Introduction

One of the common downstream analyses of ChIP-seq data is comparing the binding profile of a given factor in different conditions (e.g. healty - disease, wild-type - mutant, different cell types). In this practical we will do differential binding analysis on a H3K27ac (active enhancer mark) dataset comparing two mutant Drosophila cell lines (Tl10b and gd7), which correspond to mesodermal and ectodermal precursor cells (GSE68983) [Koenecke et al., Genome Biology (2016)].

Differential binding

Finding differentially bound regions in the genome is analogous to identifying differentially expressed genes in RNA-seq data. In both cases we are dealing with count data summerised over features (genes/transcripts in the case of RNA-seq and peaks in the case of ChIP-seq). In both cases the biological replicates show larger variability than technical replicates, therefore the negative binomial model is the appropriate statistical test to compare binding affinities accross samples. There are a number of differential expression R packages that use the negative binomial model e.g. `DESeq2` and `EdgeR`. These methods are wrapped in the `DiffBind` package that is geared towards analysing differential binding in ChIP-seq data and provides a number of analytical plots as well. In this practical we will use the `DiffBind` R package to identify differentially acetylated regions in the two Drosophila cell lines. Reads were downloaded and converted into fastq format using the following command (not to be performed at the course):

```
for i in SRR2031888 SRR2031889 SRR2031896 SRR2031897 SRR2031909
   SRR2031910 SRR2031917 SRR2031918
do
fastq-dump -F --gzip $i
done
```

Let's move into the folder containing the data.

```
cd ~/Desktop/Differential_binding/
```

Quality control

In order to proceed with the dataset it is important to test the quality of the reads, for which we use `FASTQC` on the fastq files. We have already performed this step using the following command and you can check the results in the html files:

```
for i in SRR2031888 SRR2031889 SRR2031896 SRR2031897 SRR2031909
   SRR2031910 SRR2031917 SRR2031918
```

```
do
fastqc $i
done
```

Questions

1. How is the sequencing quality? ______________________________________
   ____

Alignment

These files are fine to proceed and map them to the Drosophila genome (dm3). We will map reads only to chr2L to keep the dataset small. For mapping we use the aligner called `bowtie` with the options -v 2, which allows for 2 mismatches in the alignment and -p 20, which specifies that we use 20 cores for mapping. At the same time we use samtools remove low quality alignments and to sort the bam file according to chromosome and position. The sorted bam files can be filtered for PCR duplicates using the samtools markdup function. These steps were also done for you before the course, you will use bowtie in the next practical yourself.

```
for i in *fastq.gz
do bowtie -v 2 -p 20 -S /mnt/large/data/bowtie_indexes/dm3_chr2L/dm3 $i
   | samtools view -Sb -q 10 | samtools sort > ${i::-9}.bam
done

for i in *.bam
do samtools rmdup -s $i ${i::-4}dedup.bam
done
```

Files were subsequently renamed to reflect celltype and sample or input information (gd7_K27ac_rep1, gd7_K27ac_rep2, gd7_input_rep1, gd7_input_rep2, tl10b_K27ac_rep1, tl10b_K27ac_rep2, tl10b_input_rep1, tl10b_input_rep2). The tl10b_input_rep1 has very low read counts therefore we use the escond replicate as input for both K27ac samples.

Peak calling

The first step in the differential binding analysis is identifying the regions to be tested. It could be regions that you are interested in (e.g. promoters, enhancers etc.), but in most cases we test differential binding at the peaks called in each dataset to be compared. In order to make sure that the peaks are not PCR artefacts, we use the deduplicated files for peak calling.

Open the Terminal. First, go to the right folder, where the data are stored.

```
cd ~/Desktop/Differential_binding/
```

Call peaks for each sample using MACS2. H3K27ac marks are relatively sharp, therefore we use the standard callpeaks function as you did yesterday, but as a bonus exercise you could use the --broad option that is specifically designed for broad histone modifications and joins smaller peaks into wider peaks. For effective genome size we use 90% of chr2L which 21Mb.

```
macs2 callpeak -t gd7/gd7_K27ac_rep1_chr2Ldedup.bam -c
   gd7/gd7_input1_chr2Ldedup.bam --format BAM --name
   dedup_peaks/gd7_H3K27ac_rep1 --gsize 21000000 --qvalue 0.01
   --call-summits
macs2 callpeak -t gd7/gd7_K27ac_rep2_chr2Ldedup.bam -c
   gd7/gd7_input2_chr2Ldedup.bam --format BAM --name
   dedup_peaks/gd7_H3K27ac_rep2 --gsize 21000000 --qvalue 0.01
   --call-summits
macs2 callpeak -t tl10b/tl10b_K27ac_rep2_chr2Ldedup.bam -c
   tl10b/tl10b_input_chr2Ldedup.bam --format BAM --name
   dedup_peaks/tl10b_H3K27ac_rep2 --gsize 21000000 --qvalue 0.01
   --call-summits
macs2 callpeak -t tl10b/tl10b_K27ac_rep1_chr2Ldedup.bam -c
   tl10b/tl10b_input_chr2Ldedup.bam --format BAM --name
   dedup_peaks/tl10b_H3K27ac_rep1 --gsize 21000000 --qvalue 0.01
   --call-summits
```

