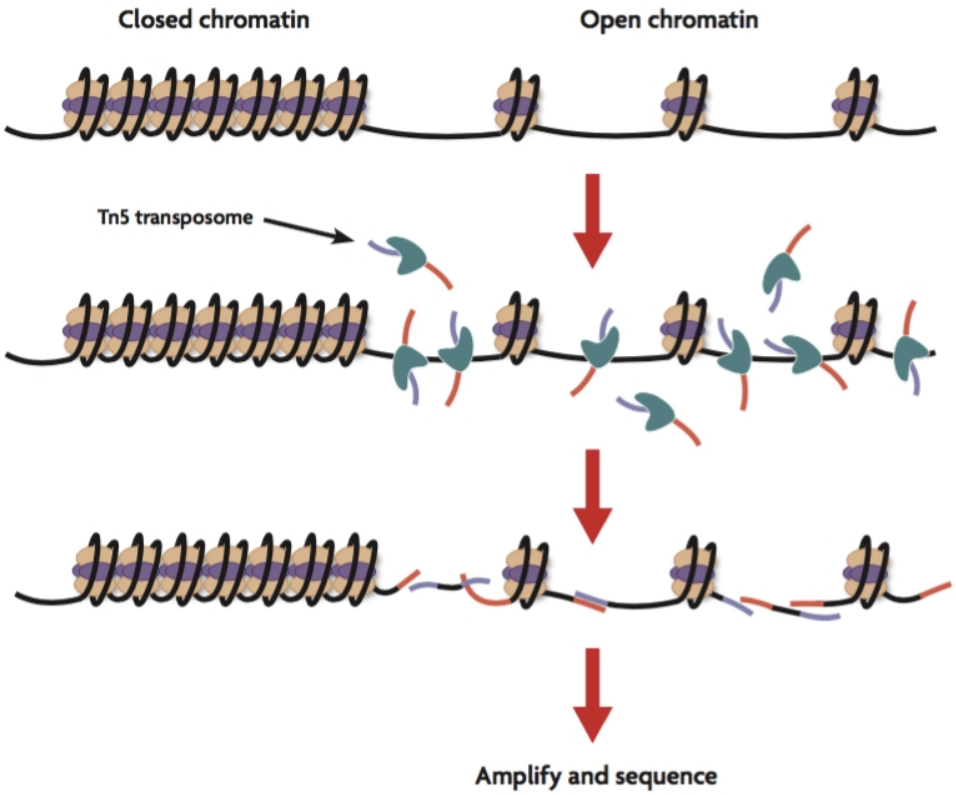
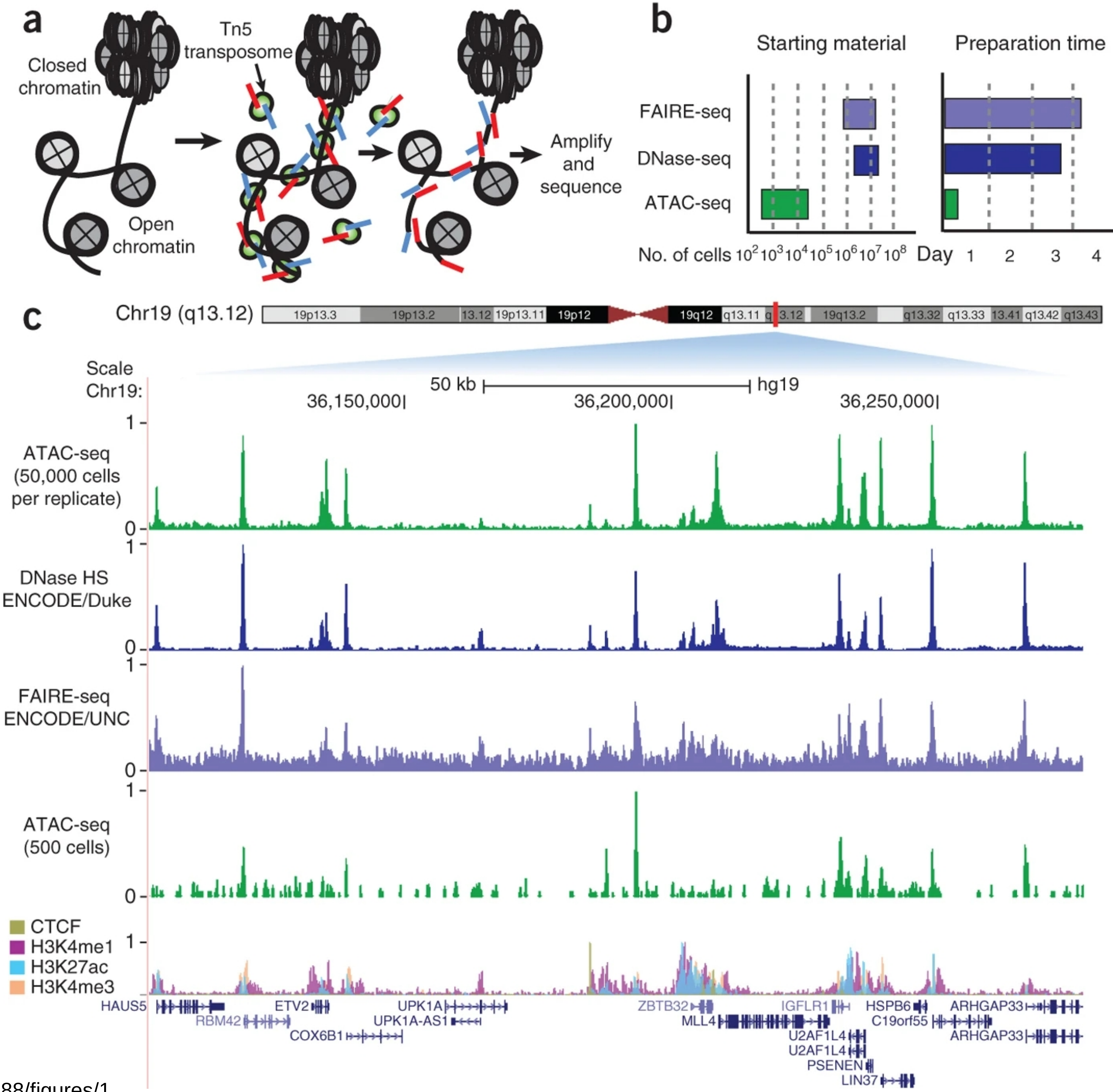
ATAC-Seq Anaysis pipeline at the MPI-AGE Bioinformatics Core Facility



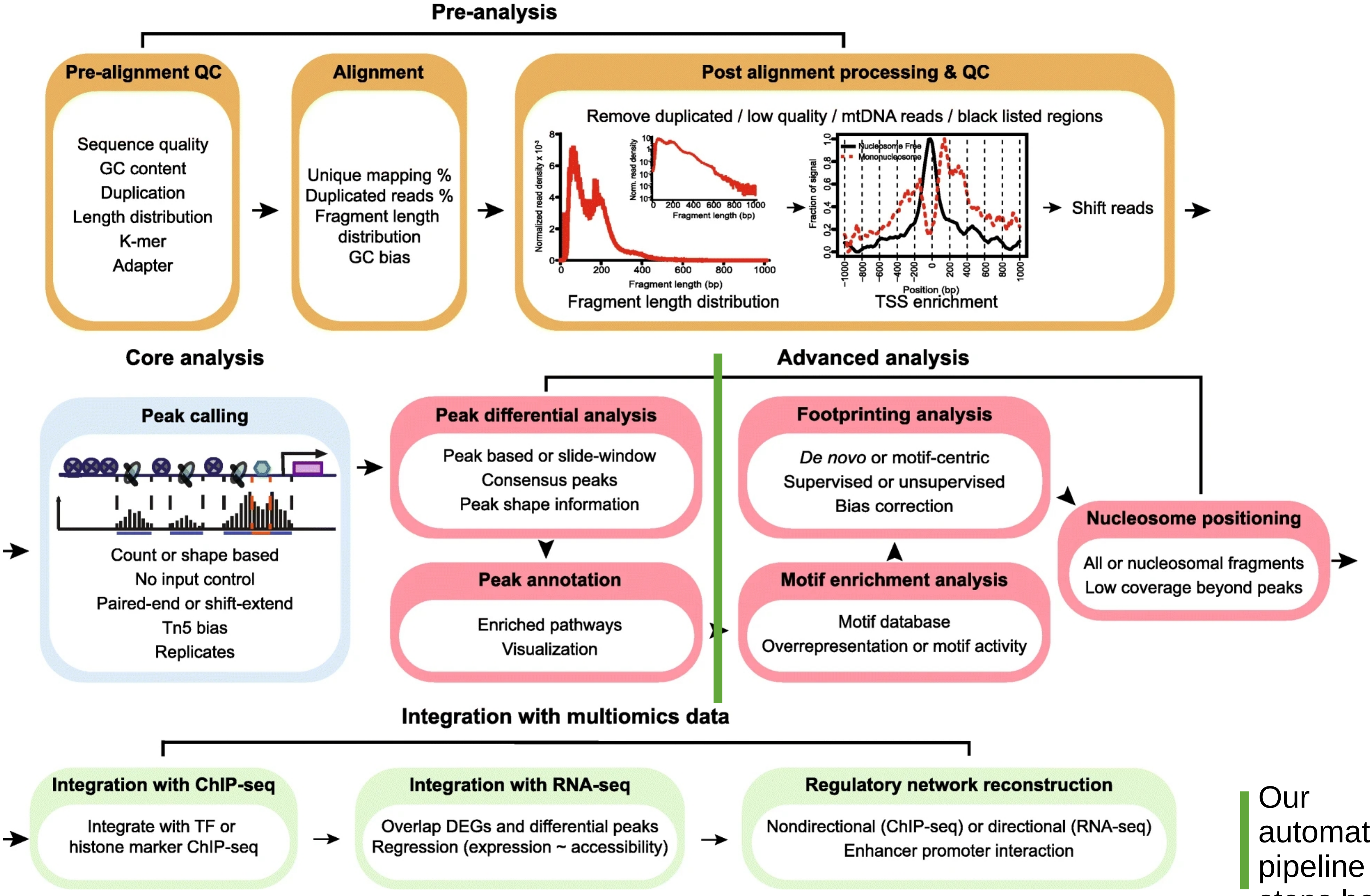
https://www.activemotif.com/blog-atac-seq

