Downstream Analysis

Shoko Hirosue
MRC Cancer Unit, University of Cambridge

CRUK CI Bioinformatics Summer School July 2020

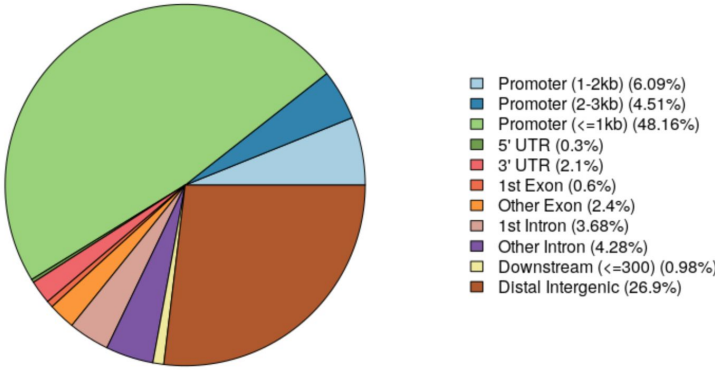
What can we do with ChIP seq?

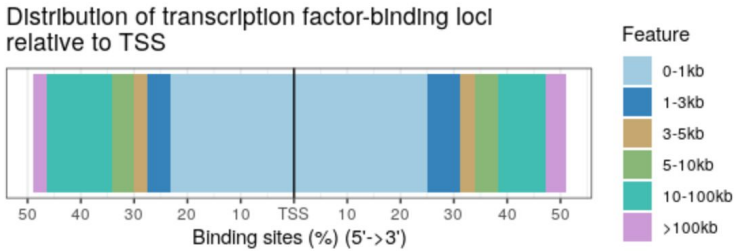
1. Annotation of genomic features to peaks
2. Functional enrichment analysis: Ontologies, Gene Sets, Pathways
3. Normalization and Visualization
4. Motif identification and Motif Enrichment Analysis

1. Annotation of Genomic Features to Peaks

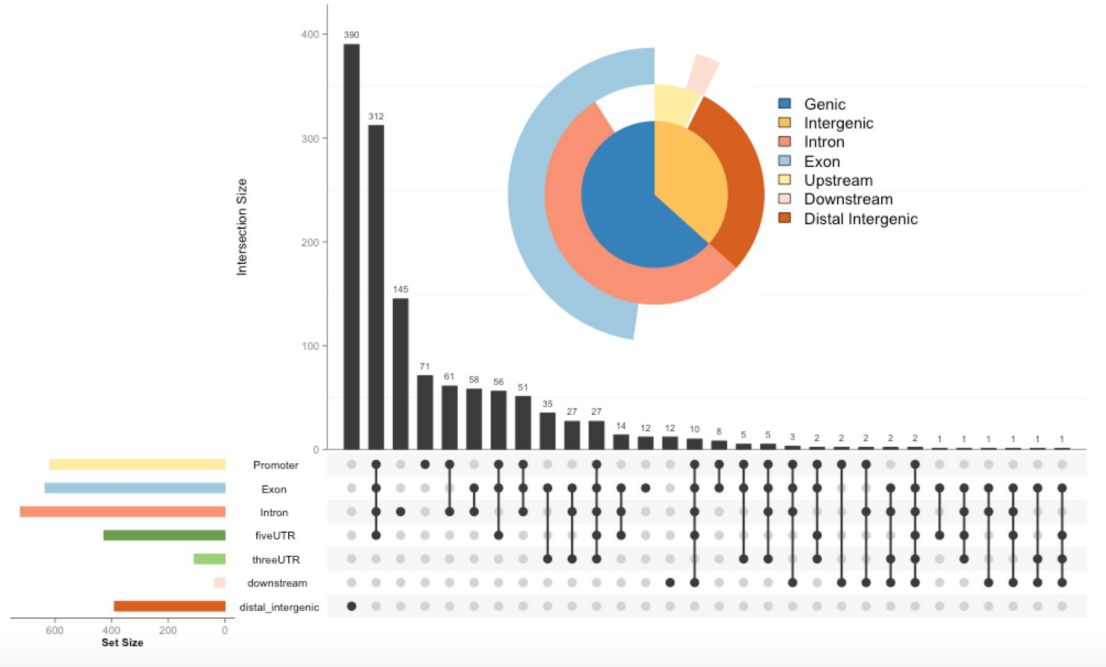
1. Annotation of genomic features to peaks