The next steps will be performed within Rstudio, please open it and set the working directory to ~/Desktop/Differential_binding/. For the DiffBind analysis we need to prepare a comma separated file containing information about the samples to compare. We prepared this file earlier and you can view it by loading it into R.

```
library(readr)
sheet=as.data.frame(read_csv("H3K27ac_diffbind2.csv"))
```

Questions

1. Which column can we use to contrast the two cell types? ____________________

   ____

Differential interactions within DiffBind

As the immunoprecipitation largely enriches a small fraction of the genome, it is expected that in the ChIP samples we see exact duplicates, and removing those can reduce your power to identify significant binding differences, as it caps the dynamic range. In order to see the real level of changes, we use the mapped files including duplicates. Note, that a higher than expected level of duplication indicates that there are extensive PCR amplification artefacts, and in this case it is advisable to remove the duplicates.

Questions

1. Would you include or exclude duplicate reads in this analysis given the FASTQC results? ____

   ____

We load the `Diffbind` library.

```
library(DiffBind)
```

We create a `DiffBind` object, by reading the pre-made sample description file.

```
H3K27ac=dba(sampleSheet="H3K27ac_diffbind2.csv")
```

Then we check the correlation between the peak sets (occupancy analysis).

```
plot(H3K27ac)
```

This first plot does not take into account the read counts at those peaks, so in the next step we count the reads falling in peak regions. Using the `minOverlap` parameter we can adjust in how many samples the analysed peaks have to be present. By setting it to 2 we consider peaks that are shared by biological replicates.

```
H3K27ac=dba.count(H3K27ac, minOverlap=2)
```

We specify the factor that separates our samples into the groups we want to compare. `minMembers` allows us to set the number of replicates we require per group for the analysis.

```
H3K27ac=dba.contrast(H3K27ac, categories=DBA_CONDITION, minMembers=2)
```

We use the DESeq2 method to identify the differentially acetylated reagions.

```
H3K27ac=dba.analyze(H3K27ac, method=c(DBA_DESEQ2))
```

We can extract the differentially bound regions

```
H3K27ac.DB=dba.report(H3K27ac)
```

Finally we create plots depicting the binding affinity differnces at the differentially acetylated regions. The MA plot shows the average normalised read counts on the X-axis and the log-fold change on the Y-axis, and it can be used to assess the effect of the normalisation on the data as well as observing which are the significant differentially bound regions.

```
dba.plotMA(H3K27ac)
```

PCA (principle component analysis) plot shows how the analysed samples cluster according to the normalised read counts at the significantly differential regions (FDR<0.05).

```
dba.plotPCA(H3K27ac, contrast=1, label=DBA_CONDITION)
```

Boxplots are useful to view how read distributions differ between classes of binding sites. The first two boxes show normalised readcounts accross all regions in the analysis and the following four show those regions are significantly more bound in one condition and the other. If there are differences accross all regions, then the normalisation is not appropriate or the regions selected are specific to one of the condition.

```
dba.plotBox(H3K27ac)
```

Another plot that gives an idea about how different the identified regions are is the heatmap, which shows significantly differentially bound regions as rows and the samples as columns. The samples are clustered according to the normalised read counts at these regions.

```
dba.plotHeatmap(H3K27ac, contrast=1, correlations=FALSE)
```

Understanding the DiffBind plots

1. Is the normalisation appropriate for these samples? ______________________

   ____

2. Do the significantly differentially bound regions show consistent differences between the two cell types? ______________________________

   ____

✓ CONGRATULATIONS! You've made it to the end of the practical.

We hope you enjoyed it! Don't hesitate to ask any questions and feel free to contact us any time (email addresses on the front page).

Bonus Exercise I

Within DiffBind there is the possibility to use EdgeR as the method for differential binding analysis (DBA_EDGER). Use EdgeR instead of DESeq2 in the analysis.

Questions

1. How do the results from DESEq2 and EdgeR compare? Which identified more regions? Which set of identified differentially bound regions segregate the two cell types better in PCA? ____

   ____
2. Do both methods use acceptable normalisation? ________________________

   ____

Bonus Exercise II

Given the high rate of PCR replicates in some of our samples it would be advisable to remove duplicates from the BAM files used in the differential analysis. Run the analysis without duplicaates included. Hint: We have already removed the duplicates for the MACS2 peak calling, therefore you only need to change the name of the path to the bam files in the `DiffBind` samplesheet.

Questions

1. Do the samples cluster better with or without duplicates? ________________

   ____
2. How does removing the duplicates affect the observed level of change? ________

   ____

Bonus Exercise III

As H3K27ac peaks are broader than most transcription factor peaks, it would be worthwhile to compare the results we obtained with using standard callpeaks in MACS2 with those you would get if the peak regions were called with the –broad option. Call broad peaks and use those as the regions tested for differential acetylation. Note: `broadPeaks` file format should still be indicated as `narrowPeaks` in the sample sheet, as the columns are the same, and there is no separate option within `DiffBind` for `broadPeaks`.

Questions

1. How do the results compare in terms of number of differential regions and clustering of biological replicates? ____
   ____