This overview clarifies the different type of expected signals, input type, and preparation time of ATAC-seq, Dnase, and FAIRE-Seq



https://www.nature.com/articles/nmeth.2688/figures/1

Overview



Pre-analysis

Pre-alignment QC

Sequence quality
GC content
Duplication
Length distribution
K-mer
Adapter

Alignment

Unique mapping %
Duplicated reads %
Fragment length distribution
GC bias

Post alignment processing & QC

Remove duplicated / low quality / mtDNA reads / black listed regions

Fragment length distribution

TSS enrichment

Shift reads

Core analysis

Peak calling

Count or shape based
No input control
Paired-end or shift-extend
Tn5 bias
Replicates

Advanced analysis

Peak differential analysis

Peak based or slide-window
Consensus peaks
Peak shape information

Peak annotation

Enriched pathways
Visualization

Footprinting analysis

De novo or motif-centric
Supervised or unsupervised
Bias correction

Motif enrichment analysis

Motif database
Overrepresentation or motif activity

Nucleosome positioning

All or nucleosomal fragments
Low coverage beyond peaks

Integration with multiomics data

Integration with ChIP-seq

Integrate with TF or histone marker ChIP-seq

Integration with RNA-seq

Overlap DEGs and differential peaks
Regression (expression ~ accessibility)

Regulatory network reconstruction

Nondirectional (ChIP-seq) or directional (RNA-seq)
Enhancer promoter interaction

Our automatic pipeline stops here

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1929-3

Experimental Setup

- ATAC Seq libraries
- Multiple conditions
- At least two replicates per condition
- Recommendation:
  - 50 Mio mapped reads for reliably detecting changes in open chromatin
  - 200 Mio mapped reads for TF footprinting
  - Paired end
  - Fragments can range between
    - < 100 BP for nucleolsome free regions
    - ~ 200 BP for mono-nucleosomes
    - ~ 400 BP for di-nucleosomes
    - etc.

Pipeline

1) Quality control with `fastqc`

2) Adapter trimming with `flexbar`

3) Mapping with `bowtie2`

4) Cleaning reads

- Remove mito reads awk `'{if (${chromosome} != "MT" ) print }'`
- MAPQ > 10 `samtools view -q 10`
- Properly paired `samtools view -f 2`
- Removing PCR duplicated with `Picard`

5) General QC of ATAC seq with `ATACSeqQC`

6) Peak calling with `MACS2`

7) Converting peaks to bigwigs for UCSC genome browser

8) Finding differential binding sites using `DiffBind`

9) Annotating peaks using `ChIPseeker`

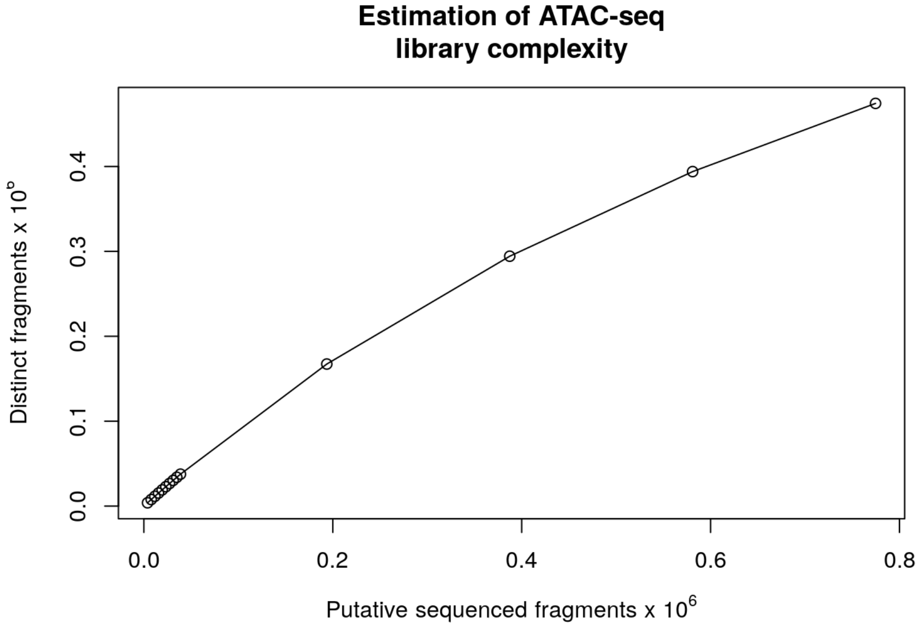
Results

```
[top/series]/
  ├── ATACSeqQC_output
    ├── ATACSeqQC results per sample
    ├── chromosome_heatmaps.[sample]/
    └── splited
       ├── [sample]/
       └── [sample].shifted.bam
  ├── bigwig_output
    └── peaks in bigwig format
  ├── diffbind3_output
    ├── differential binding results
    └── annotated peaks
  └── macs2_output
     └── peaks data
```

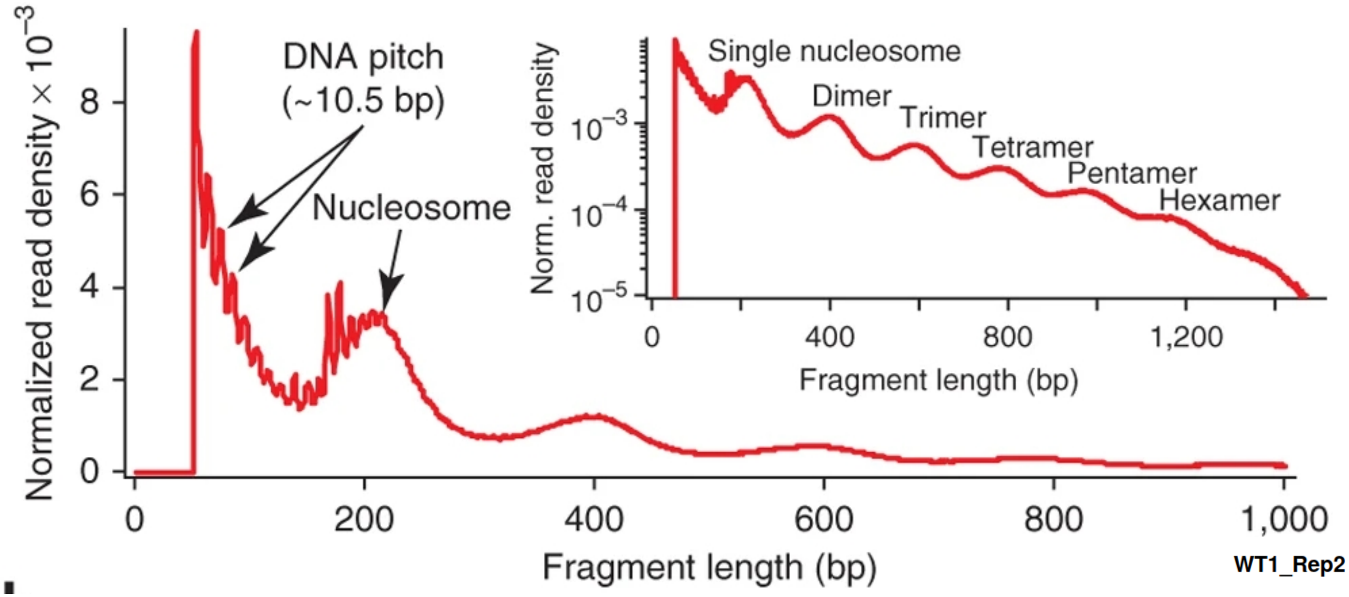
ATACSeq QC

- To help researchers quickly assess the quality of ATAC-seq data, we have developed the ATACseqQC package for easily making diagnostic plots following published guidelines.
- When running the pipeline please make sure you are using the correct refrences

- The first step is the estimation of library complexity.
- This is assessed using the bam files before pcr duplicate removal.
- There is no specific explanation given for the interpretation of this result