ChIPSeeker





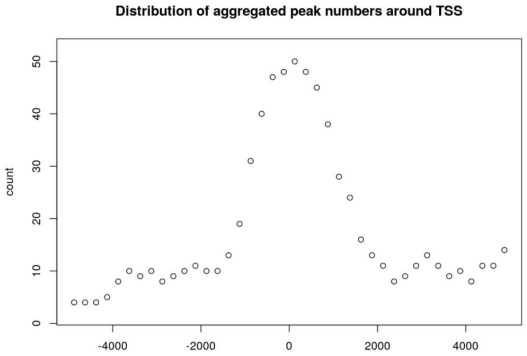
Distribution of Binding Sites

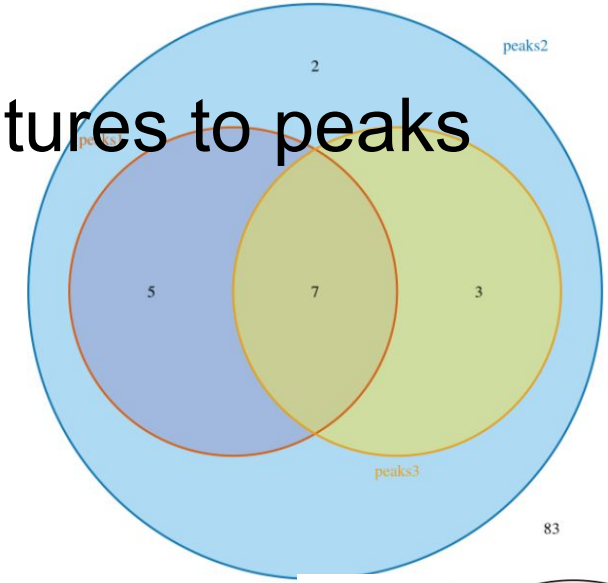


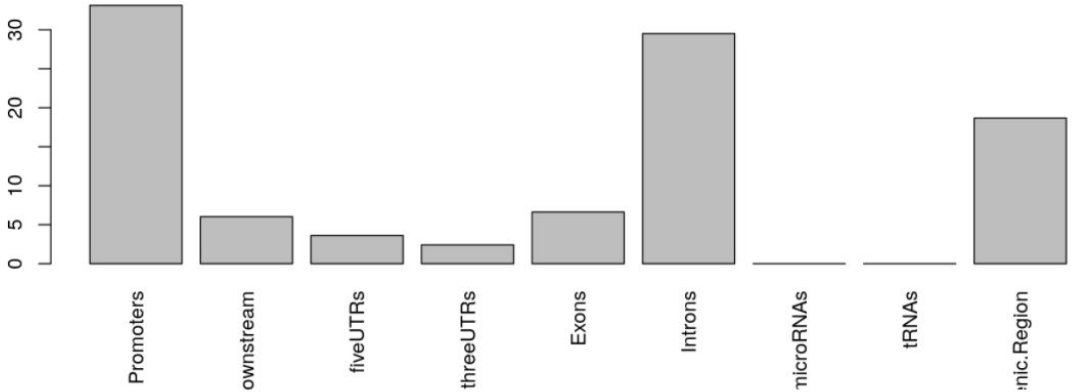
Yu et al., 2015, Bioinformatics

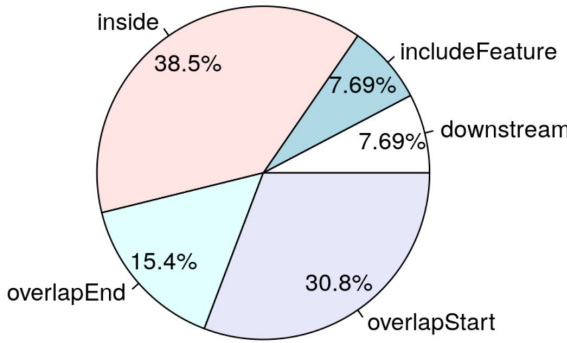
