545	30.342
Chr7:3994...	7	3994951	3995850		0	SDK1	ENSG...	sidekick cell adhesion molecule 1 [Source:HGN...	+	gene	overlapping	0	70.773	251.34
Chr3:4693...	3	46935301	46936200		0	CCDC12	ENSG...	coiled-coil domain containing 12 [Source:HGN...	-	gene	overlapping	0	26.266	143.803
Chr8:2292...	8	229201	230250		0	RPL23AP53	ENSG...	ribosomal protein L23a pseudogene 53 [Sourc...	-	gene	overlapping	0	183	50.783
Chr17:743...	17	74353801	74354700		0	KIF19	ENSG...	kinesin family member 19 [Source:HGNC Sym...	+	gene	overlapping	0	76.798	250.521

Close Save