ATACSeq QC



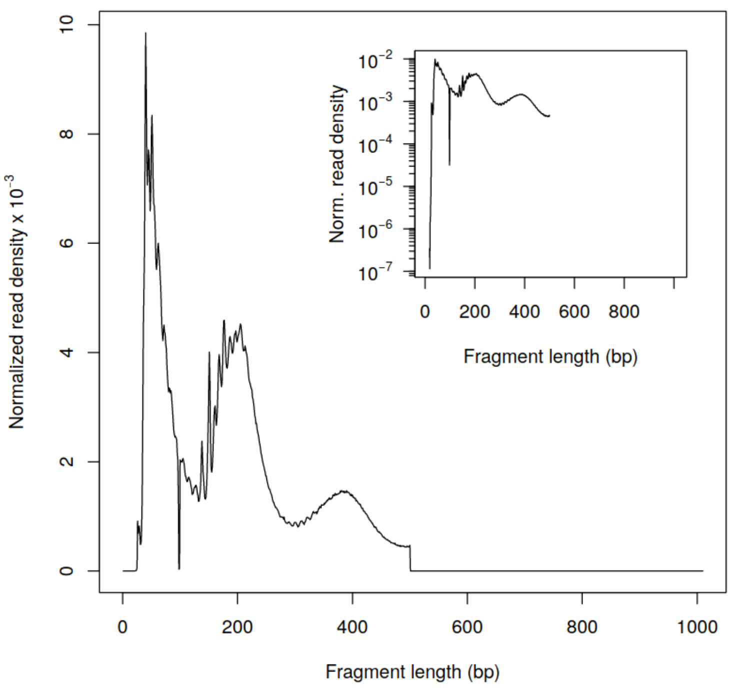
From their tutorial: Fragment size distribution

First, there should be a large proportion of reads with less than 100 bp, which represents the nucleosome-free region. Second, the fragment size distribution should have a clear periodicity, which is evident in the inset figure, indicative of nucleosome occupacy (present in integer multiples).

Top: representative figure with explanation from https://www.nature.com/articles/nmeth.2688

Bottom: example of fragment size distribution of WT sample generated by Costas

WT1_Rep2 fragment sizes



ATACSeq QC

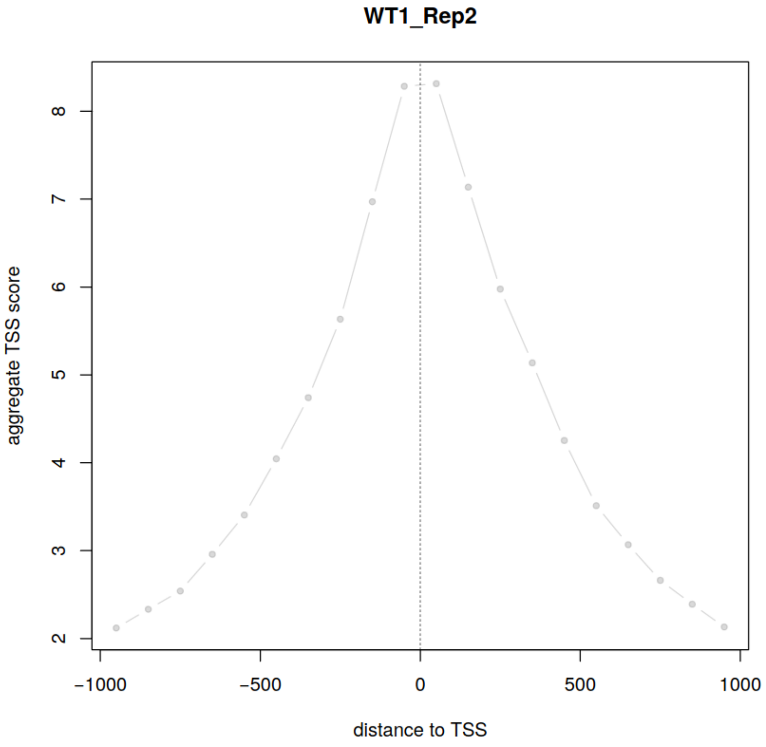
From their tutorial: Transcription Start Site (TSS) Enrichment Score

TSS enrichment score is a ratio between aggregate distribution of reads centered on TSSs and that flanking the corresponding TSSs.

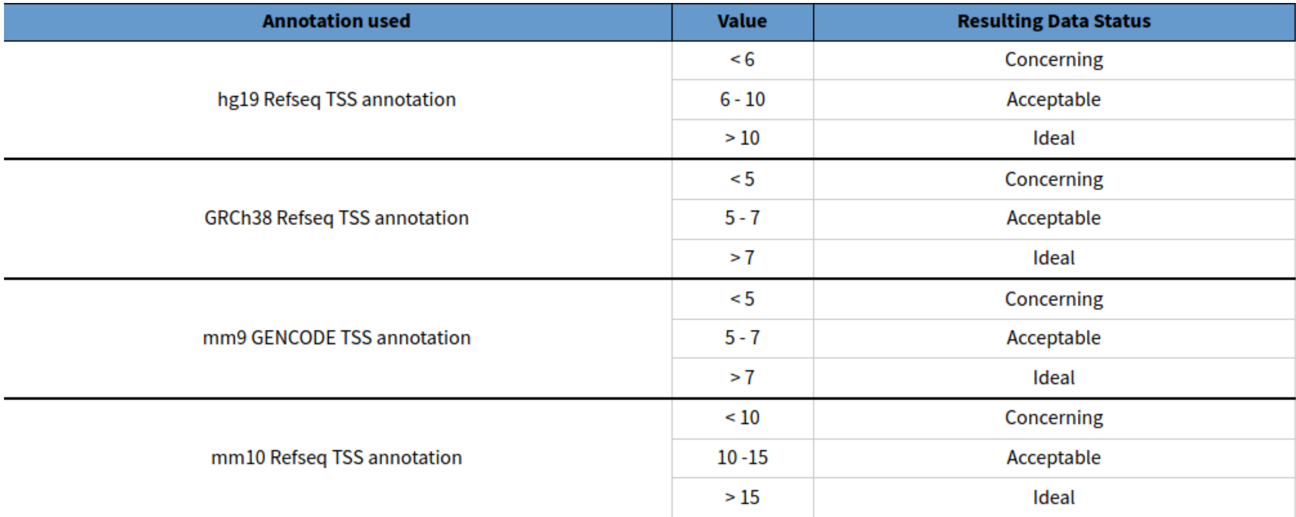
TSS score = the depth of TSS (each 100bp window within 1000 bp each side) / the depth of end flanks (100bp each end).

TSSE score = max(mean(TSS score in each window)).

TSS enrichment score is calculated according to the definition at https://www.encodeproject.org/data-standards/terms/#enrichment.



Transcription start site (TSS) enrichment values are dependent on the reference files used; cutoff values for high quality data are listed in the following table from https://www.encodeproject.org/atac-seq/

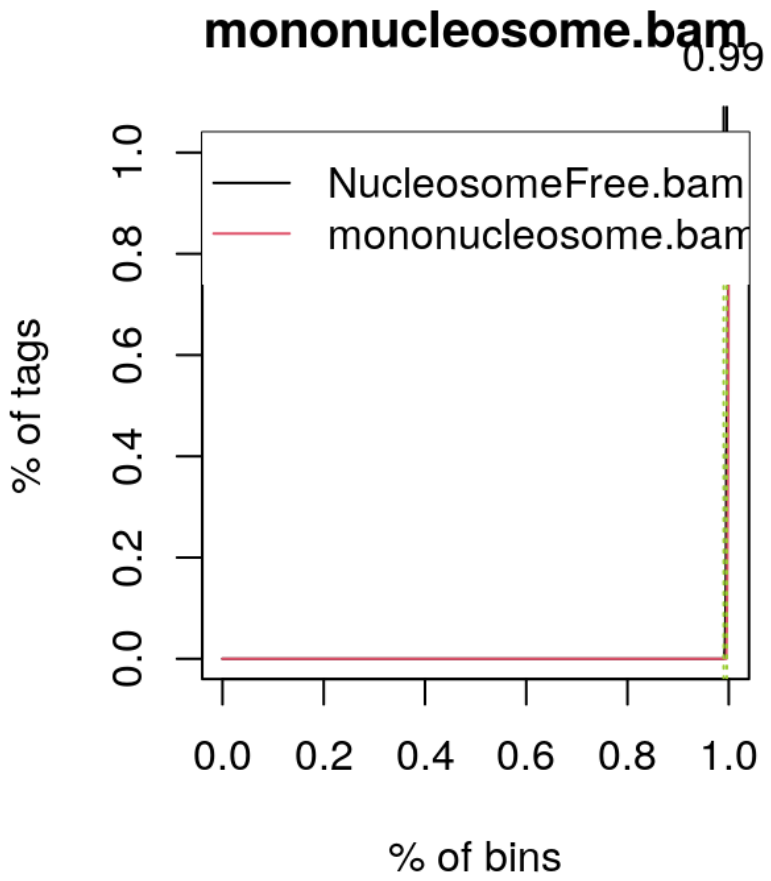


Annotation used	Value	Resulting Data Status
hg19 Refseq TSS annotation	< 6	Concerning
	6 - 10	Acceptable
	> 10	Ideal
GRCh38 Refseq TSS annotation	< 5	Concerning
	5 - 7	Acceptable
	> 7	Ideal
mm9 GENCODE TSS annotation	< 5	Concerning
	5 - 7	Acceptable
	> 7	Ideal
mm10 Refseq TSS annotation	< 10	Concerning
	10 -15	Acceptable
	> 15	Ideal