1. Annotation of genomic features to peaks

ChIPSeqAnno





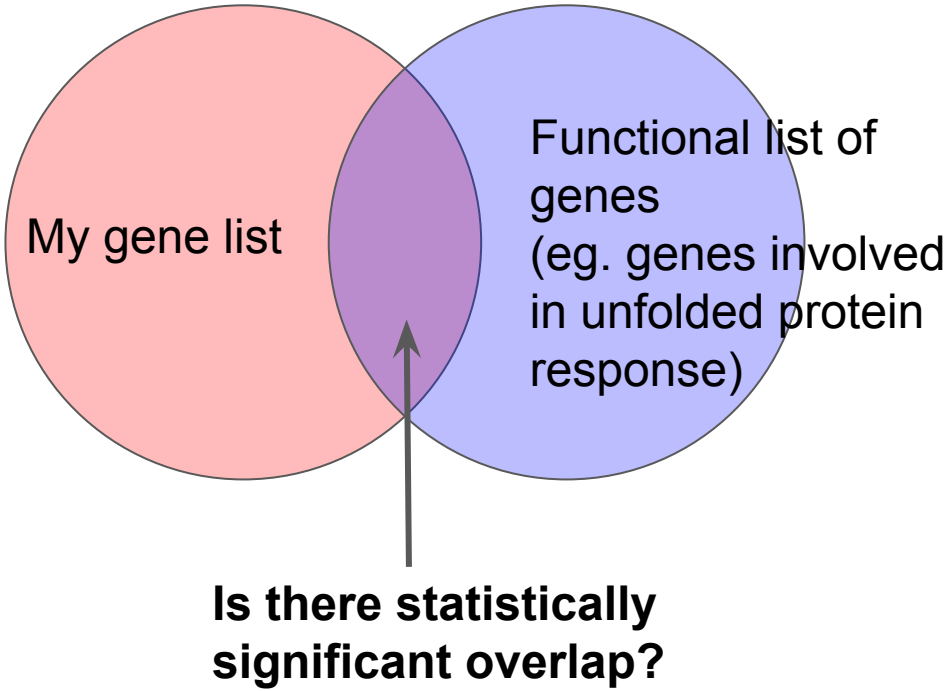




Zhu et al. 2010. *BMC Bioinformatics*

2. Functional Enrichment Analysis

2. Functional enrichment analysis



Databases of functional list of genes

- GO
- KEGG
- Reactome
- ...

