UNIVERSITÄT BERN

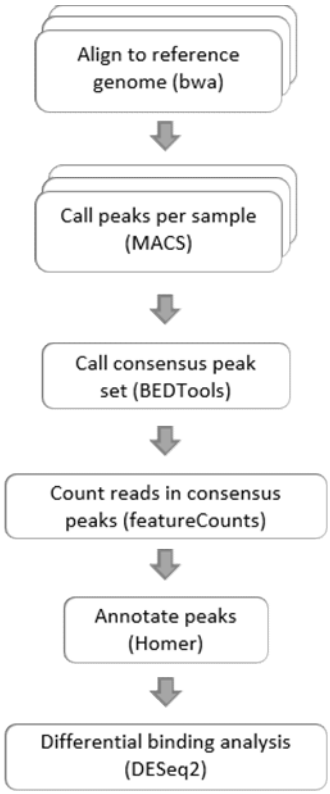
ChIP-seq Differential binding analysis

Interfaculty Bioinformatics Unit, University of Bern

01.06.2021

Swiss Institute of Bioinformatics

Workflow



The analysis uses the nf-core chip-seq v. 1.2.2 pipeline. The steps run by the pipeline are described here https://nf-co.re/chipseq.

In short, reads are mapped to the reference genome using bwa v0.7.17-r1188 and peaks are called in each sample separately with MACS2 v2.2.7.1. The sample-specific peaks are then combined into a consensus peak set with BEDTools v2.29.2 and the number of reads in each sample and peak interval is re-called with featureCounts v2.0.1. The resulting count matrix is used as input to DESeq2 v. 1.26.0 in R v3.6.3 for differential binding analysis. The peaks are annotated with Homer v.4.11.

Output files

UNIVERSITÄT BERN

Quality reports

- multiqc: detailed overview of various quality measures. See video linked in the report for more info.

Differential binding (DB) results:

- One table per pairwise contrast (see following slides for details).

Peak info

- <antibody>.consensus_peaks.annotatePeaks.txt: Annotation of peak intervals (see following slides for details)
- <antibody>. consensus_peaks.bed: File specifying the position of the peaks in bed format (https://m.ensembl.org/info/website/upload/bed.html). This can be used, for example, to display the position of the peaks in a tool like IGV (https://software.broadinstitute.org/software/igv/)
- <antibody>.consensus_peaks.boolean.intersect.plot.pdf: Shows overlap of peaks between samples. The bottom panel shows a black dot for each sample in which a particular peak was detected. See here for more info: https://jku-vds-lab.at/tools/upset/

DA tables

UNIVERSITÄT BERN

You will receive one table per pairwise contrast (i.e comparison of 2 experimental groups). The file name is structured as follows: <experimental group 1>vs<experimental group 2>. deseq2.results.txt.

Column	Content
Geneid	Peak ID
Chr, Start, End, Strand	Location of the peak by chromsome, start and end position and strand
Length	Length of the peak region
baseMean	Mean number of reads within the peak
log2FoldChange	Log2 of the ratio of counts in group 1 versus group 2
lfcSE	Standard error for the log2 fold-change
stat	Test statistic (used to calculate P-value)
pvalue	P-value (not corrected for multiple testing)
padj	False discoveray rate adjusted P-value. This is the P-value that should be considered.
<sample>.raw	Raw counts per peak and sample
<sample>.pseudo	Normalized counts per peak and sample. These counts have been adjusted to account for differences in sequencing depth between samples

Peak annotation

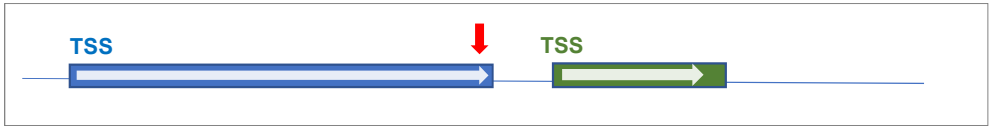
UNIVERSITÄT BERN

File: **<antibody>.consensus_peaks.annotatePeaks.txt**

There are 5 columns with peak ID and genomic position (analogous to DB table). This is followed by two empty columns (Peak Score and Focus Ratio/Region Size), and columns with the annotation as follows:

Column	Content
Annotation	Location within the gene (e.g. exon, intron etc)
Detailed.Annotation	Similar to "Annotation" but with more detail in some cases
Distance.to.TSS	Distance to nearest transcription start site (negative values mean upstream of the TSS, positive values mean downstream)
Nearest.PromoterID	
Entrez.ID	
Nearest.Unigene	
Nearest.Refseq	
Nearest.Ensembl	Various IDs and info on the gene
Gene.Name	
Gene.Alias	
Gene.Description	
Gene.Type	

The fields shaded in blue refer to the gene with which **the centre of the peak** overlaps.
The fields shaded **in green**, on the other hand, give information about the gene whose transcription start site (TSS) is closest to the centre of the peak. Often, the blue and green fields will refer to the same gene but this is not always the case.
In the example below, the peak of interest (indicated by red arrow) falls within an exon of the blue gene, and the blue fields in the table will contain information about this blue gene. The nearest TSS, however, is that of the green gene, and the green fields in the table refer to the green gene.



Often, a given position will overlap with multiple possible annotations. In this case, annotations are prioritised as explained in the Homer manual: http://homer.ucsd.edu/homer/ngs/annotation.html