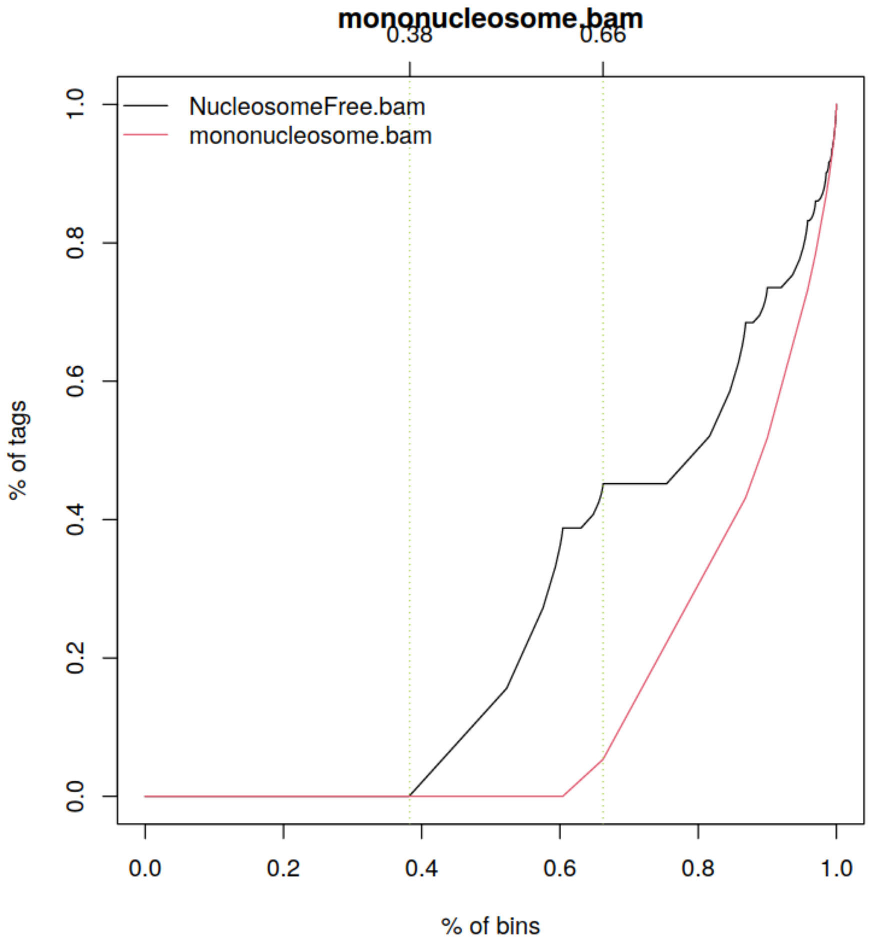
ATACSeq QC

This quality plot shows the cumulative percentage of tag allocation in nucleosome-free and mononucleosome bam files. Sadly, there is no further explanation how to interpret these plots.

ATACseqQC tutorial



Costas WT test set



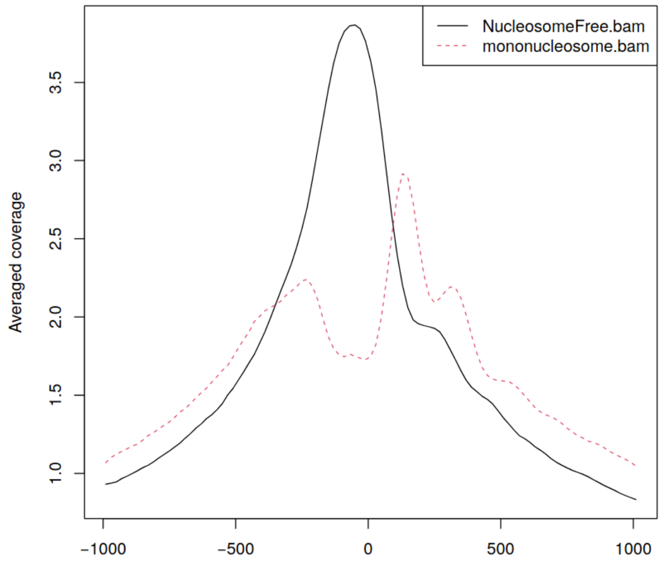
ATACSeq QC

This quality plot shows the average coverage around TSS regions of nucleosome free and mono-nucleolsome fragments. Bottom: example of ATACseqQC tutorial. Top-right, not normalized, bottom-right, normalized coverage.

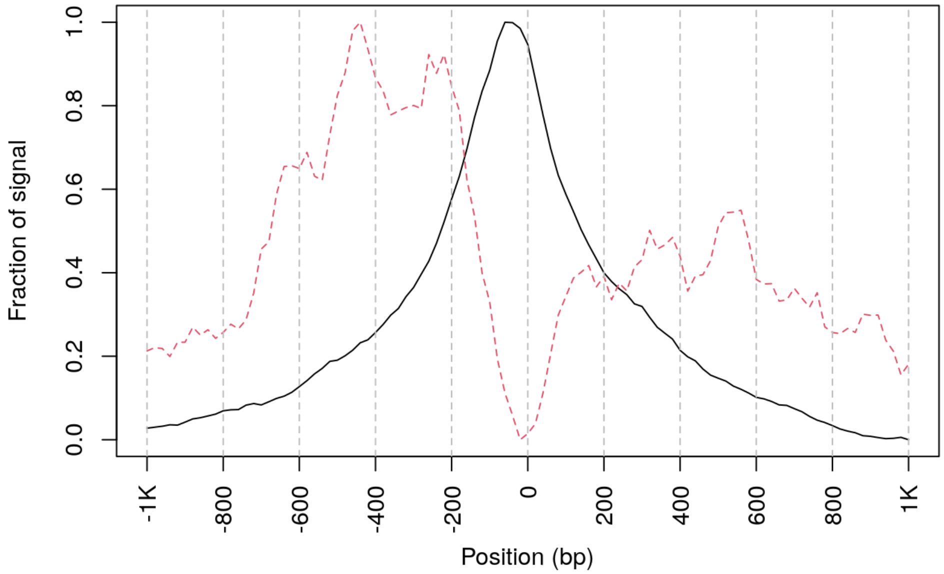
One expects a peak in the TSS region of nucleosome free track and a valley in the TSS region of the mono-nucleosome track.

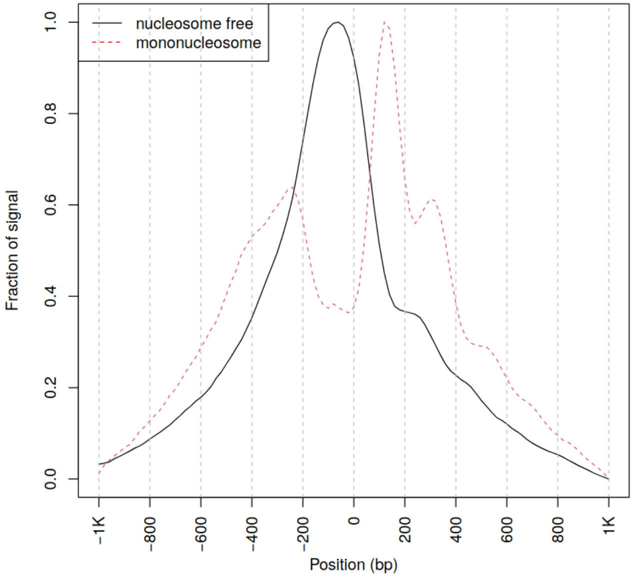
Sadly, there is no further explanation how to interpret these plots.

Costas WT test set



ATACseqQC tutorial





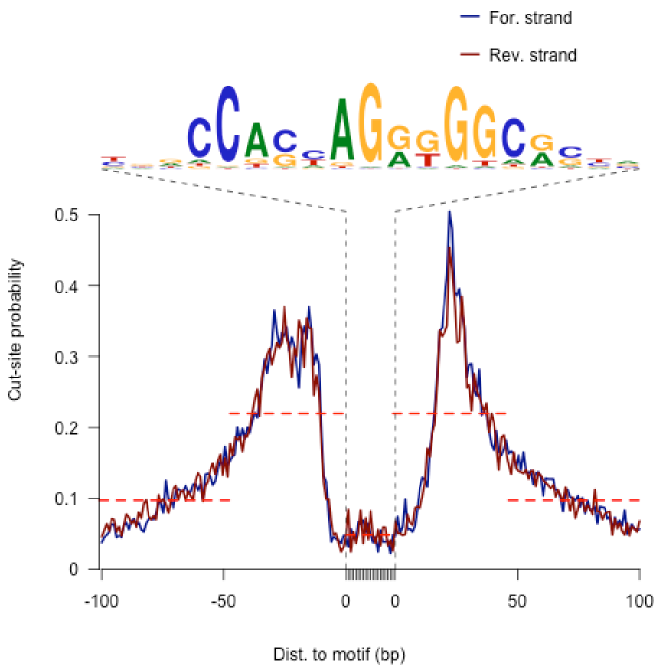
ATACSeq QC

This step is not included in the standard pipeline, but an example is given in `example_TF_scan.Rmd`.