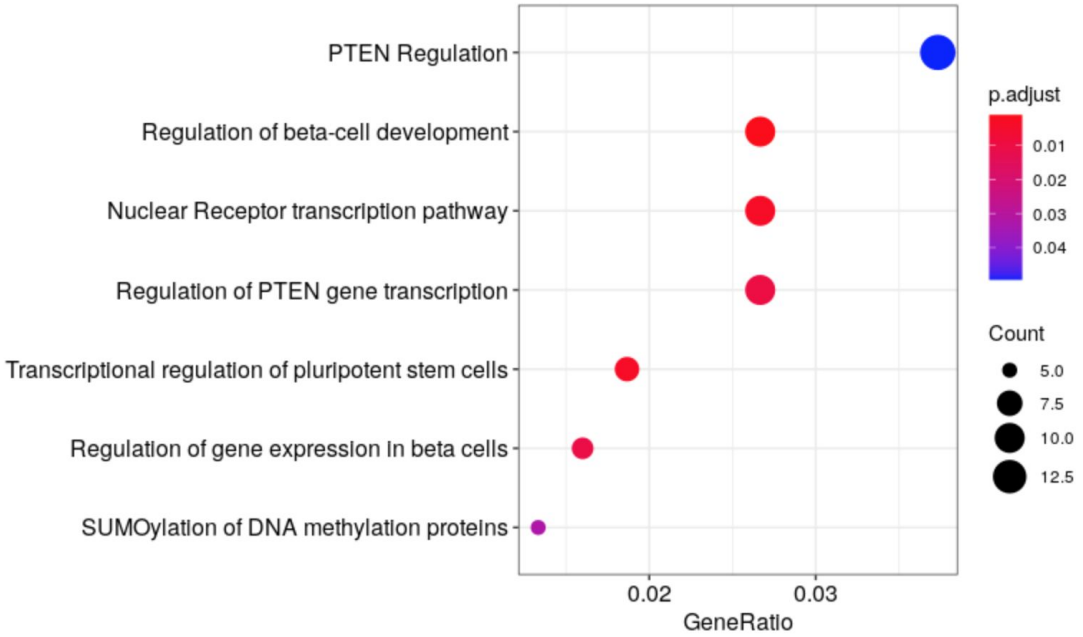
2. Functional enrichment analysis

ChIPSeeker

ClusterProfiler (GO, KEGG)

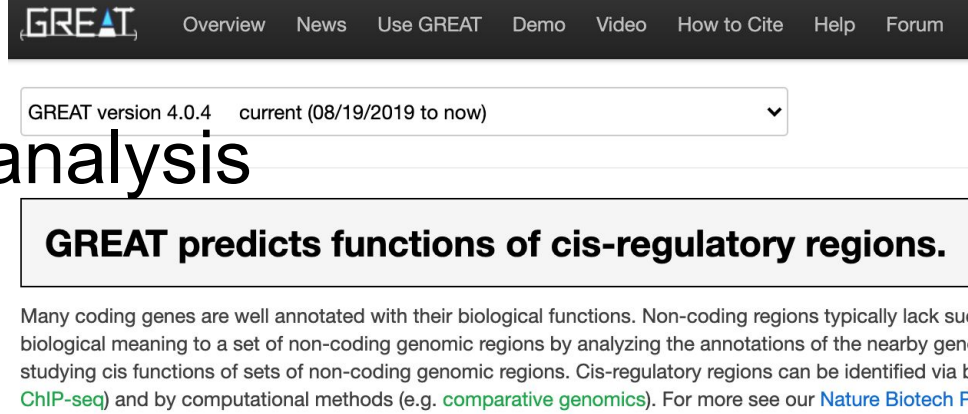
DOSE (Disease Ontology)

ReactomePA (Reactome)



Yu et al., 2015, Bioinformatics

2. Functional enrichment analysis



GREAT (http://great.stanford.edu/public/html/)

- Widely used web based tools
- Associates genomic regions with genes by defining a ‘regulatory domain’ for each gene in the genome.
  - 5 kb upstream and 1 kb downstream from its transcription start site (denoted below as 5+1 kb)
  - an extension up to the basal regulatory domain of the nearest upstream and downstream genes within 1 Mb (user can modify the length)
  - refine the regulatory domains of a handful of genes, including several global control regions20, by using their experimentally determined regulatory domains
- Incorporates annotations from 20 ontologies and is available as a web application

