

L8: Downstream analysis of ChIP-seq and ATAC-seq data

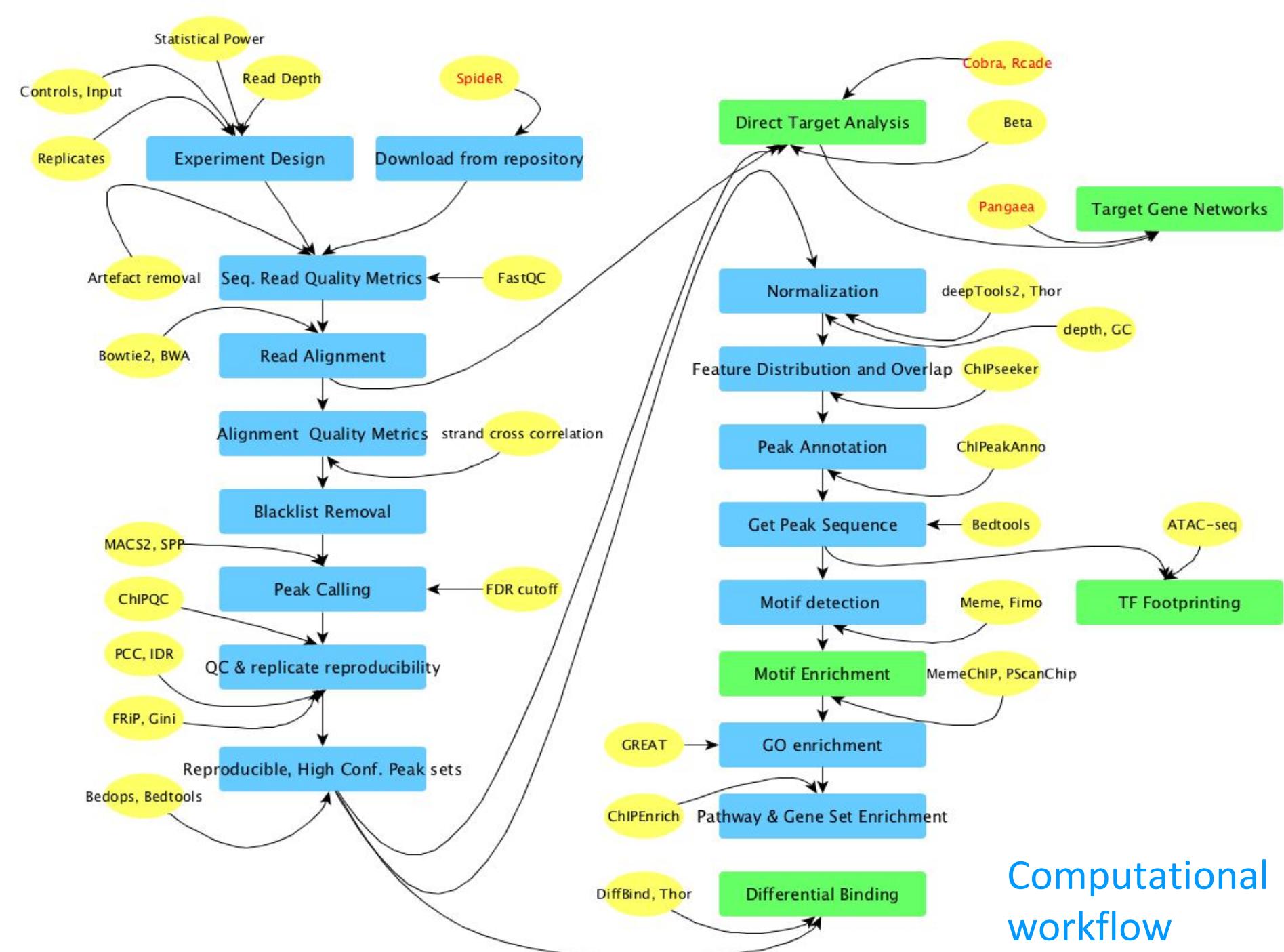
Shamith Samarajiwa
CRUK Bioinformatics Autumn School
September 2017



