# ChIP-seq Biological Replicates Exercise

Myrto Kostadima < *kostadim@ebi.ac.uk* >

Advanced ChIP-seq data analysis
University of Cambridge
February, 2018

# General information

The following standard icons are used in the hands-on exercises to help you locating:

- i Important Information
- 🗎 General information / notes
- 👣 Follow the following steps
- Q Questions to be answered
- ! Warning – PLEASE take care and read carefully
- ✸ Optional Bonus exercise
- ✸ Optional Bonus exercise for a champion

## Resources used

Samtools: http://samtools.sourceforge.net/

BEDTools: http://code.google.com/p/bedtools/

CHANCE: https://github.com/songlab/chance

deepTools: https://deeptools.readthedocs.io/en/latest/index.html

IDR: https://github.com/nboley/idr/

## Additional resources:

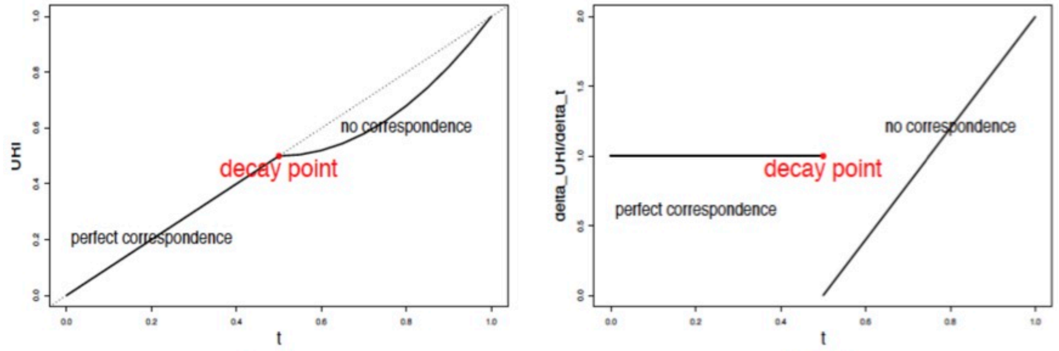
Original Data from: http://encodeproject.org

# Introduction

Many projects use biological or technical replicates to test the validity of ChIP-seq experiments, for example, all ENCODE experiments are performed with at least two replicates, either isogenic replicates for cell lines or biological replicates for primary tissues. The goal of this practical is to run the ENCODE method for consolidating ChIP-seq peak calls across biological replicates, using a method developed within the project, called the Irreproducible Discovery Rate (IDR).

## Irreproducible Discovery Rate (IDR) - Theory

The IDR method was developed by Qunhua Li and Peter Bickel's group and is extensively used by the ENCODE and modENCODE projects and is part of their ChIP-seq guidelines and standards. The method compares two lists of ChIP-seq peaks, and statistically assesses the point where the ranking in the list is no longer conserved between the replicates. The IDR method can be represented graphically as below. First, the peaks from the two replicates are sorted by some metric (e.g. p value). You can then plot for each top *X* list, the number of peaks shared between the replicates (Figure 1a).



a. Ψ

b. Ψ′

Figure 1: Taken from http://www.personal.psu.edu/users/q/u/qul12/IDR101.pdf.

In this idealised experiment, where the top ranked peaks are the same up to a point (*the decay point*), and after the ranking is random, the line will remain close to the diagonal up to the decay point, and after will move away from the line, before returning to the line when all peaks are included (we use relaxed peak calling thresholds and assume that the set of peaks is the same in both replicates).

Figure 1b shows the gradient or slope of the line in Figure 1a, and so we would expect this to be relatively flat until the decay point, when the gradient becomes smaller.

The authors have used this principle to define numerical cutoffs on peaks called after merging the replicates, called IDR thresholds, where an IDR of 0.05 means that there is a 5% chance that a peak called is not reproducible.

In this practical, we will again be using ChIP-seq data for the PAX5 transcription factor, generated on the GM12878 lymphoblastoid cell line.