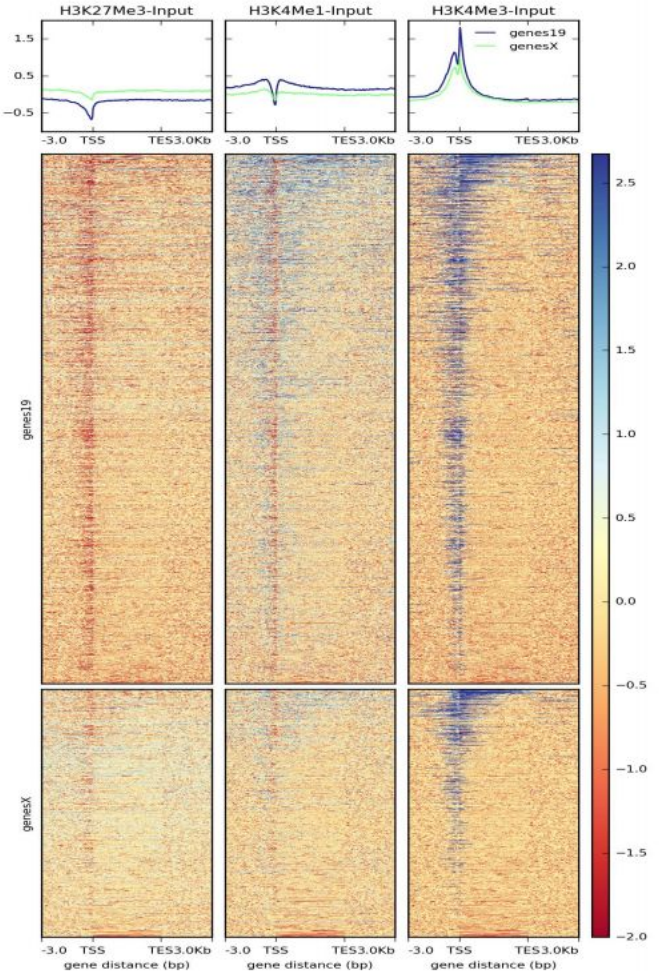
McLean et al. 2010, Nat Biotech

3. Normalization and Visualization

3. Normalization and visualization

Deeptools

- Plot signal profiles
- Customized heat-maps
- PCA, correlation and fingerprint plots (chip enrichment)

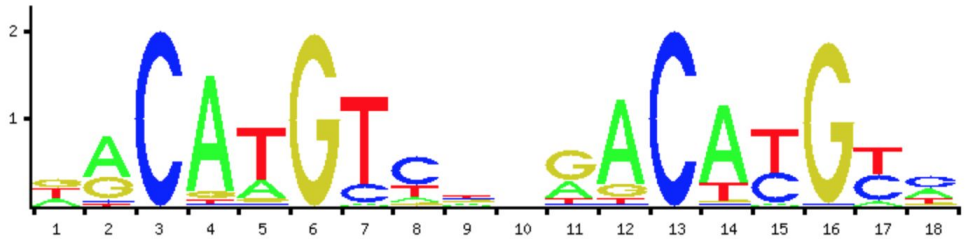


Ramírez et al., 2016, Nucleic Acids Res.

4. Motif Analysis

Motifs are genomic sequences that specifically bind to transcription factors.

There are many possible bases at certain positions in the motif, whereas other positions have a fixed base.



Sequence logo diagram for TP73. The height of the letter represents the frequency of the nucleotide observed.

