Babraham Bioinformatics

Exercises:
Analysing ChIP-Seq data

Version 2021-04

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

If you have data open in seqmonk already then close and re-open the program. Load in the seqmonk project file from the Human CTCF data folder.

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

In this study you can hopefully see that peak calling, specifically narrow peak calling, would be appropriate. We are going to look at two aspects of peak calling. We will start by doing a manual peak calling for one of the samples using the MACS peak caller built into seqmonk.

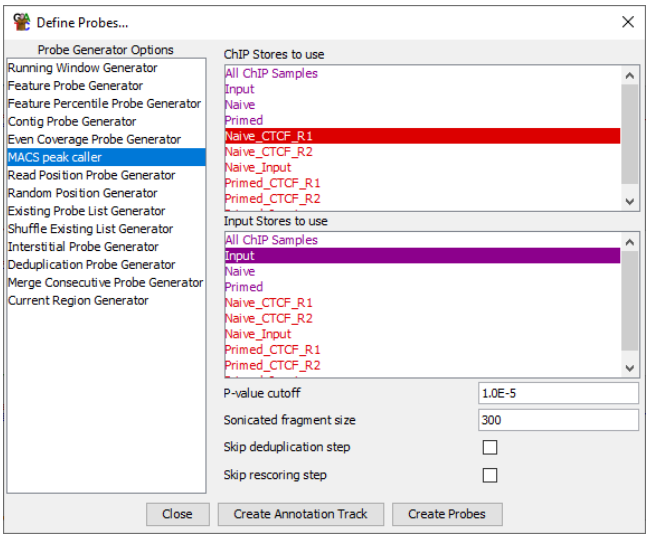
For a final analysis we will import a set of MACS2 peaks called on the command line from the underlying BAM files and will import these to create a merged set of peaks which we can then quantitate and analyse.

Step 2.1

We will use the MACS peak caller probe generator within seqmonk to call peaks for one of the samples so you can see how the process works. We'll use the Naïve CTCF R1 sample.

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

In the options, the ChIP store will be the Naïve CTCF R1 sample 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.

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

Step 2.2 Importing MACS2 Peaks

We now want to make a final set of positions which we can analyse to compare enrichment between naïve and primed. You have been provided with a set of peak call files from MACS2 which were run on the underlying BAM files for this data.

`naive_1_peaks.narrowPeak.txt`

`naive_2_peaks.narrowPeak.txt`

`primed_1_peaks.narrowPeak.txt`

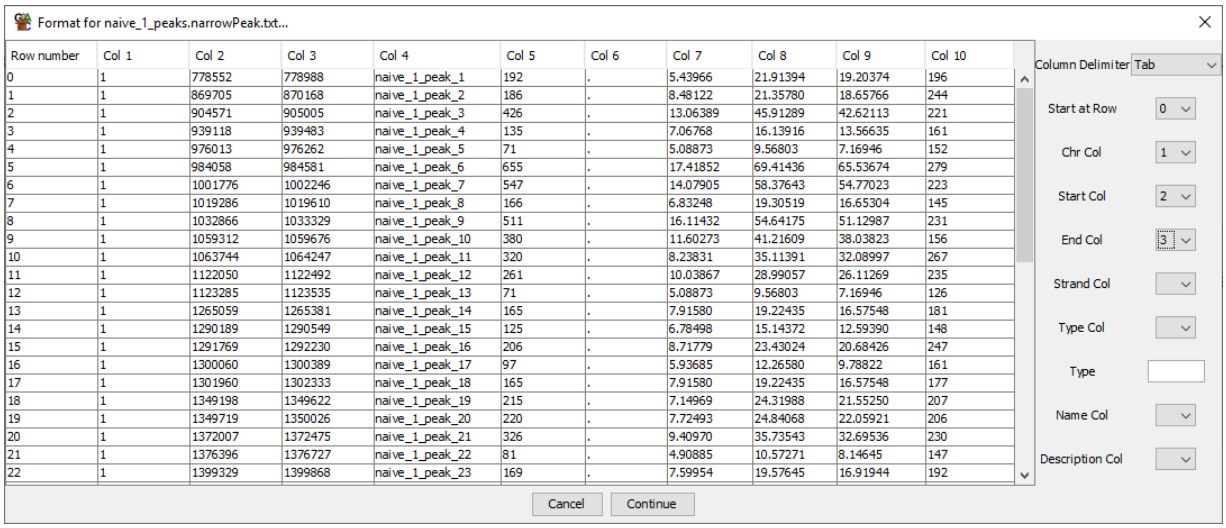
`primed_2_peaks.narrowPeak.txt`

We are going to import these peaks as annotation tracks and then create a merged set of peaks from them.