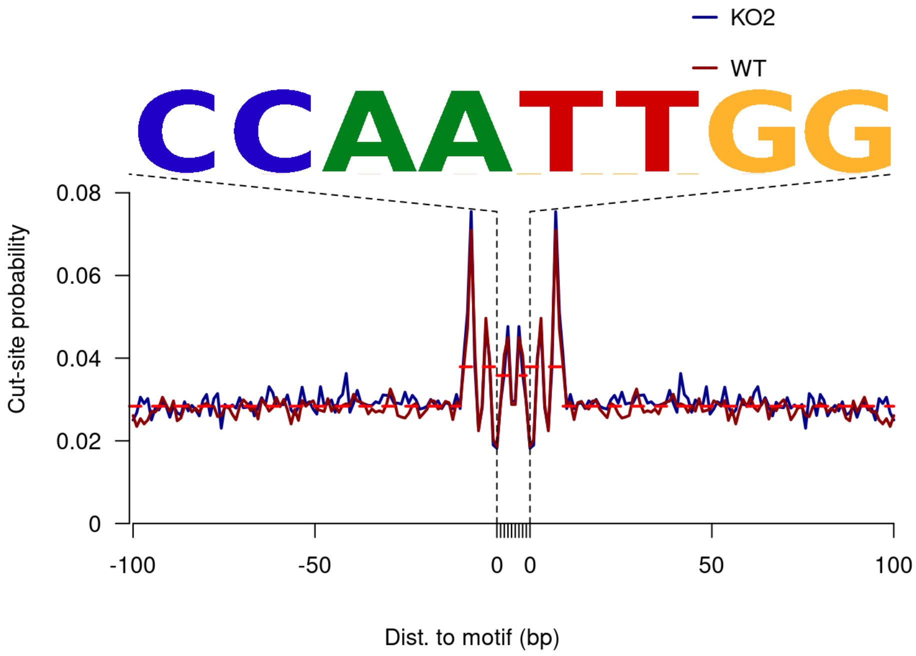
From their tutorial: plot Footprints

ATAC-seq footprints infer factor occupancy genome-wide. The factorFootprints function uses matchPWM to predict the binding sites using the input position weight matrix (PWM). Then it calculates and plots the accumulated coverage for those binding sites to show the status of the occupancy genome-wide.

ATACseqQC tutorial



Costas WT test set



DiffBind

From their Tutorial:

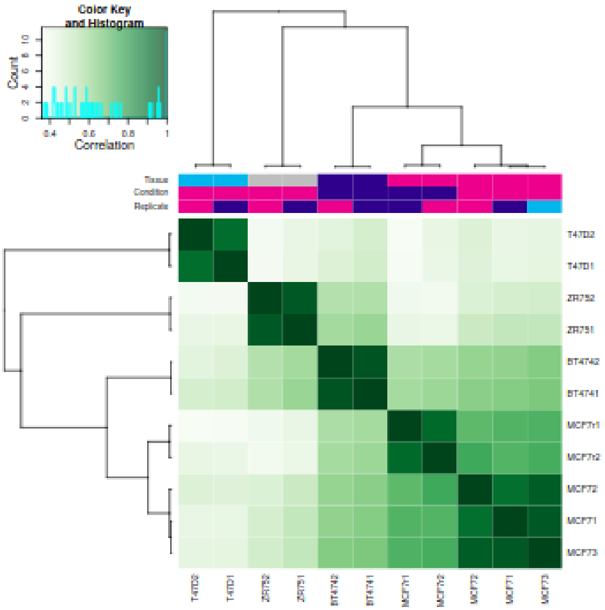
Bioconductor packageDiffBind provides functions for processing DNA data enriched for genomic loci, including ChIP-seq data enriched for sites where specific protein/DNA binding occurs, or histone marks areenriched, as well as open-chromatin assays such as ATAC-seq

The primary emphasis of the package is on identifying sites that are differentially boundbetween sample groups. It includes functions to support the processing of peak sets, in-cluding overlapping and merging peak sets, counting sequencing reads overlapping intervalsin peak sets, and identifying statistically significantly differentially bound sites based on ev-idence of binding affinity (measured by differences in read densities). To this end it usesstatistical routines developed in an RNA-Seq context (primarily the Bioconductor packages edgeR and DESeq2). Additionally, the package builds onRgraphics routines to provide aset of standardized plots to aid in binding analysis.

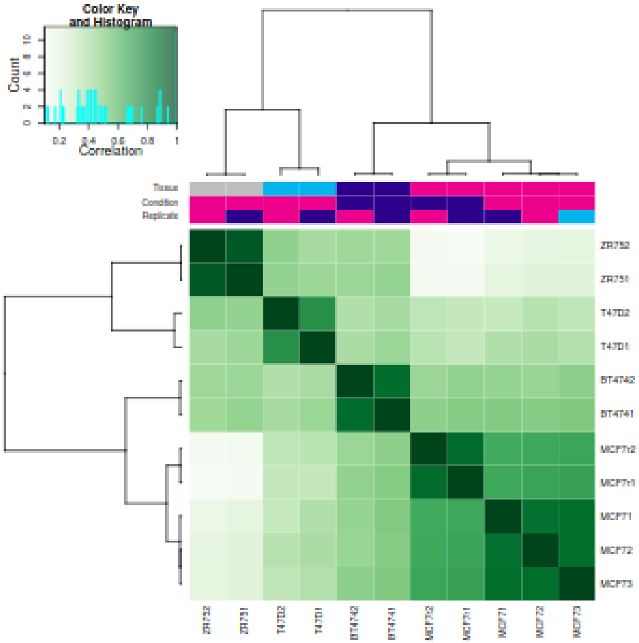
DiffBind

The pdf general_QC.pdf contains four plots using all samples from an experiment.

The first picture in `general_QC.pdf` represents a correlation heatmap based on cross-correlations of overlapping peaks.

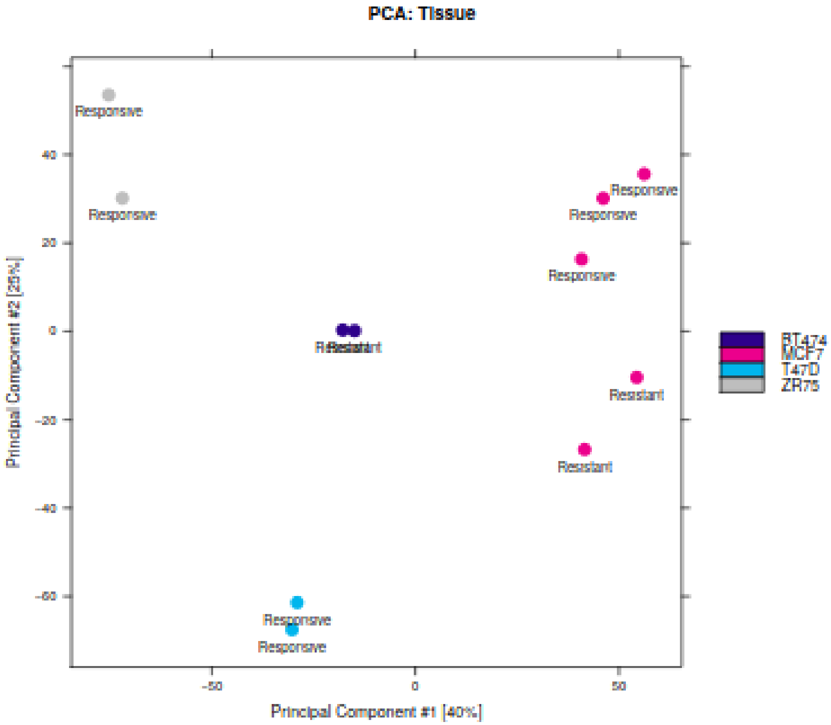


The second picture in `general_QC.pdf` represents the correlation heatmap after counting reads. The correlations are based on absolut read counts in consensus peaks.



DiffBind

The thrid picture in `general_QC.pdf` represents a PCA analysis after read counting and normalization.



The fourth picture in `general_QC.pdf` represents the correlation heatmap after counting reads and normalization. The correlations are based on normalized read counts in consensus peaks.