Summary

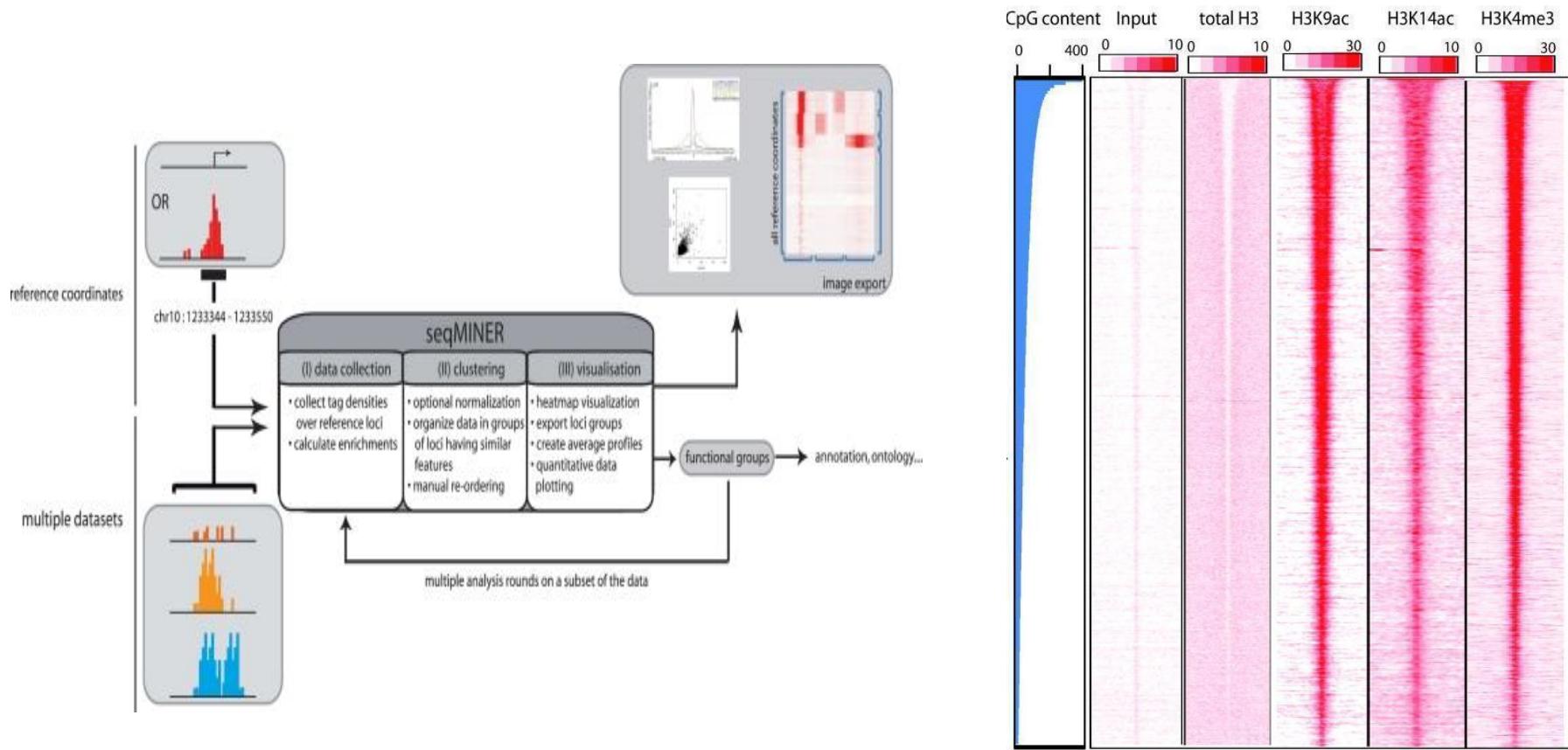
Downstream analysis for extracting meaningful biology :