To import the peaks select:

File > Import Annotation > Text (Generic)

Then select all of the narrowPeak files from the MACS2_Peaks folder. You will then see a preview of the top of the first file and you'll need to say which columns are the chromosome, start and end. These are chromosome = col 1, start = col 2, end = col3



Row number	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10
0	1	778552	778988	naive_1_peak_1	192	.	5.43966	21.91394	19.20374	196
1	1	869705	870168	naive_1_peak_2	186	.	8.48122	21.35780	18.65766	244
2	1	904571	905005	naive_1_peak_3	426	.	13.06389	45.91289	42.62113	221
3	1	939118	939483	naive_1_peak_4	135	.	7.06768	16.13916	13.56635	161
4	1	976013	976262	naive_1_peak_5	71	.	5.08873	9.56803	7.16946	152
5	1	984058	984581	naive_1_peak_6	655	.	17.41852	69.41436	65.53674	279
6	1	1001776	1002246	naive_1_peak_7	547	.	14.07905	58.37643	54.77023	223
7	1	1019286	1019610	naive_1_peak_8	166	.	6.83248	19.30519	16.65304	145
8	1	1032866	1033329	naive_1_peak_9	511	.	16.11432	54.64175	51.12987	231
9	1	1059312	1059676	naive_1_peak_10	380	.	11.60273	41.21609	38.03823	156
10	1	1063744	1064247	naive_1_peak_11	320	.	8.23831	35.11391	32.08997	267
11	1	1122050	1122492	naive_1_peak_12	261	.	10.03867	28.99057	26.11269	235
12	1	1123285	1123535	naive_1_peak_13	71	.	5.08873	9.56803	7.16946	126
13	1	1265059	1265381	naive_1_peak_14	165	.	7.91580	19.22435	16.57548	181
14	1	1290189	1290549	naive_1_peak_15	125	.	6.78498	15.14372	12.59390	148
15	1	1291769	1292230	naive_1_peak_16	206	.	8.71779	23.43024	20.68426	247
16	1	1300060	1300389	naive_1_peak_17	97	.	5.93685	12.26580	9.78822	161
17	1	1301960	1302333	naive_1_peak_18	165	.	7.91580	19.22435	16.57548	177
18	1	1349198	1349622	naive_1_peak_19	215	.	7.14969	24.31988	21.55250	207
19	1	1349719	1350026	naive_1_peak_20	220	.	7.72493	24.84068	22.05921	206
20	1	1372007	1372475	naive_1_peak_21	326	.	9.40970	35.73543	32.69536	230
21	1	1376396	1376727	naive_1_peak_22	81	.	4.90885	10.57271	8.14645	147
22	1	1399329	1399868	naive_1_peak_23	169	.	7.59954	19.57645	16.91944	192

You can then press “Continue” until all of the peak sets are imported. Since you’re importing multiple annotation tracks they won’t automatically be added to your view. To make them appear you need to select.

View > Set Annotation Tracks

Then select the 4 peak tracks from the Available Tracks and add them to the displayed tracks. You should see them show up in the chromosome view.

Have a look at the sets of calls. Do you see the same positions coming up between the replicates for each condition? Do you see differences between naïve and primed?

Do you always see what you’d think was a biologically relevant enrichment increase where you have a call?

Do you see obvious differences in the enrichment between samples where you have differences in the presence / absence of a call in different samples?

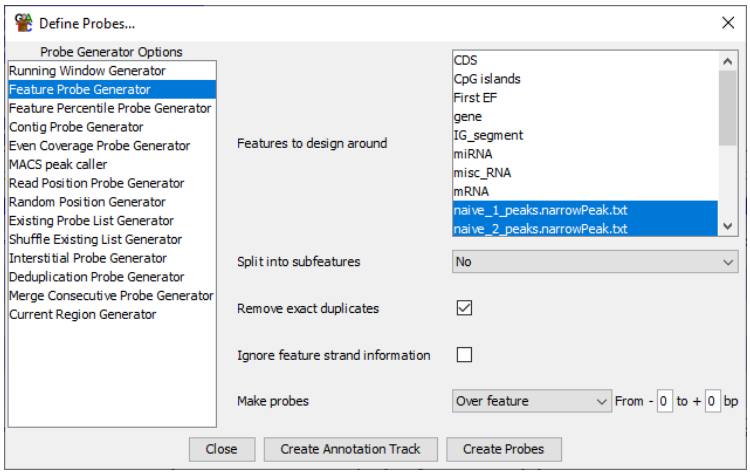
Step 2.3 Creating a merged peak set

We now want to make a merged set of peaks from all of the predictions. We’ll need to do this in two steps.

Firstly we will use the feature probe generator to put probes over all of the peaks. To do this select:

Data > Define Probes > Feature Probe Generator

Select all 4 peak tracks in the “Features to design around” list and press create probes.