BAM files for the two replicates and a matched control Input file have been downloaded from http://www.encodeproject.org, named `PAX5_BR1TR1_ENCFF614SHU.bam`, `PAX5_BR2TR1_ENCFF475VQL.bam` and `Control.bam`, respectively.

We will first calculate genome-wide correlation of the read coverage of the two *PAX5* ChIP-seq replicates, a basic measure of reproducibility across experiments.

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

```
cd ~/Desktop/ChIP-seq_idr/
```

The first step towards calculating pair-wise correlations between the two biological replicates is to compute the read coverages over genomic regions using the BAM files. The analysis can be performed for the entire genome by running the tool *multiBamSummary* in 'bins' mode. By default the coverage calculation is done for consecutive bins of 10 kilobases long.

```
multiBamSummary bins --bamfiles
   processed_data/bam/PAX5_BR1TR1_ENCFF614SHU.bam
   processed_data/bam/PAX5_BR2TR1_ENCFF475VQL.bam --labels PAX5_BR1TR1
   PAX5_BR2TR1 -out processed_data/coverage/PAX5_read_counts.npz
   --outRawCounts processed_data/coverage/PAX5_read_counts.tab
```

Finally we compute the Pearson correlation of these two *PAX5* ChIP-seq profiles, using *plotCorrelation*.

```
plotCorrelation -in processed_data/coverage/PAX5_read_counts.npz
   --corMethod pearson --skipZeros --plotTitle "Pearson Correlation of
   Average Reads per Bin" --whatToPlot scatterplot -o
   plots/PAX5_pearson_corr.png --outFileCorMatrix
   processed_data/coverage/PAX5_pearson_corr.tab --removeOutliers
```

## Questions

1. What is the correlation between these two samples? ____________________
2. Is this higher or lower than you expected? ____________________

   ____________________________________________

# Working with two biological replicates - CHANCE statistics

CHANCE (ChIP-seq ANalytics and Confidence Estimation) is a standalone package for ChIP-seq quality control and protocol optimization [Diaz et al., Genome Biology (2012)]. The authors provide a user-friendly graphical software that among others functionalities it estimates the strength and quality of immunoprecipitations (CHANCE is available at https://github.com/songlab/chance). For the purposes of our practical we will calculate the enrichment of our two PAX5 biological replicates over our Control sample, not using the tool itself, but another popular tool, called deepTools (https://deeptools.readthedocs.io/en/latest/).

Firstly we need to make sure that all our BAM files, for both replicates and Control samples are indexed. To do that please have a look at the `processed_data/bam` folder if all BAM files have their respective *.bam.bai* file. If any index file is missing, please create it using `samtool sindex`.

Once all index files have been created run the following command to calculate the enrichment of sample `PAX5_BR1TR1_ENCFF614SHU` within bin sizes of 1000bp.

```
plotFingerprint -b processed_data/bam/PAX5_BR1TR1_ENCFF614SHU.bam
   processed_data/bam/Control.bam --binSize 1000 --labels PAX5_BR1TR1
   Control --JSDsample processed_data/bam/Control.bam
   --outQualityMetrics PAX5_BR1TR1_ENCFF614SHU_fingerprint.txt -plot
   PAX5_BR1TR1_ENCFF614SHU_fingerprint.png
```

For a detailed explanation of the output format of the *_fingerprint.txt* file, please see here https://deeptools.readthedocs.io/en/latest/content/feature/plotFingerprint_QC_metrics.html.

Now run the bove command for the other replicate too. Once the tool has finished open both the PNG files created.

## Questions

1. Is the enrichment similar for both PAX5 biological replicates? If not, what differences do you observe? ________________________________

# Working with two biological replicates - Irreproducible Discovery Rate (IDR) analysis

Peak calling with *MACS2* is performed for both of the replicates individually and after merging the alignments from both replicates, using e.g. samtools merge, to create a pooled peak set.