- Normalization and Visualization
- Annotation of genomic features to peaks
- Feature distribution of binding sites
- Feature overlap analysis
- Functional enrichment analysis: Ontologies, Gene Sets, Pathways
- Motif identification and Motif Enrichment Analysis
- Differential binding analysis
- Integration with transcriptomic data to Identify direct targets
- Network Biology applications



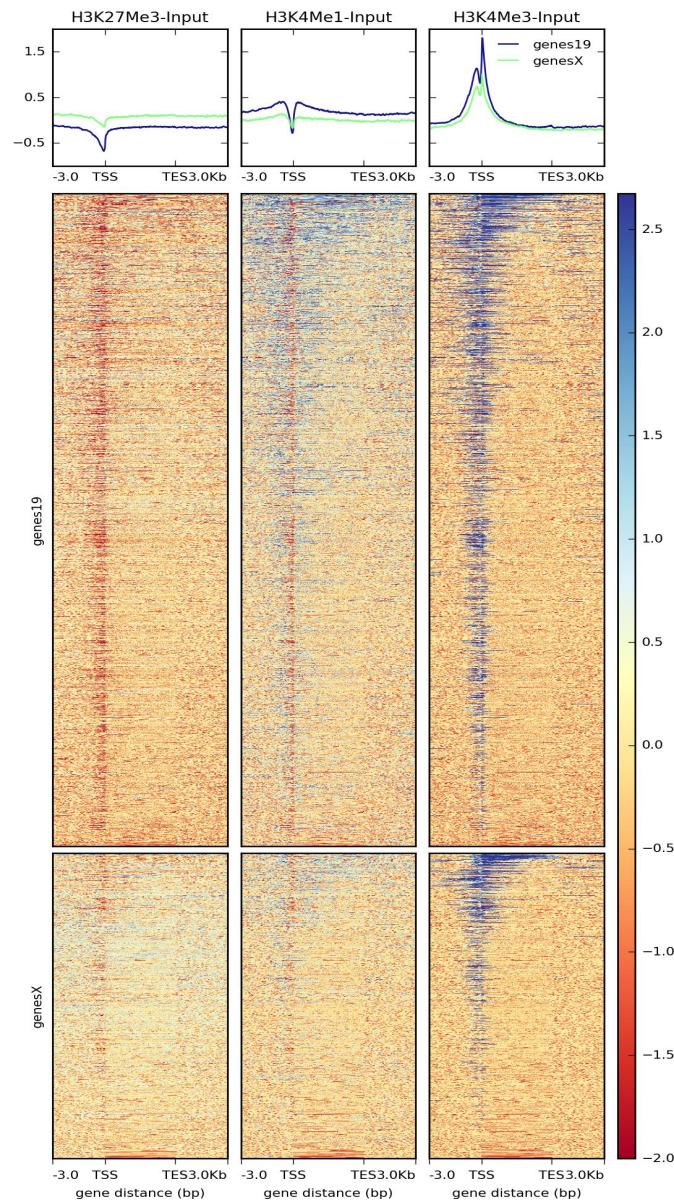
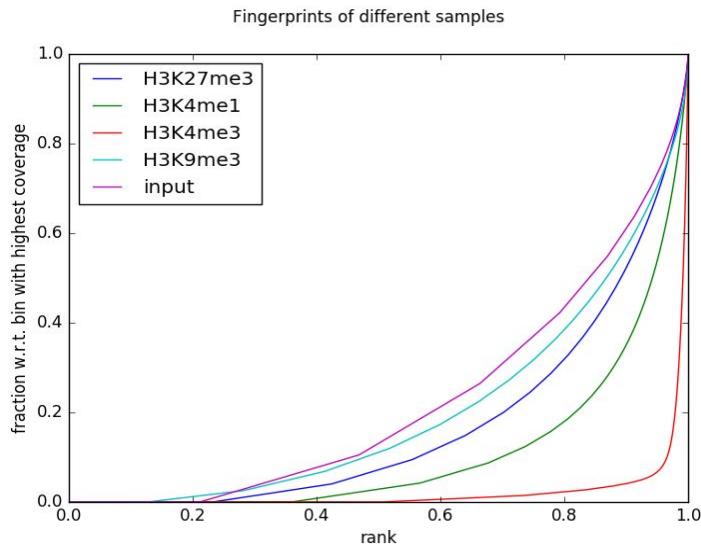
Compare, Normalize & Visualize 1

- **seqMiner** enables qualitative comparisons between a reference set of genomic positions and multiple ChIP-seq data-sets.
- Useful for comparing and visualizing replicates or conditions.