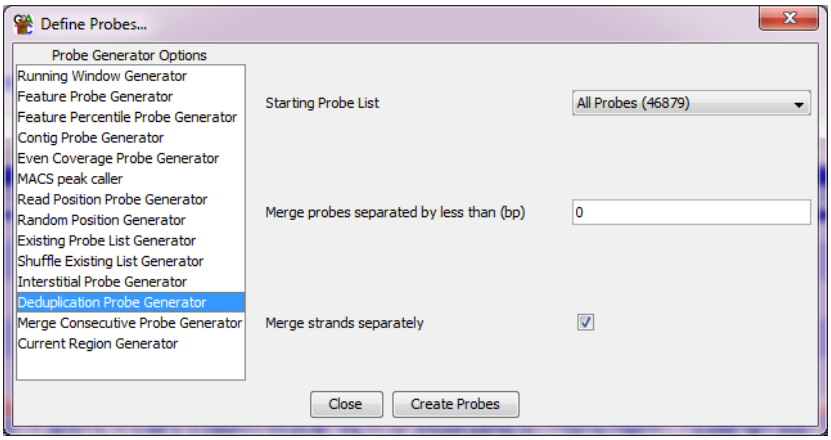


Use the “Fixed value quantitation” to give a value of 1 to each probe (these probes aren’t going to be around for long so it really doesn’t matter what quantitation they get!).

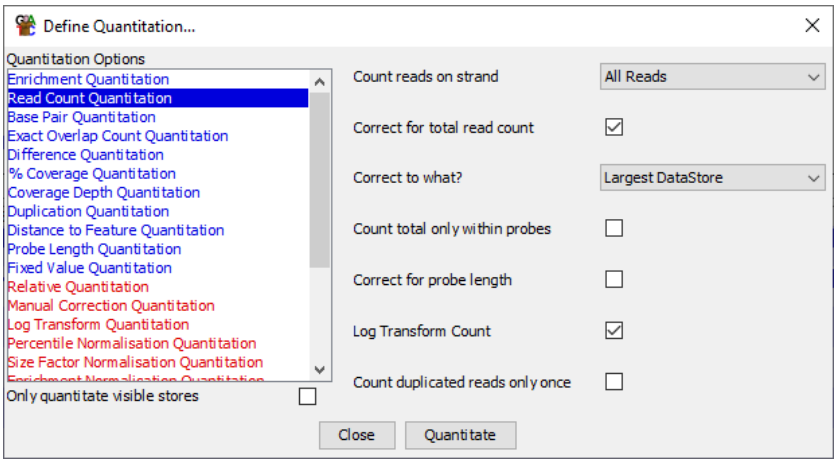
Next we need to merge overlapping peaks together. To do this select

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.



When the quantitation options appear again you should quantitate with read count quantitation on a log scale.



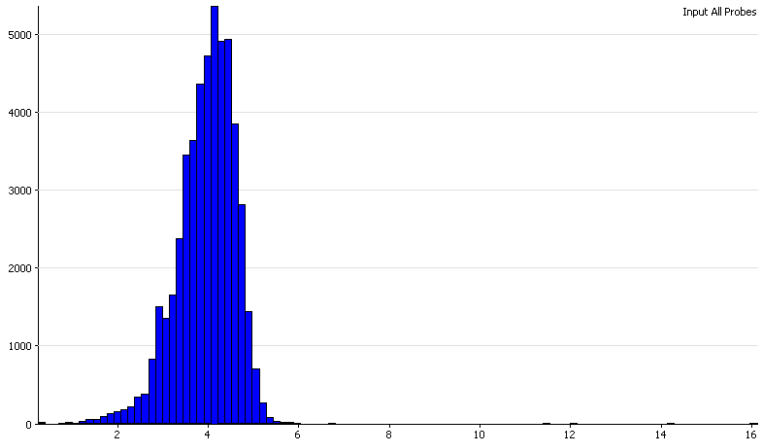
Look carefully through your final set of peaks. Compare it to the individual peak tracks you have for each sample, and the data you can see and check that it looks like you have captured all of the potentially interesting places in the genome.

You can also have a look at a scatterplot of the 3 groups of samples against each other – comparing the inputs to each of the ChIPs and the two ChIPs against each other.

Exercise 3: Filtering, Quantitation and Normalisation

Step 3.1

Before we go any further we'll do one little bit of cleaning. If you find the Input Replicate set in the Data View and right click on it and select "Probe Value Histogram" you should see something like this:

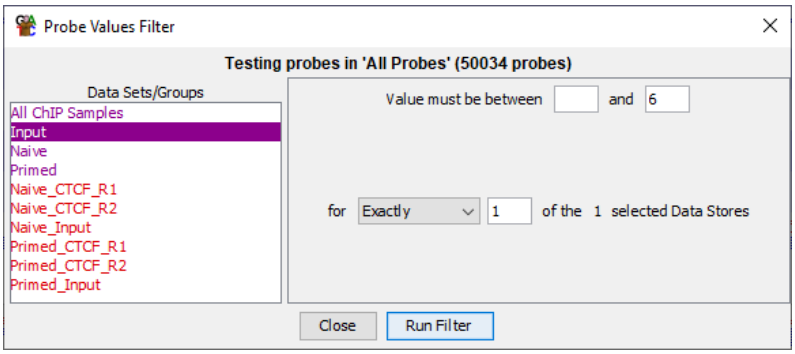


What this says is that most input regions (as expected) have relatively low and consistent values. However there are a small number of input regions with very (and in some case stupidly) high read counts. Most of these come from the mitochondrion, but there are others. Since they will skew our plotting it makes sense to get rid of them. We wouldn't want to pursue hits with odd input read levels anyway.