Not done in tutorial, hence no figure here

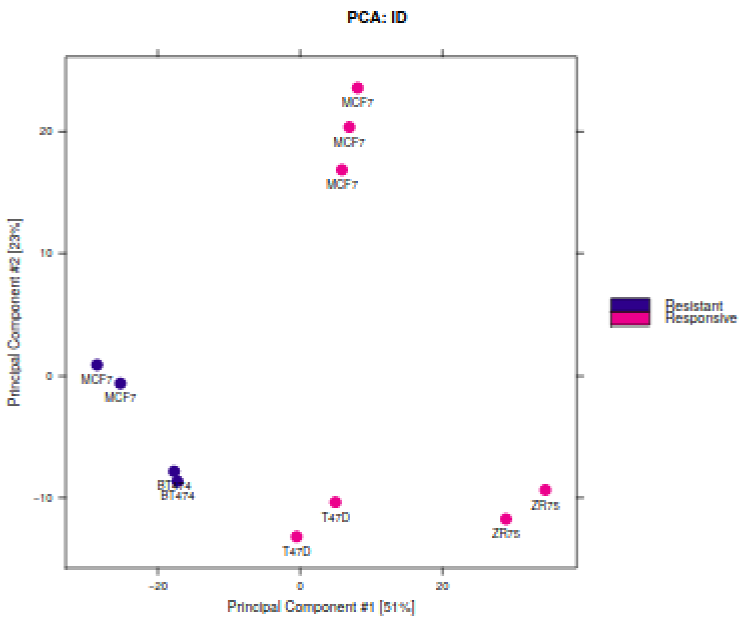
DiffBind

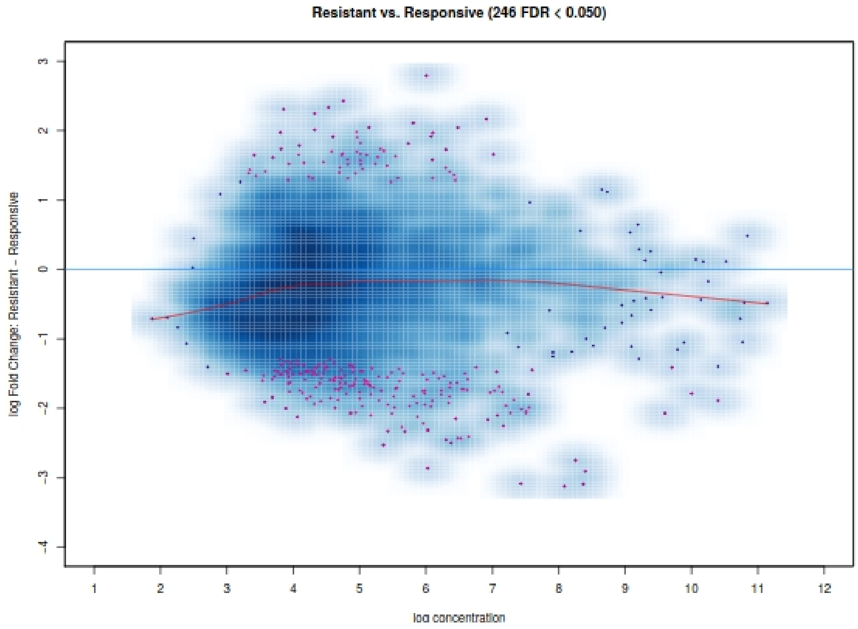
The next pictures come from pairwise camparisons `A.vs.B.pdf`.

The first pictures is of a PCA using only the differentially expressed peaks. There should be a clear separation of the sample on PC1.

The second picture is a standard MA plot.

The third picture plots the log concentrations of both conditions against each other. (not pictured here)



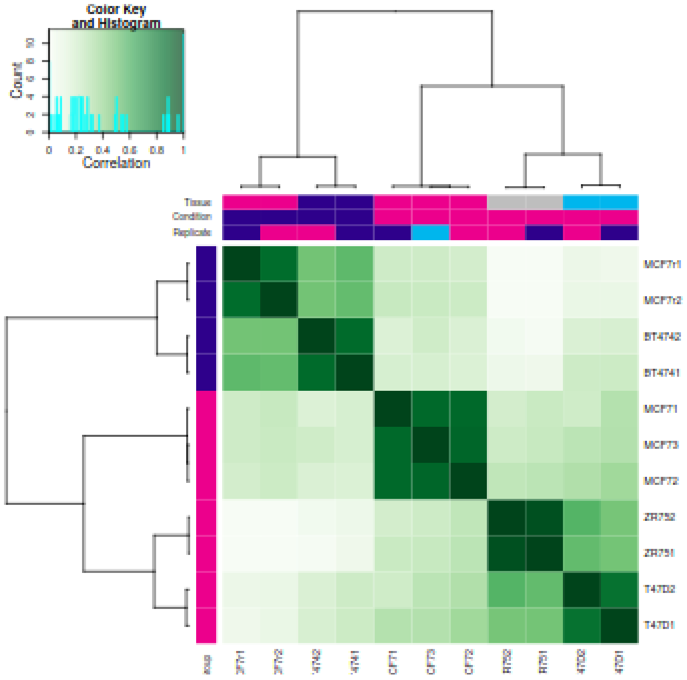


DiffBind

On page 4 is the correlation heatmap based only on the normalized expression of significantly different peaks.

From their tutorial:

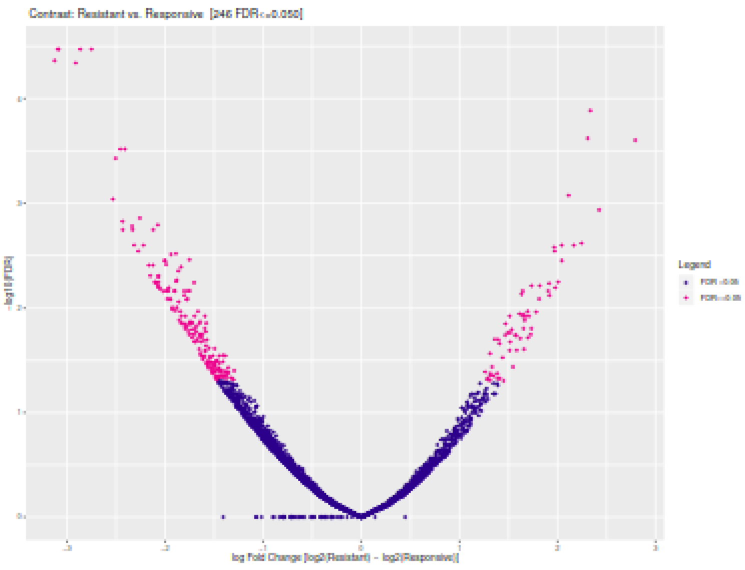
Note this is plot is not a "result" in the sense that the analysis is selecting for sites thatdiffer between the two conditions, and hence are expected to form clusters representing theconditions.



On page 5 is a standard volcano plot. These tend to look a little different to normal RNAseq or proteomics volcano plots.

From their tutorial:

The plot shows the predominance of lower binding in the Resistant case-evidenced by the greater number of significant sites on the negative side of the Fold Change(X) axis.

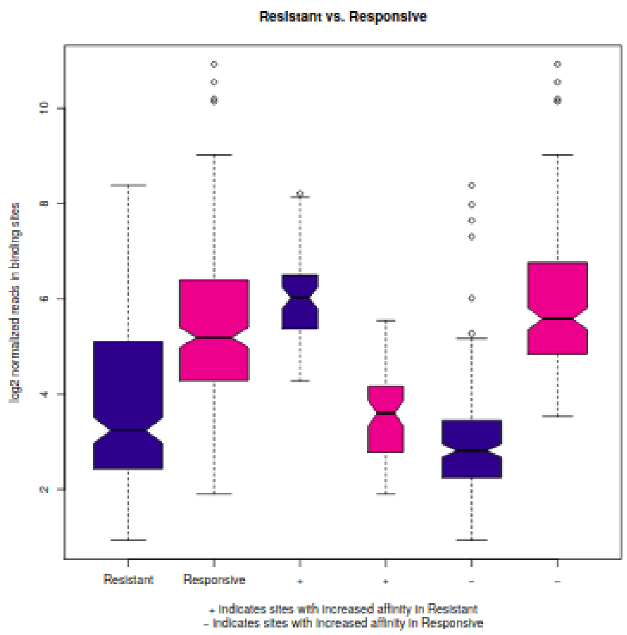


DiffBind

The plot on page 6 is a Box plot of read distributions for significantly differentially bound (DB) sites.

From their tutorial:

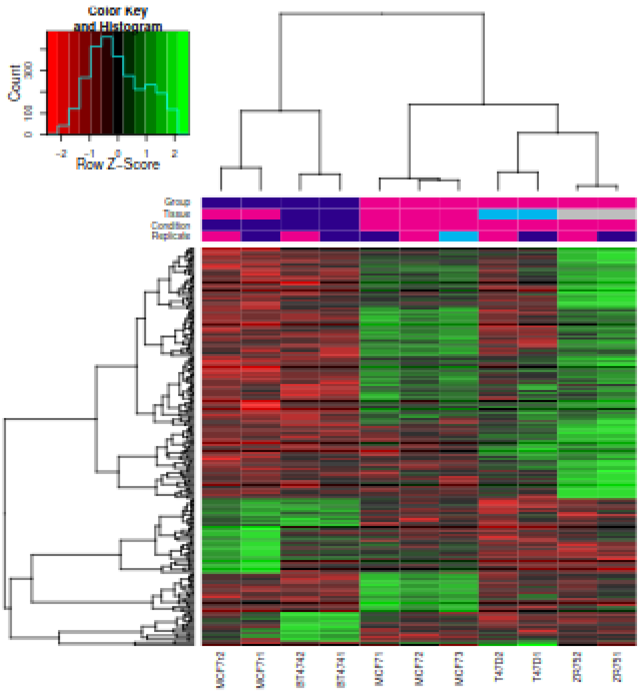
Tamoxifen Resistant samples are shown in blue, and Responsive samples are shown in red. Left two boxesshow distribution of reads over all DB sites in the Resistant and Responsive groups; middle two boxesshow distributions of reads in DB sites that increase in affinity in the Resistant group; last two boxesshow distributions of reads in DB sites that increase in affinity in the Responsive group.