Compare, Normalize & Visualize 2

- **deepTools2** sequence depth or input normalization, GC bias correction
- Plot signal profiles
- Customized heat-maps
- PCA, correlation and fingerprint plots (chip enrichment)



Peak annotation 1

- **ChIPpeakAnno (BioC)** map peaks to nearest feature (TSS, gene, exon, miRNA or custom features)
 - extract peak sequences
 - find peaks with bidirectional promoters
 - obtain enriched gene ontology
 - map different annotation and gene identifiers to peaks
- Use **biomaRt** package to get annotation from Ensembl.
- **IRanges, GenomicFeatures, GO.db, BSgenomes, multtest (BioC)**
- converts BED and GFF data formats to *RangedData* object before calling *peak annotate* function.

Peak annotation 2

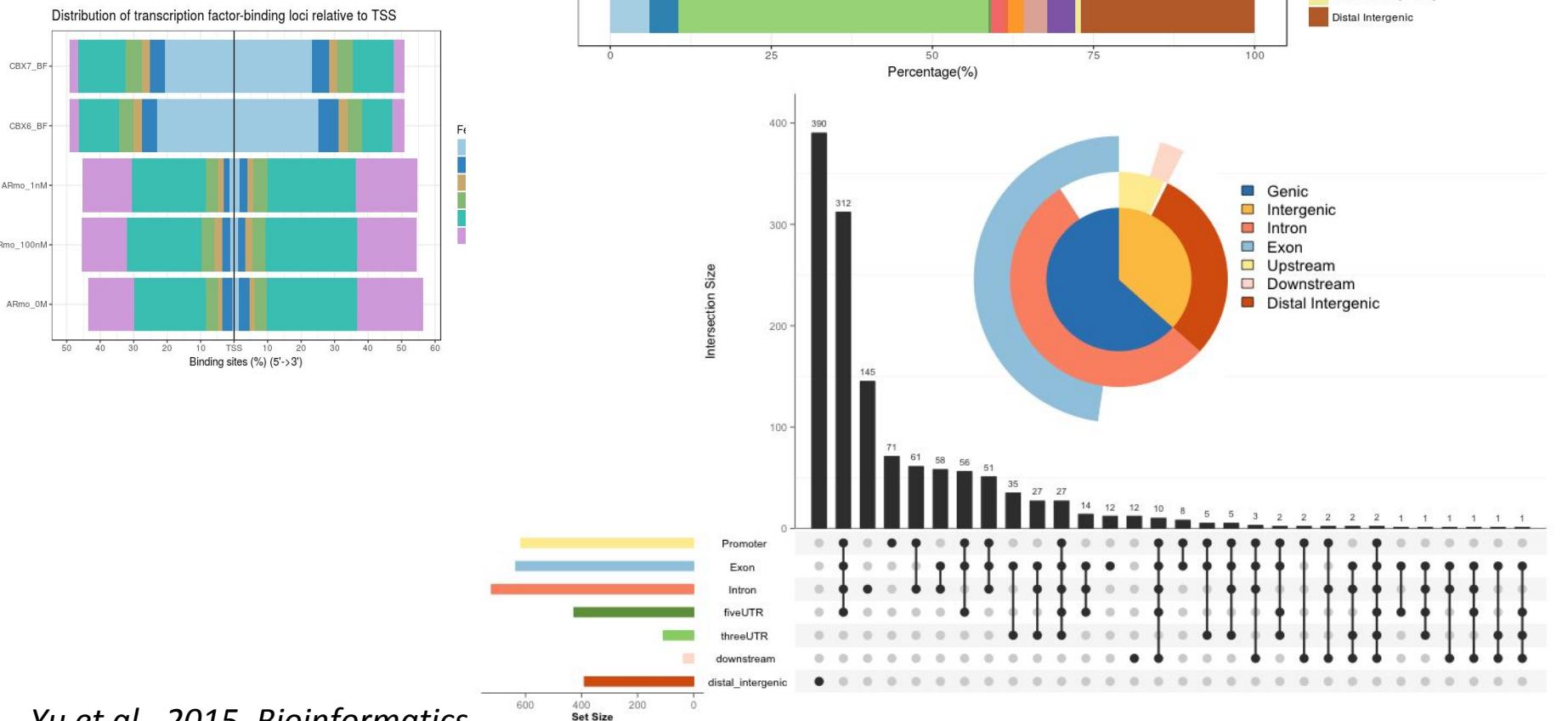
PeakAnalyzer

- A set of high-performance utilities for the automated processing of experimentally-derived peak regions and annotation of genomic loci.
- Consists of PeakSplitter and PeakAnnotator.
- Biologist' friendly tool.
- Get latest genome annotation files from Ensembl (gtf format) or UCSC (BED format).
- Map to either nearest downstream gene, TSS or user defined annotation.
- Determine overlap between peak sets.
- Split peaks to sub-peaks. May be useful for *de novo* motif analysis.

Salmon-Divon et al., 2010, BMC Bioinformatics.

Peaks distribution across features