To get rid of these select:

Filtering > Filter on Values > Individual Probes

You can then select only the probes with input values below 6

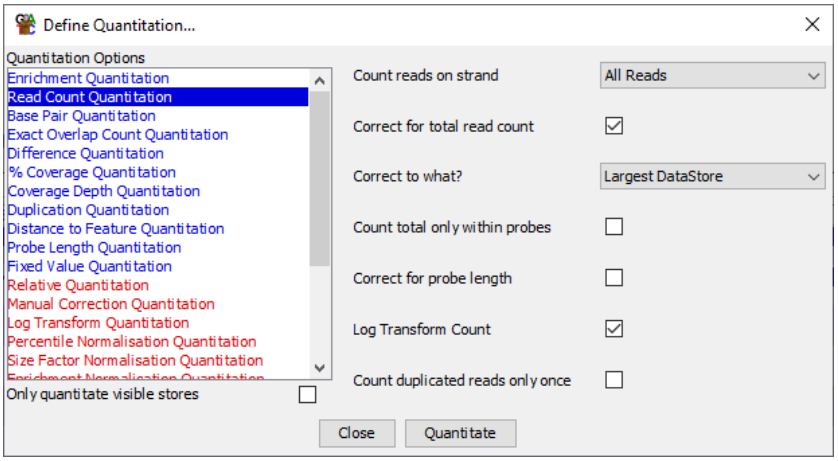


Save this list as "Sensible Input" and then select it to show only those probes.

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 log 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.

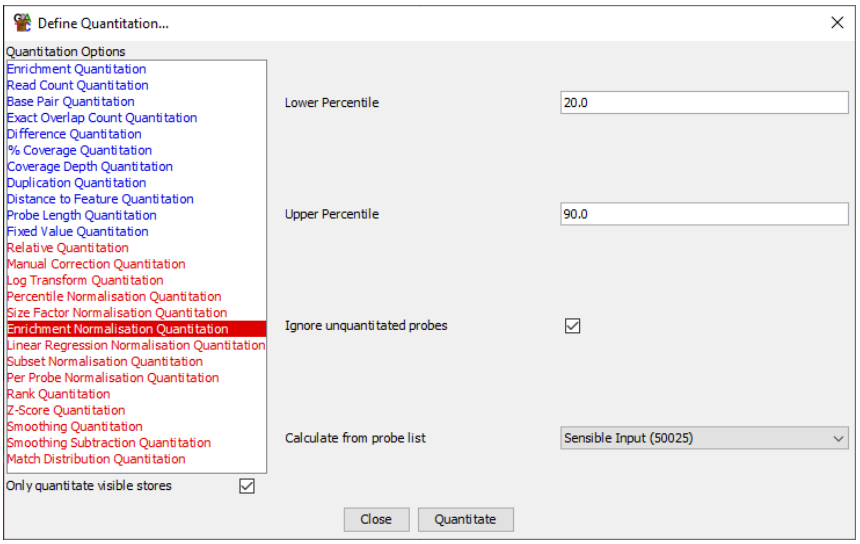
To do this we are going to use the enrichment normalisation quantitation. This sets two points of reference – a lower one for a background level, and an upper one for a highly enriched region. The tool then scales all samples between these two points.

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 > Set Data Tracks** to remove them from the current view.

We can then normalise the remaining 4 ChIP samples. Select

Data > Quantitate Existing Probes > Enrichment Normalisation Quantitation.

We are going to normalise between the 20th and 90th 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 Viewing the normalised data

Now that we have values which are more directly comparable you can go back to re-draw the scatterplot of the native vs primed. Have a look and see if the data appears to be well normalised now. See if you think there will be large or small number of hits.

Have a look at some of the largest changes and get a feel for what these look like in the real data.

Exercise 4: Differential Enrichment Analysis

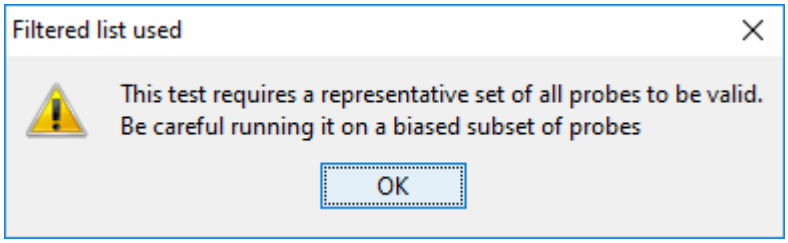
Step 4.1 LIMMA statistical testing

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.