On the last page there is a binding affinity heatmap showing affinities for differentially bound sites.

From their tutorial:

Samples cluster first by whether they are responsive to tamoxifen treatment, then by cell line,then by replicate. Clusters of binding sites show distinct patterns of affinity levels



ChIPseeker

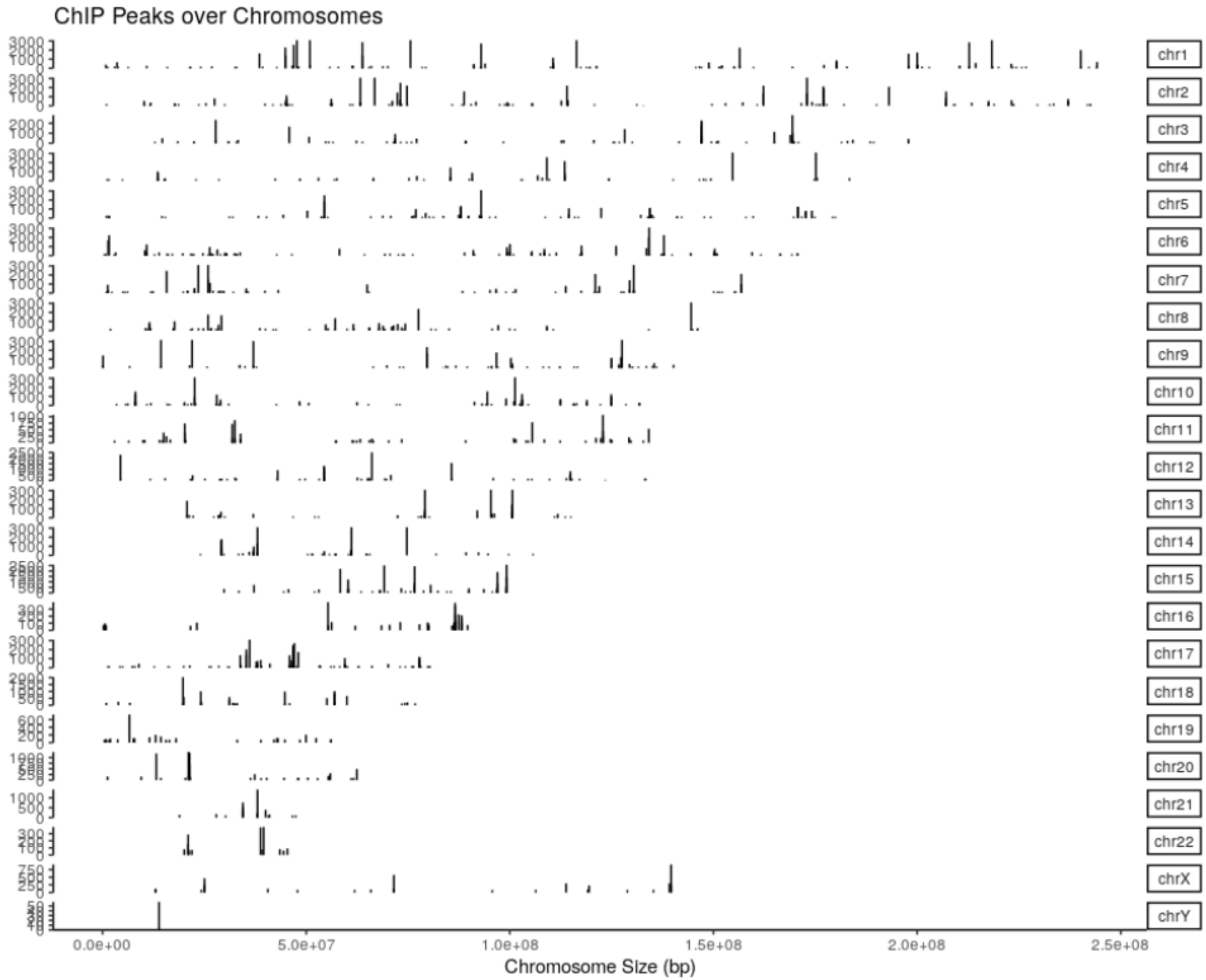
From their tutorial:

ChIPseeker is an R package for annotating ChIP-seq data analysis. It supports annotating ChIP peaks and provides functions to visualize ChIP peaks coverage over chromosomes and profiles of peaks binding to TSS regions. Comparison of ChIP peak profiles and annotation are also supported. Moreover, it supports evaluating significant overlap among ChIP-seq datasets. Currently, ChIPseeker contains 17,000 bed file information from GEO database. These datasets can be downloaded and compare with user’s own data to explore significant overlap datasets for inferring co-regulation or transcription factor complex for further investigation.

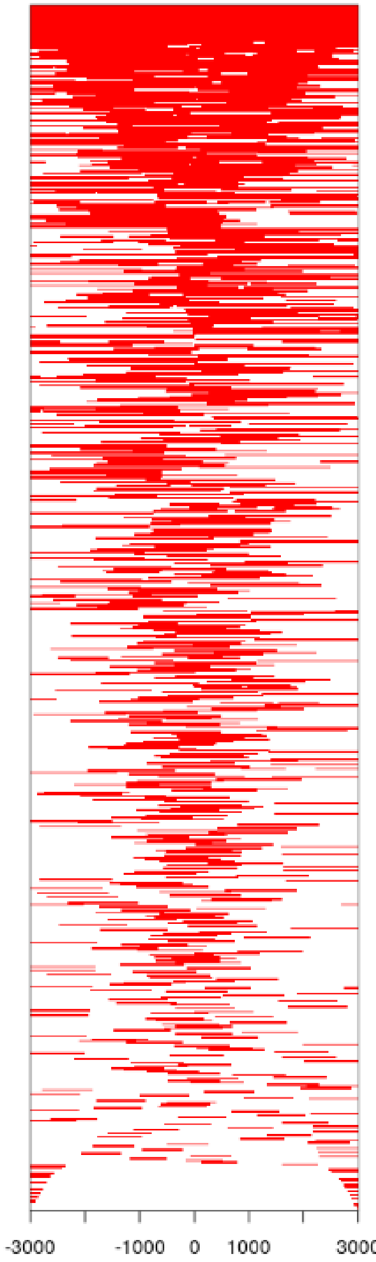
Here I will quickly go over the plots generated automatically. The script [top/tmp/peakAnno.R] can be extended for more custom analysis.

ChIPseeker

`covplot.A.vs.B.pdf` visualizes the location of all peaks along the chromosomes. The hight of each peak represents the average binding affinity ‘Conc’ calculated by DiffBind3.



ChIPseeker

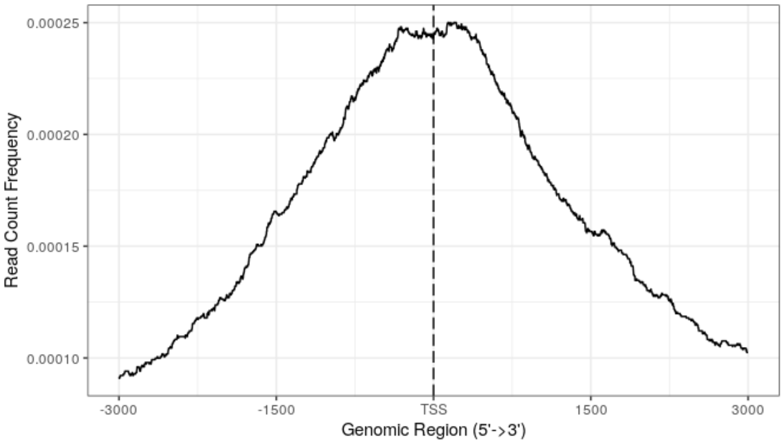


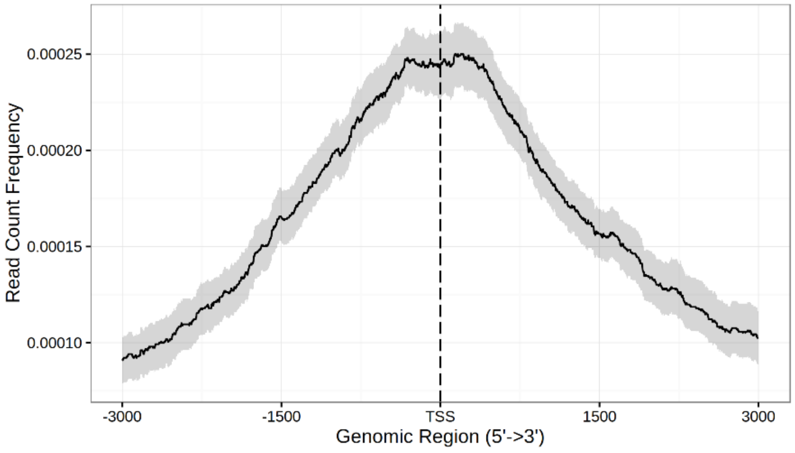
`tagMatrix.A.vs.B.pdf` visualizes the peak density around a TSS.

In our pipeline we look at the region from -2000 to + 2000 around the TSS.

In the first image each line represents a peak. The data is automatically sorted in such a way that it looks like a reverse volcono.