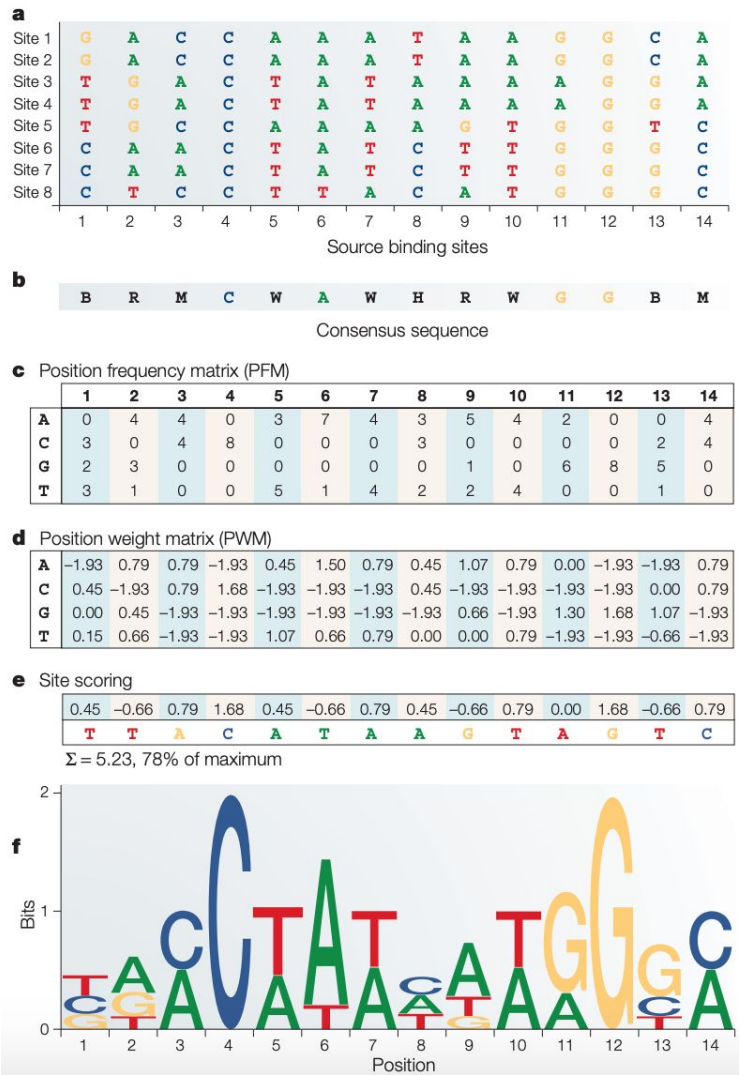
4. Motif Analysis

There are many other formats (eg. c, d, e of the right figure) to show the motif information (eg. **PWM**)

TFBS databases

- JASPAR
- TRANSFAC
- Swissregulon
- HOCOMOCO
- HOMER

Wasserman & Sandelin, 2004, Nat Rev Genet.



4. Motif Analysis

Two different ways of motif detection in sequences

1. Known Transcription Factor Binding Sites (TFBS) detection - Use prior information about TF binding motifs (PWMs)
2. De novo motif identification – Pattern discovery methods

Adapted from Shamith Samarajiwa’s slides

4. Motif Analysis

Motif Enrichment Analysis

- Identifies over and under-represented known motifs in a set of regions
- -> background is required.
- Picking the right background model will determine the success of the motif enrichment analysis:
  - All promoters from protein coding genes
  - Open chromatin regions

Adapted from Shamith Samarajiwa’s slides

4. Motif Analysis

Motif Enrichment Analysis

- Identifies over and under-represented known motifs in a set of regions
- -> background is required.
- Picking the right background model will determine the success of the motif enrichment analysis:
  - All promoters from protein coding genes
  - Open chromatin regions
  - Shuffled test sequence set
  - A sequence set similar in nucleotide composition, length and number to the test set
  - Higher order Markov model based backgrounds

Adapted from Shamith Samarajiwa’s slides

4. Motif Analysis

HOMER (http://homer.ucsd.edu/homer/)

- Perform both known TFBS detection and de-novo motif identification
- Motif Enrichment analysis
- If you do not give background regions, the background sequences will be randomly selected from the genome, matched for GC% content

- findMotifs.pl discover motifs in promoter
- findMotifsGenome.pl discover motifs in genomic regions