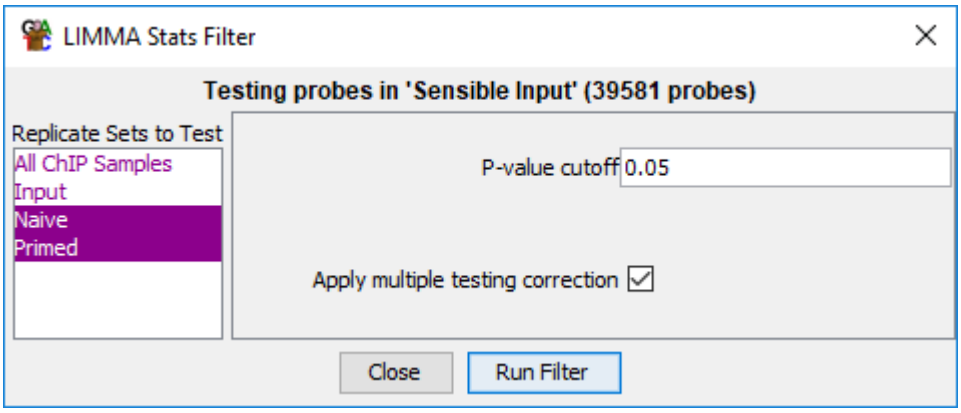
To do this select:

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?

You will see that you get a large number of hits from the analysis. If you wanted to focus on the most promising ones you could do additional filtering:

- FDR filtering to select peaks with a more stringent cutoff
- Value filtering to remove peaks with lower levels of observation
- Fold change filtering to remove peaks with lower levels of absolute enrichment
- Intensity difference filtering to select peaks with the highest enrichment relative to their observation level.

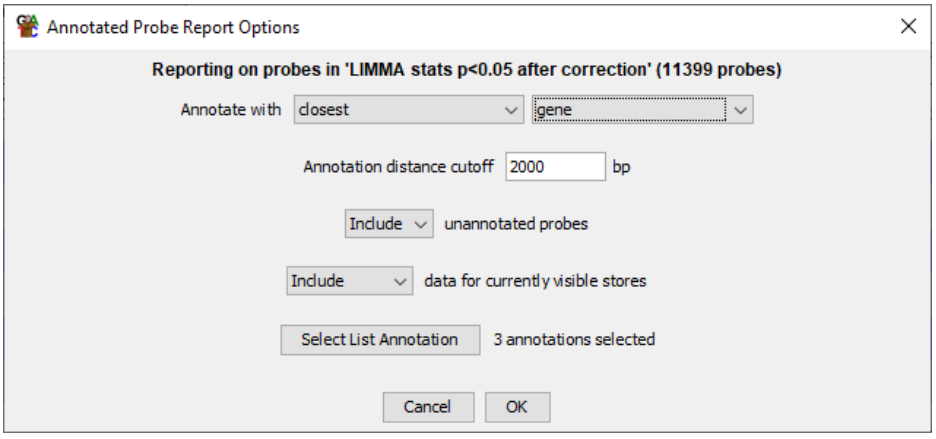
Step 5.2 Exporting Peaks

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).

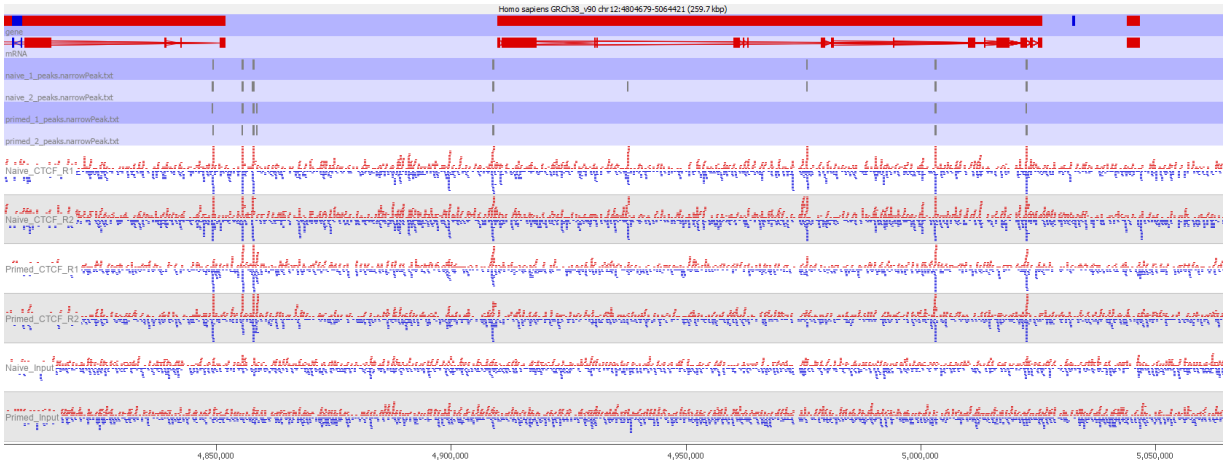


As well as being able to save the final list of interesting peaks to a text file, you can also use this display to order the peaks by their FDR value, so you can look at the most significant peaks.