The second and third image show the average signal distribution of reads around the TSS region including the confidence interval.





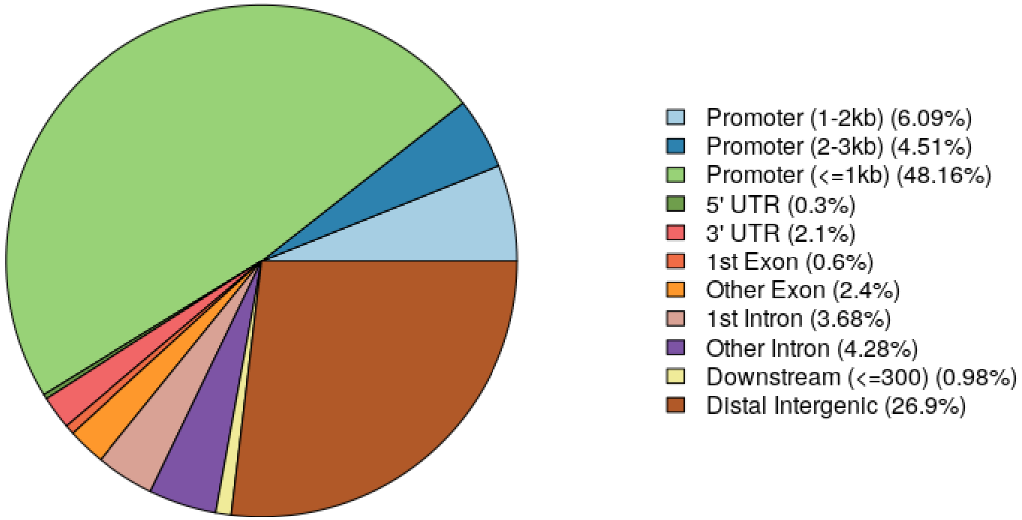
ChIPseeker

`annotated.A.vs.B.pdf` visualizes distribution of peak annotation to different gene regions amd the disribuion of TF binding loci relative to the TSS

From their tutorial:

Visualize Genomic Annotation

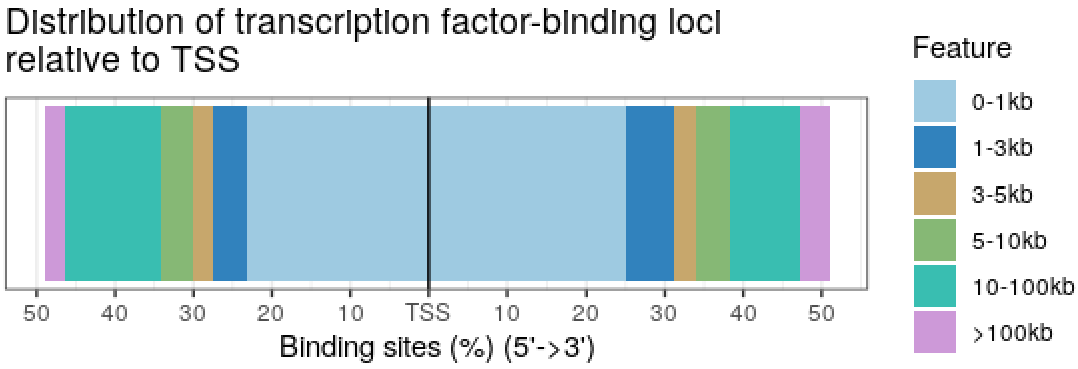
To annotate the location of a given peak in terms of genomic features, annotatePeak assigns peaks to genomic annotation in “annotation” column of the output, which includes whether a peak is in the TSS, Exon, 5’ UTR, 3’ UTR, Intronic or Intergenic. Many researchers are very interesting in these annotations. TSS region can be defined by user and annotatePeak output in details of which exon/intron of which genes as illustrated in previous section.



From their tutorial:

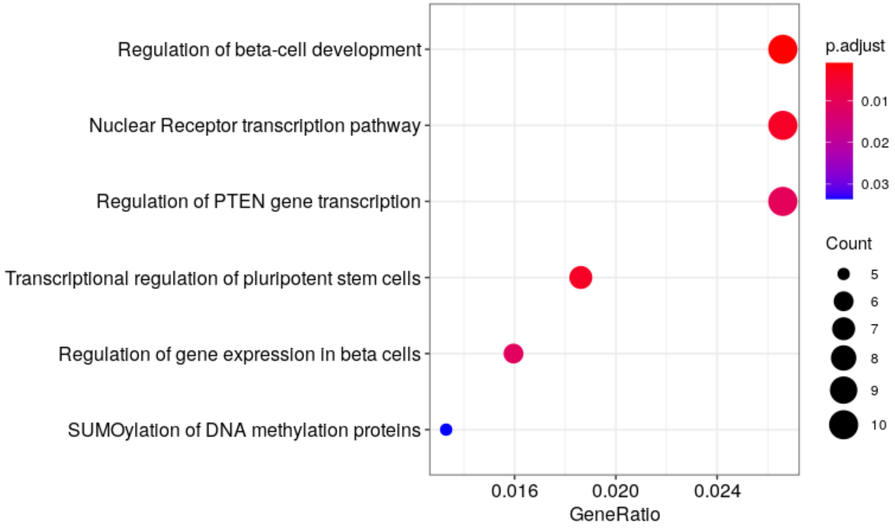
Visualize distribution of TF-binding loci relative to TSS

The distance from the peak (binding site) to the TSS of the nearest gene is calculated by annotatePeak and reported in the output. We provide plotDistToTSS to calculate the percentage of binding sites upstream and downstream from the TSS of the nearest genes, and visualize the distribution.



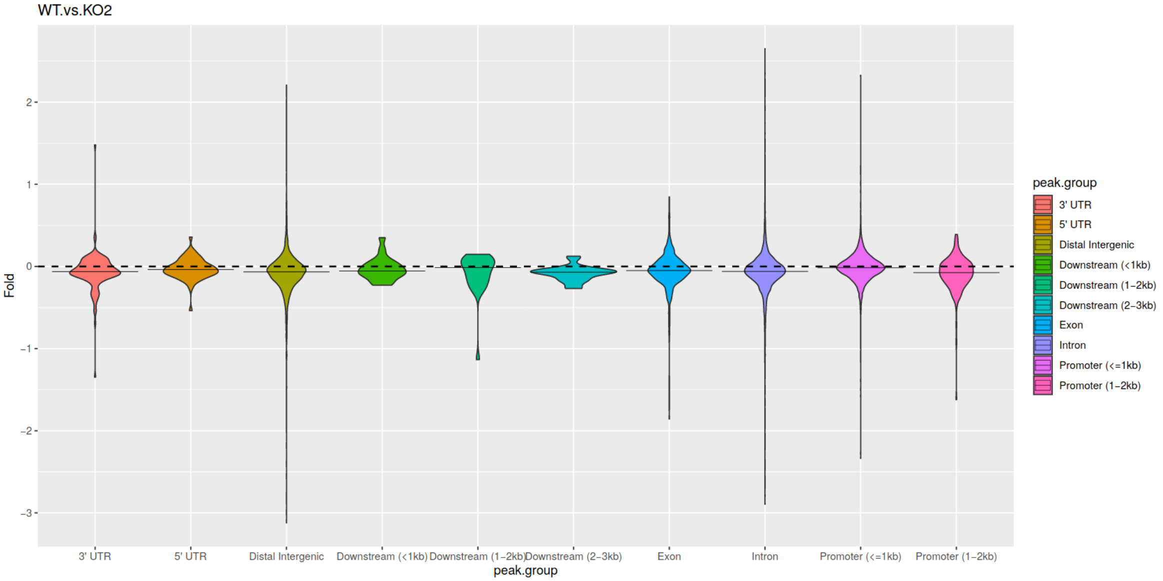
ChIPseeker

Using the `ReactomePA` library a functional enrichment of all peaks with a gene association is performed. The result is visualized in `peak_functional_enrichment.A.vs.B.pdf.` Please note that the automatic pipeline does not perform a functional annotation of differential peaks as they have not lead to any results in test projects.



Although the `ChIPseeker` tutorial continues with valuable analysis, the automation is stopped with this last plot as all further steps are all custom analysis for specific set ups.

In this last picture `feature_foldChange_violin.A.vs.B.pdf`, peaks are binned according to their annotation. Violins are drawn using the log2fold change values in order to identify an enrichment or depletion of binding afinity in a specif gene region



Reference

- Flexbar
  - https://pubmed.ncbi.nlm.nih.gov/28541403/
- Bowtie2
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322381/
- Picard
  - “Picard Toolkit.” 2019. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute
- ATACseqQC
  - https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4559-3
  - https://bioconductor.org/packages/release/bioc/vignettes/ATACseqQC/inst/doc/ATACseqQC.html
- MACS2
  - https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html
  - https://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-9-r137
- DiffBind
  - https://www.nature.com/articles/nature10730
  - https://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf
- ChIPseeker
  - https://academic.oup.com/bioinformatics/article/31/14/2382/255379
  - https://www.bioconductor.org/packages/release/bioc/vignettes/ChIPseeker/inst/doc/ChIPseeker.html
- Resource for ATACseq in general:
  - https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1929-3