As we have covered peak calling in the ChIP-seq practical, these data sets are provided for you in the `processed_data/macs2` directory, labelled `PAX5_BR1TR1_peaks.bed`, `PAX5_BR2TR1_peaks.bed` and `PAX5_pooled_peaks.bed` respectively.

The narrowPeak format in which the peaks are reported is an ENCODE format, using an extension of the BED format for providing peaks and associated scores and p values. See https://genome.ucsc.edu/FAQ/FAQformat.html#format12 for full information.

The ChIP-seq experiment and alignment to the genome sequence is affected by sources of artefact read mapping caused by biases in chromatin accessibility and ambiguous alignment. These spurious regions can be removed by filtering out any peaks that overlap a blacklist, believed to contain experiment and cell type independent areas of high artefactual signal. The ENCODE blacklist has been generated by combining regions of known repeats and manually curated genomic regions of ubiquitous open chromatin and input sequence signal. The file can be downloaded from:

https://www.encodeproject.org/annotations/ENCSR636HFF/

We now remove any peaks which overlap with the ENCODE blacklist.

```
bedtools intersect -a processed_data/macs2/PAX5_BR1TR1_peaks.bed -b
   annotation/GRCh38_blacklisted_regions.bed -v >
   processed_data/macs2/PAX5_BR1TR1_peaks_filtered.bed
```

Next, we need to sort the peak files into significance order. For *MACS2*, the p-value works best as the ranking for the IDR procedure. See https://sites.google.com/site/anshulkundaje/projects/idr#TOC-Peak-callers-tested-with-IDR for the best measures to use with other peak callers.

Here, we sort by p value and then use the 100,000 most significant peaks:

```
sort -k 8fr processed_data/macs2/PAX5_BR1TR1_peaks_filtered.bed | head
   -n 100000 > processed_data/macs2/PAX5_BR1TR1_top100000_peaks.bed
sort -k 8fr processed_data/macs2/PAX5_BR2TR1_peaks_filtered.bed | head
   -n 100000 > processed_data/macs2/PAX5_BR2TR1_top100000_peaks.bed
sort -k 8fr processed_data/macs2/PAX5_pooled_peaks_filtered.bed | head
   -n 100000 > processed_data/macs2/PAX5_pooled_top100000_peaks.bed
```

Finally, we can perform the IDR analysis on the peaks called in the two technical replicates, using the peak list from the merged replicates

```
idr --samples processed_data/macs2/PAX5_BR1TR1_top100000_peaks.bed
   processed_data/macs2/PAX5_BR2TR1_top100000_peaks.bed --peak-list
   processed_data/macs2/PAX5_pooled_top100000_peaks.bed --idr-threshold
   0.05 --output-file idr/PAX5_replicates_idr.txt --plot
```

## Understanding the IDR output plot

1. How reproducible are the peaks called in these two technical replicates for PAX5 binding in the GM12878 cell line? ______________________________________________

   ____________________________________________________________________

2. What factors could lower the reproducibility between two ChIP-seq experiments? ______________________________________________

   ____________________________________________________________________

✔ 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

The IDR statistics can also be used to flag data sets with low reproducibility. This may be due to one of the two replicates being of lower ChIP enrichment, hence having a high signal-noise ratio. In this case, the standard IDR protocol would record few reproducible peaks, despite one replicate having high information content.

ENCODE has developed a rescue strategy in this case by using pseudo-replicates. These pseudo-replicates are generated by pooling all the reads, and then randomly splitting them into two files. These pseudo-replicates do not represent true biological or experimental replicates, but attempt to model the stochastic noise in the sampling of sequenced reads from a population of fragments.

The pseudo-replicates analysis uses a lower IDR threshold than biological replicates, due to the reduced noise, typically 0.0025.

Using https://sites.google.com/site/anshulkundaje/projects/idr and modifiying the code above, run the IDR analysis for pseudo-replicates of the GM12878 PAX5 ChIP-seq data.

## Questions

1. Does the IDR method select more peaks using the original technical replicates or the pseudo-replicates?