Heinz et al. Mol Cell, 2010

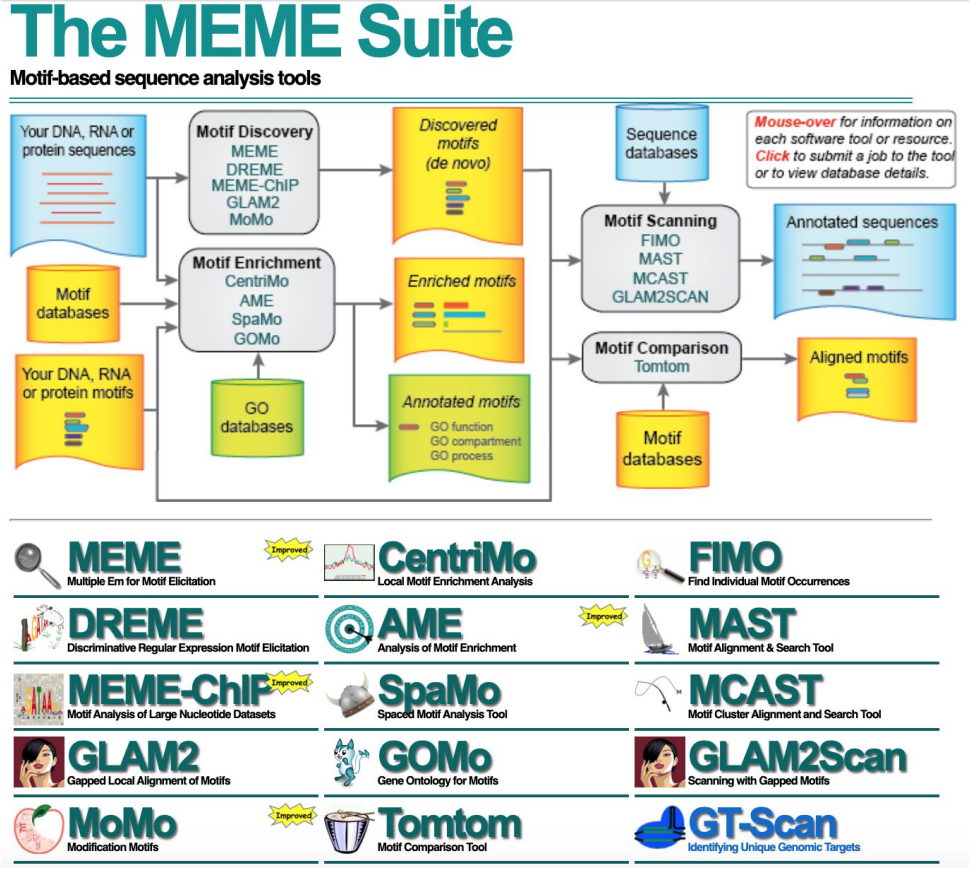
4. Motif Analysis

MEME Suite
(http://meme-suite.org/)

Given a set of genomic regions, it performs

- De-novo motif identification (MEME, DREME)
- Compare identified motifs to known motifs (TOMTOM)

- Known TFBS detection (Centrimo, AME)



4. Motif Analysis

Limitations

”Futility Theorem” of motif finding

Extremely high false positive rate in TFBSs (Transcription Factor Binding Sites) prediction, as the methods detect potential binding sites, NOT NECESSARILY those of **functional importance**

Wasserman and Sandelin, 2004, Nat Rev Genet

References

- CRUK summer school 2019 materials (https://bioinformatics-core-shared-training.github.io/cruk-summer-school-2019/)
- Yu et al., 2015, Bioinformatics. “ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization”
- Zhu et al. 2010. “ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data.” *BMC Bioinformatics*
- McLean et al. 2010, Nat Biotech. “GREAT improves functional interpretation of *cis*-regulatory regions”
- Ramírez et al., 2016, Nucleic Acids Res. “deepTools2: a next generation web server for deep-sequencing data analysis”
- Wasserman & Sandelin, 2004, Nat Rev Genet. “Applied bioinformatics for the identification of regulatory elements”
- Heinz et al. Mol Cell, 2010. “Simple Combinations of Lineage-Determining Transcription Factors Prime *cis*-Regulatory Elements Required for Macrophage and B Cell Identities”
- Bailey et al. 2009, *Nucleic Acids Research.* "MEME SUITE: tools for motif discovery and searching"