1 Project 2: ChIP-seq Biological Replicates

1.1 Introduction

This practical has been adapted from materials developed by Angela Goncalves, Myrto Kostadima, Steven Wilderand Maria Xenophontos.

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 technical 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)**.

1.2 Preparing your environment

First, go to the group_projects folder.

```
cd /home/manager/course_data/group_projects
```

Check to see if the ChIPSeq-Project2 folder exists.

```
ls ChIPSeq-Project2
```

If this folder doesn't exist, please check with your course instructor.

Once you have the data, go into the ChIPSeq-Project2 directory.

```
cd ChIPSeq-Project2
```

1.3 Irreproducible Discovery Rate

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).

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).

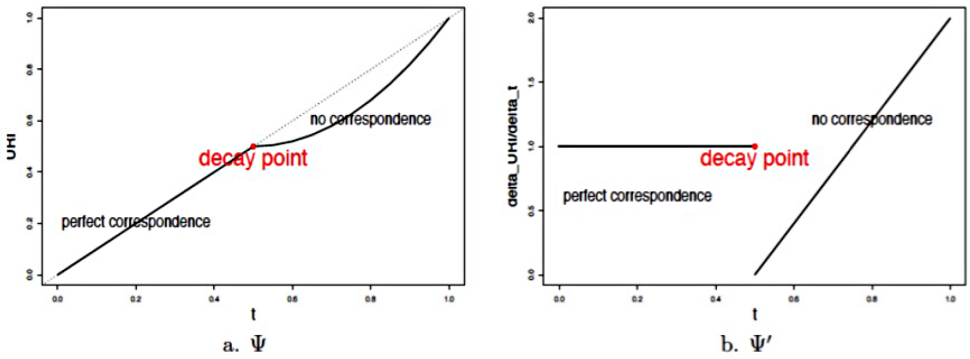

figure 1

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

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 technical replicates and a matched control Input file have been downloaded from http://www.encodeproject.org, named `ENCFF00NZI.bam`, `ENCFF00NZL.bam` and `ENCFF0000CL.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.

Calculate genome-wide coverages for one of the replicates, ENCFF000NZI. The other one has already been done.

```
genomeCoverageBed -bg -ibam ENCFF000NZI.bam -split \
      -g genome/hg19.chrom.sizes > ENCFF000NZI.bedgraph
```

Now, we will convert the files from bedGraph https://genome.ucsc.edu/FAQ/FAQformat.html#format1.8 genome indexed and compressed UCSC BigWig https://genome.ucsc.edu/goldenPath/help/bigWig.html for the same replicate.

Sort ENCFF000NZI.bedgraph.

```
LC_COLLATE=C sort -k1,1 -k2,2n ENCFF000NZI.bedgraph > ENCFF000NZI.sorted.bedgra
```

Convert the bedGraph file to BigWig.

```
bedGraphToBigWig ENCFF000NZI.sorted.bedgraph genome/hg19.chrom.sizes ENCFF000NZ
```

Finally, we compute the correlation of these two PAX5 ChIP-Seq profiles.

```
bigWigCorrelate ENCFF000NZI.bw ENCFF000NZL.bw
```

Q1: What is the correlation between these two samples?

Q2: Is this higher or lower than you expected?

Q3: What factors could explain a change in the correlation statistic?

1.4 Calculating fragment lengths from ChIP-Seq data

As part of the ChIP-seq protocol, a size selection is carried out, so that only fragments with lengths within a specified narrow range are selected for sequencing. This desired length is chosen by the experimenter, and for transcription factor ChIP-seq is usually around the size of a nucleosome (147 base pairs).

However, the true sizes chosen can differ from the targeted lengths, and this can affect peak calling, so we try to estimate the true fragment length using the ChIP-seq data, usually by comparing the reads aligning to the forward and reverse strands.

As the fragments span the transcription-factor bound region, but only the ends are sequenced, the reads map upstream of the binding site on the forward strand and downstream on the reverse strand. One way to calculate the fragment length is to calculate the correlation between the read density across the two strands, when you shift one strand's signal relative to the other by different amounts. The correlation should have a maximum at the fragment length.

For data sets with lower enrichment, or high amplification bias, there will also be a peak in this cross-correlation profile at a shift equal to the read length, due to reads mapping to repeats, hence the reads on both strands aligning to the same location.

Phantompeakqualtools uses the ratio of the height of the fragment length peak and the height of the read length peak to give a measure of sample quality.