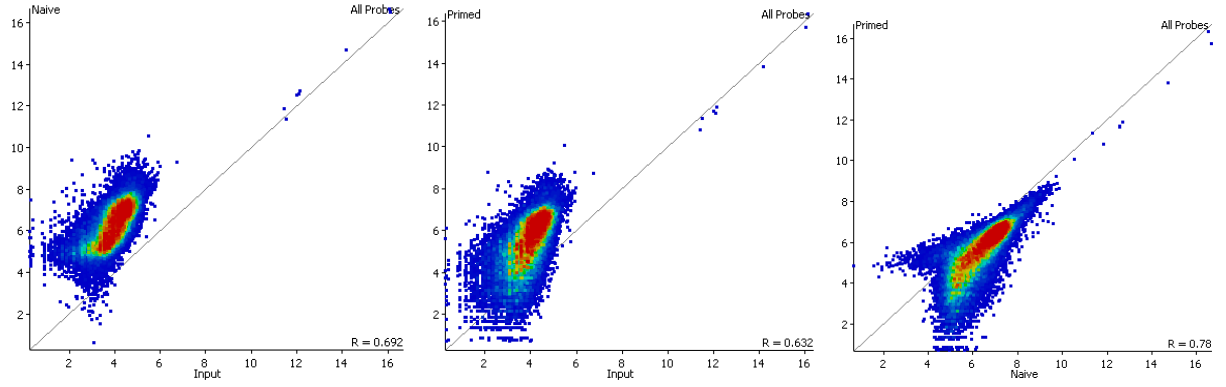
Example Plots

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

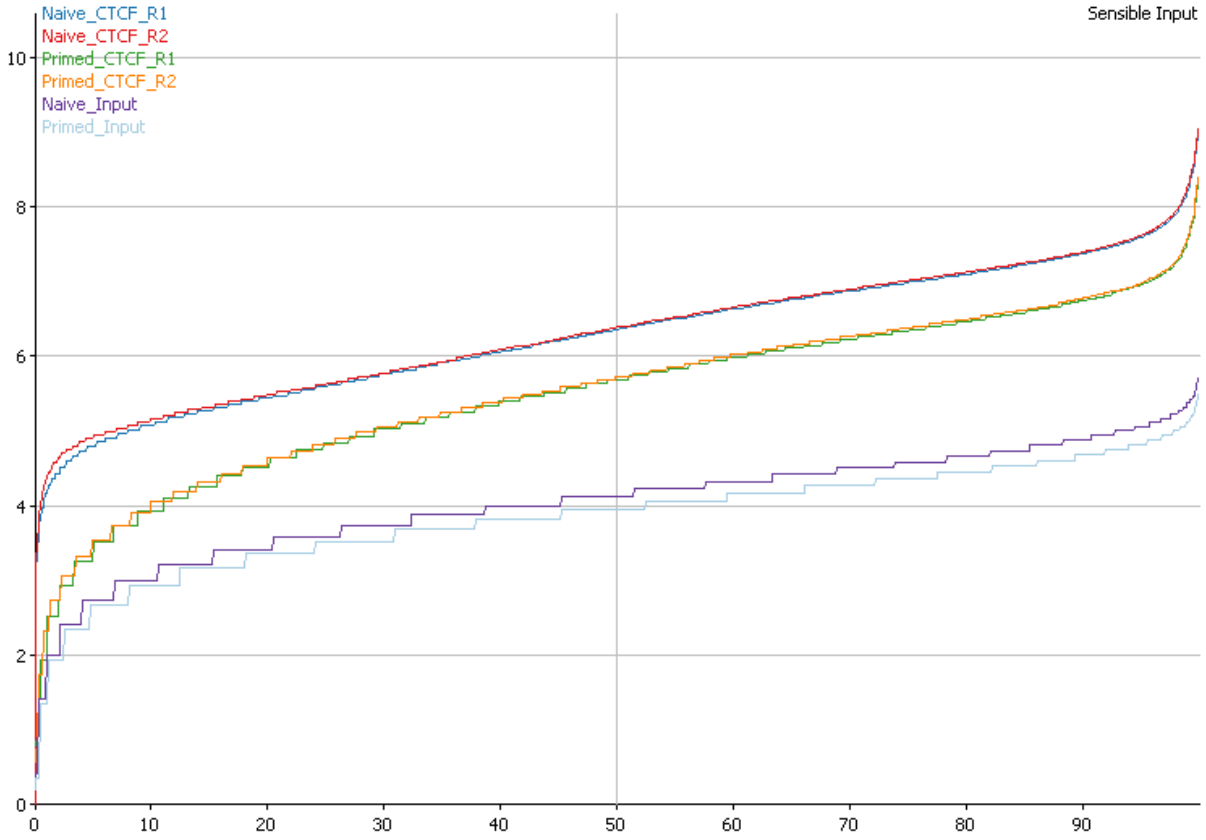
Step 2.3 Merged Peaks



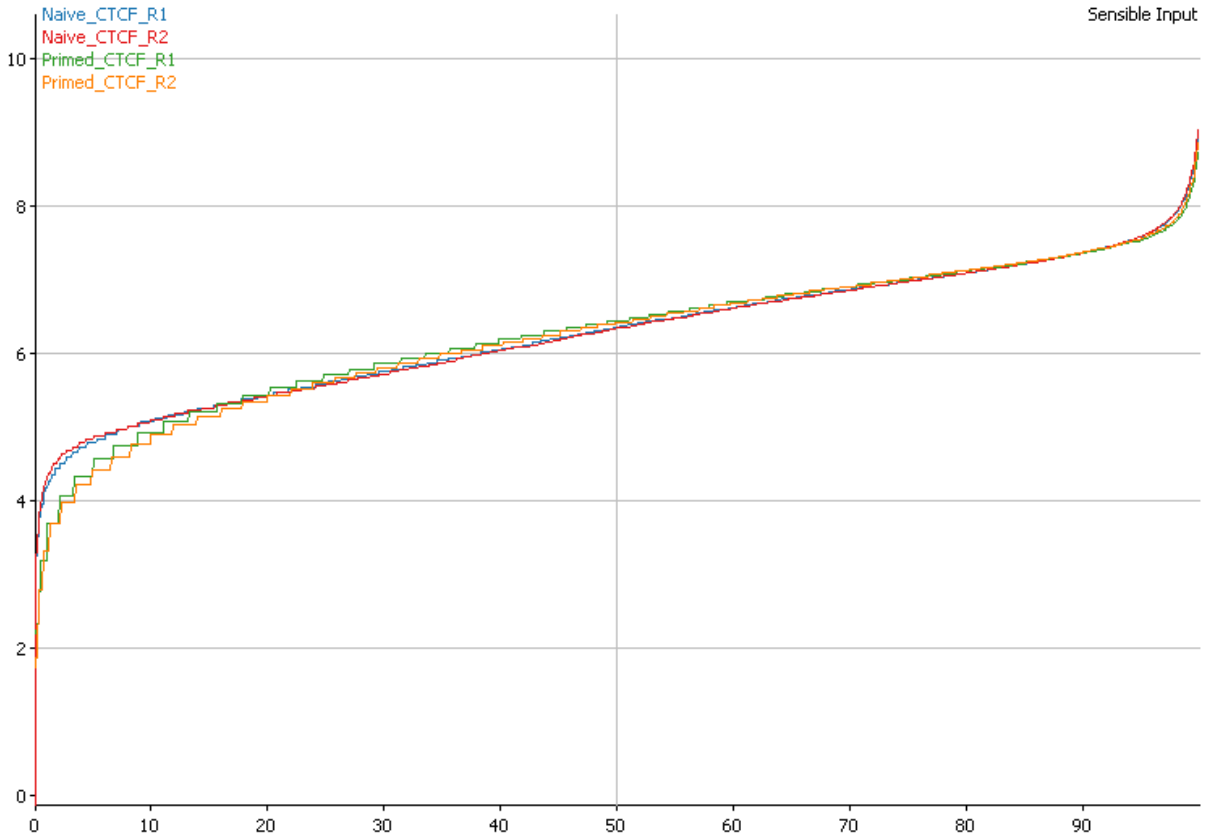
Step 2.3 Raw Quantitation Comparison



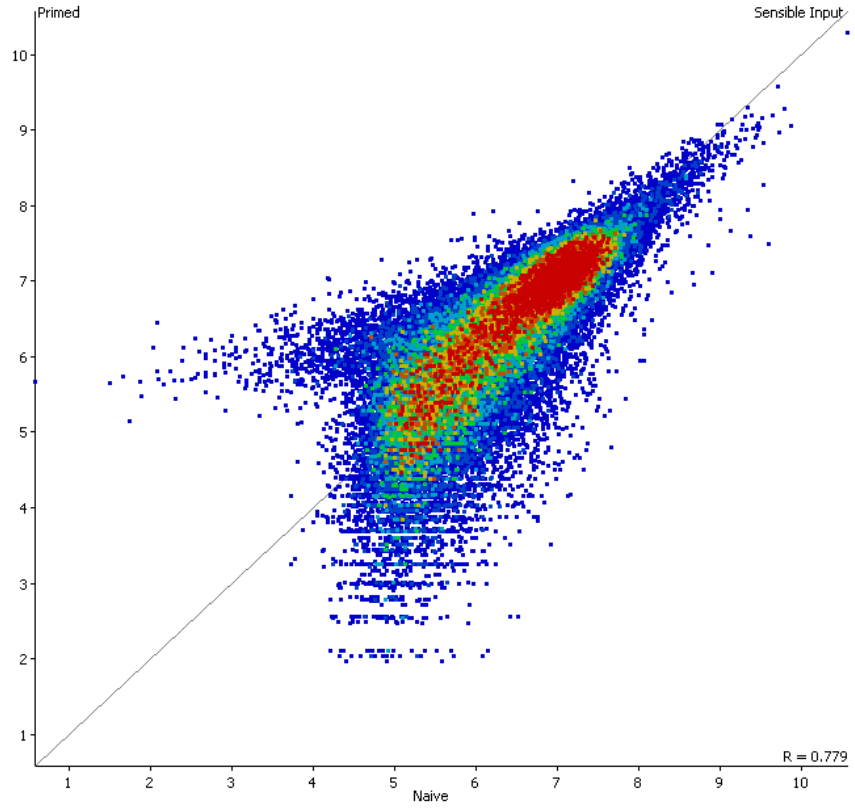
Step 3.2 Cumulative distribution plot before normalisation