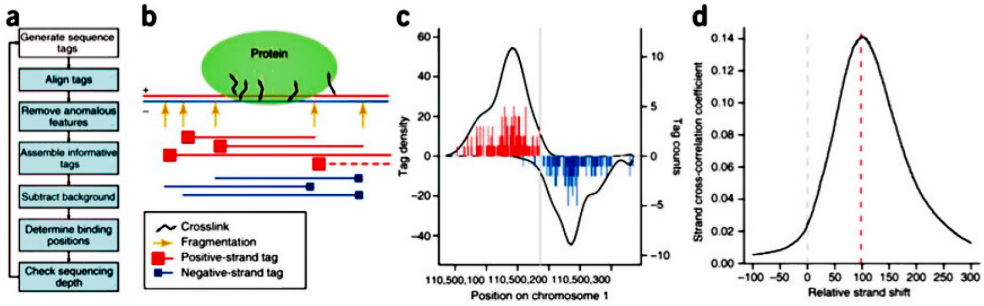

figure 2

Figure 2: Strand-specific profile.

Calculate the fragment length for ENCFF000NZI.bam.

```
R CMD BATCH -vanilla -args '-s=-100:5:500 -rf
          -c=ENCFF000NZI.bam -out=ENCFF000NZI_spp.out
          -savp=ENCFF000NZI.pdf'
          /usr/local/bin/run_spp.R runSPP_ENCFF000NZI.log
```

Note: We have already run run_spp.R for the other PAX5 replicate and the control sample.

Phantompeakqualtools gives a quality tag to each sample between -2 and +2 (codes:-2:very Low, -1:Low, 0:Medium, 1:High, 2:very High).

Q1. What scores have our samples been given?

Q2. What explanation can you think of for these scores?

Q3. How could the quality of the samples be improved?

Now, 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 ENCFF000NZI_peaks.narrowPeak, ENCFF000NZL_peaks.narrowPeak and PAX5_pooled_peaks.narrowPeak respectively.

The narrowPeak format 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: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz

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

```
bedtools intersect -a ENCFF000NZI_peaks.narrowPeak \
       -b wgEncodeDacMapabilityConsensusExcludable.bed \
       -v > ENCFF000NZI_peaks_filtered.narrowPeak
```

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 8nr,8nr ENCFF000NZI_peaks_filtered.narrowPeak | \
      head -n 1000 > ENCFF000NZI_top100000_peaks.narrowPeak
```

```
sort -k 8nr,8nr ENCFF000NZL_peaks_filtered.narrowPeak | \
      head -n 100000 > ENCFF000NZL_top100000_peaks.narrowPeak
```

```
sort -k 8nr,8nr PAX5_pooled_peaks_filtered.narrowPeak | \
      head -n 100000 > PAX5_pooled_top100000_peaks.narrowPeak
```

Finally, we can perform the IDR analysis on the peaks called in the two technical replicates.

```
Rscript batch-consistency-analysis.r \
      ENCFF000NZI_top100000_peaks.narrowPeak \
      ENCFF000NZL_top100000_peaks.narrowPeak \
      -1 ENCFF000NZI_vs_ENCFF000NZL 0 F p.value
```

Now, produce a graphical representation of the IDR output.

```
Rscript batch-consistency-plot.r 1 PAX5_2Reps ENCFF000NZI_vs_ENCFF000NZL
```

As the output of the above command (idr/PAX5_2Reps-plot.ps) cannot be opened on these PCs, convert it to a PDF file.

```
ps2pdf idr/PAX5_2Reps-plot.ps idr/PAX5_2Reps-plot.pdf
```

1.5 Understanding the IDR output plot

The top left of the five plots is the equivalent of Figure 1a, considering all peaks, and the top centre figure is the equivalent of Figure 1a, only considering overlapping peaks called in both replicates. The number of peaks in common in the top X peaks of both replicates is shown in each case.

The two figures below these plot the slope (gradient) of the above graph, again for all peaks and matched peaks respectively. The top right figure shows the IDR value increasing as the rank of the peaks goes down. You can use this plot to see how many peaks pass the IDR test for different threshold values. A good value for the IDR cutoff in this case is 0.05. See also http://www.personal.psu.edu/users/q/u/qul12/IDR101.pdf.

Q4. How reproducible are the peaks called in these two technical replicates for PAX5 binding in the GM12878 cell line?

Q5. What factors could lower the reproducibility between two ChIP-seq experiments?

Open the file ENCFF000NZI_vs_ENCFF000NZL-overlapped-peaks.txt to see how many peaks pass the IDR threshold of 0.05.

```
less ENCFF000NZI_vs_ENCFF000NZL-overlapped-peaks.txt
```

We now use this number of reproducible peaks, generated on the individual replicate peak calls, to select the same number of peaks from the pooled replicates peak calls.

Replace [numPeaks] with the number of peaks passing the 0.05 threshold in the following command.

head -n [numPeaks] PAX5_pooled_top100000_peaks.narrowPeak > PAX5_pooled_macs2.conservative.narrowPeak

This file is the final output of IDR, generating a final peak set on the merged replicates. IDR can also be used with more than two replicates. In this case, the IDR procedure is followed for all pairwise comparisons, and the highest number of peaks passing the IDR cutoff is used for filtering the pooled peak list.

CONGRATULATIONS! You've made it to the end of the practical. We hope you enjoyed it!

1.6 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.

Q6: Does the IDR method select more peaks using the original technical replicates or the pseudo-replicates?