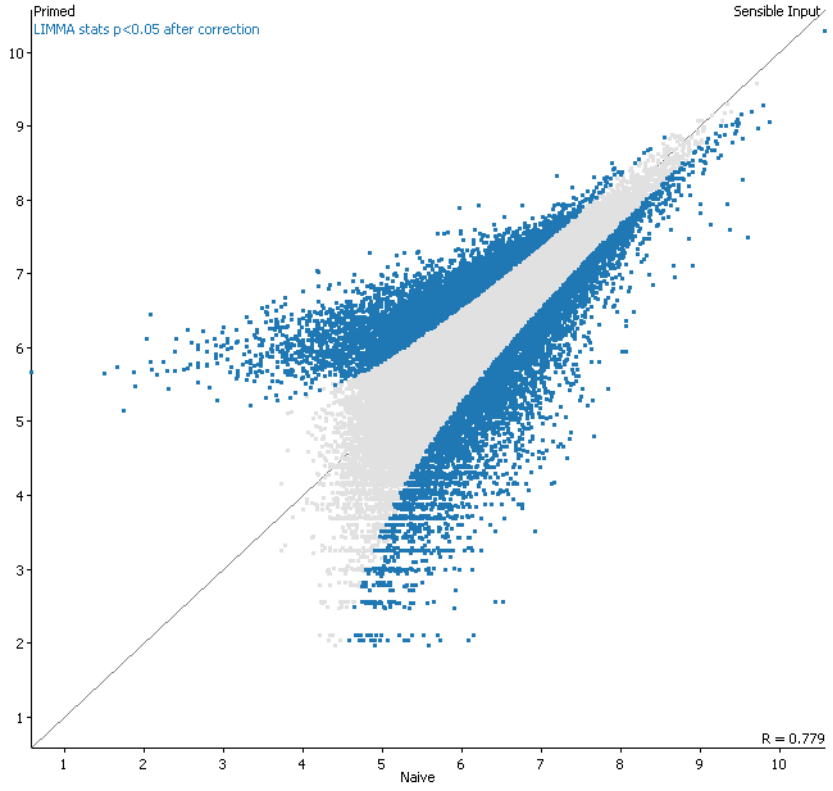
Step 3.2 Cumulative distribution plot after enrichment normalisation



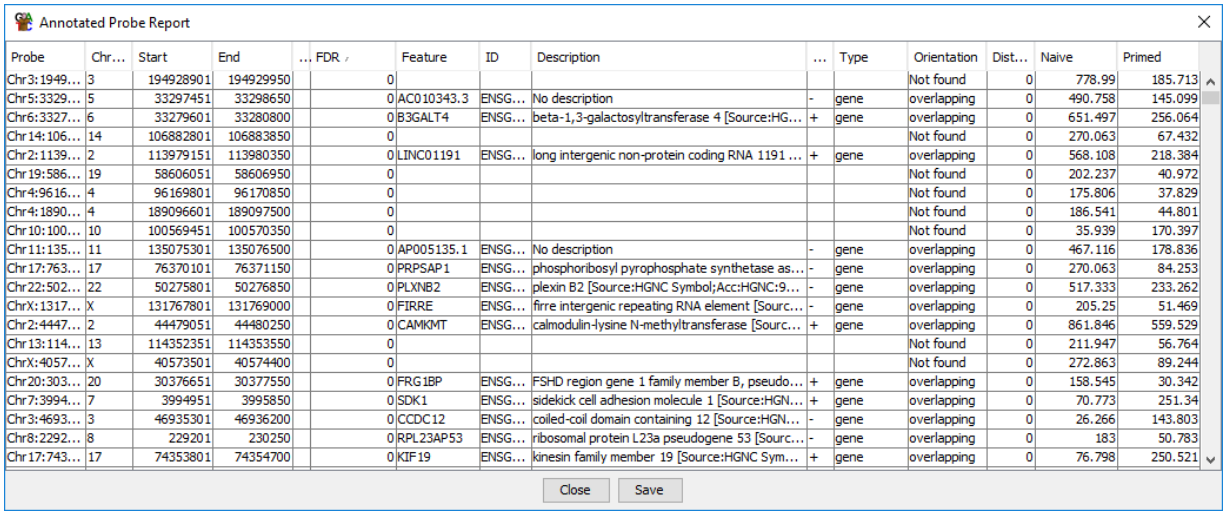
Step 3.2 Scatterplot after enrichment normalisation



Step 5.1 LIMMA hits mapped onto a scatterplot



Step 5.2 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