ChIPseeker (BioC)



Functional Enrichment Analysis 1

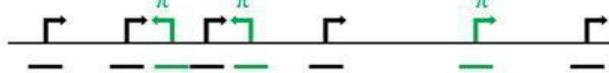
GREAT & rGREAT: Genomic Regions Enrichment of Annotations Tool

a

Hypergeometric test over genes

Step 1: Infer proximal gene regulatory domains

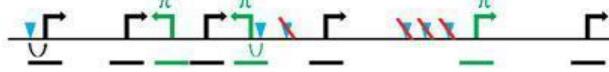
- ↷ Gene transcription start site
- π Ontology annotation (e.g., "actin cytoskeleton")
- = Proximal regulatory domain of gene with/without π



Step 2:

Associate genomic regions with genes via regulatory domains

- ↙ Genomic region associated with nearby gene
- ✗ Ignored distal genomic region



Step 3:

Count genes selected by proximal genomic regions

2 genes selected by proximal genomic regions
1 gene selected carries annotation π

Step 4: Perform hypergeometric test over genes

$$N = 8 \text{ genes in genome}$$

$$K_\pi = 3 \text{ genes in genome carry annotation } \pi$$

$$n = 2 \text{ genes selected by proximal genomic regions}$$

$$k_\pi = 1 \text{ gene selected carries annotation } \pi$$

$$P = \Pr_{\text{hyper}}(k \geq 1 \mid N = 8, K = 3, n = 2)$$

b

Binomial test over genomic regions

Step 1:

Infer distal gene regulatory domains

- ↷ Gene transcription start site
- π Ontology annotation (e.g., "actin cytoskeleton")
- = Distal regulatory domain of gene with/without π



Step 2:

Calculate annotated fraction of genome

0.6 of genome is annotated with π

Step 3:

Count genomic regions associated with the annotation

- ↙ Genomic region



5 genomic regions hit annotation π

Step 4: Perform binomial test over genomic regions

$$n = 6 \text{ total genomic regions}$$

$$p_\pi = 0.6 \text{ fraction of genome annotated with } \pi$$

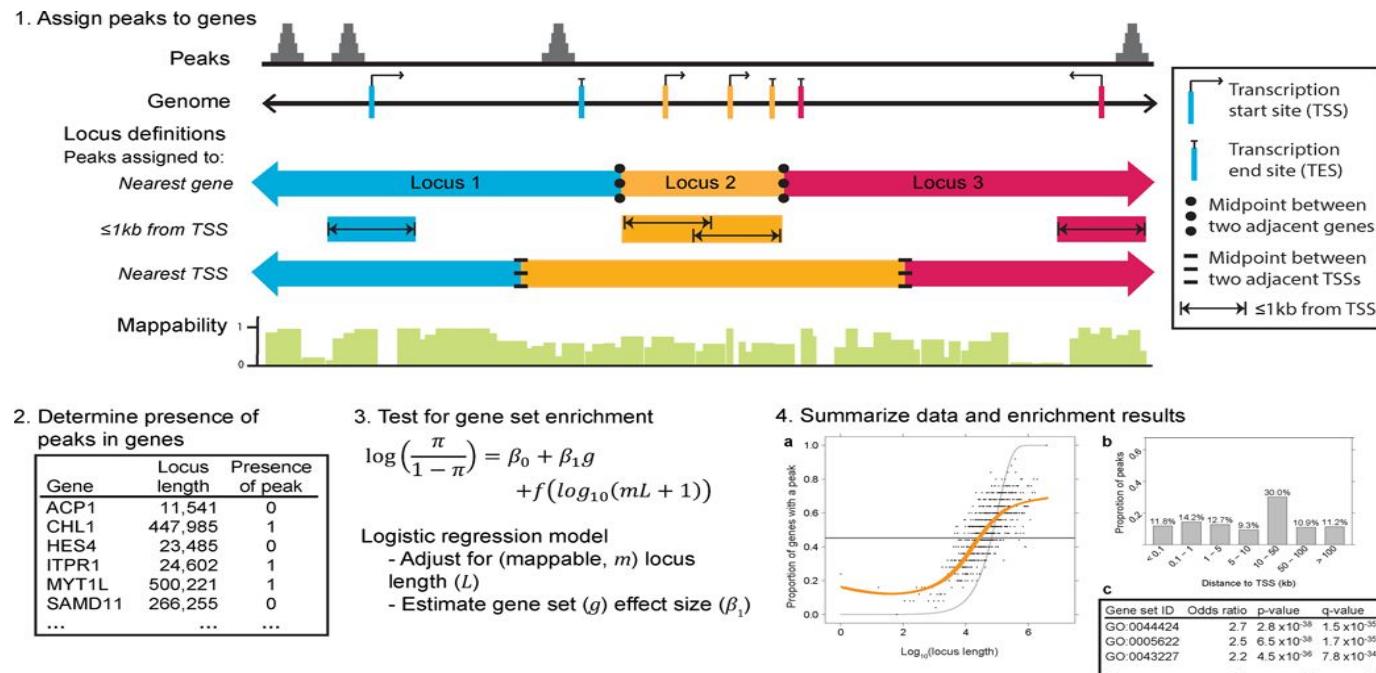
$$k_\pi = 5 \text{ genomic regions hit annotation } \pi$$

$$P = \Pr_{\text{binom}}(k \geq 5 \mid n = 6, p = 0.6)$$

Functional Enrichment Analysis 2

chipenrich

- Includes 3 different enrichment methods:
 - Broadenrich - broadpeaks or histone modifications
 - Chipenrich -TF narrow peaks 1000-10000's
 - Polyenrich -TF >100,000
- Includes annotation, and can use custom user provided annotation



Motif detection

- Don't scan a sequence with a motif and expect all sites identified to be biologically active. Random matches will swamp the biologically relevant matches! This is a well known problem in motif searching, amusingly called the "**Futility Theorem**" of motif finding. *Wasserman & Sandelin, 2004, Nat Rev Genet.*
- 1. **PWM based sequence scanning** or word search methods. These methods uses prior information about TF binding sites and therefore can only be used to detect known Transcription Factor Binding Sites (TFBS).
- 2. ***De novo*** motif identification – Pattern discovery methods:
- **Word based** – Occurrence of each ‘word’ of nucleotides of a certain length is counted and compared to a background distribution.
- **Probabilistic**- seek the most overrepresented pattern using algorithmic approaches like Gibbs sampling and Expectation maximization. These iteratively evolve an initial random pattern until a more specific one is found.
- Use ***de novo*** motif calling and alignment to build your own PWMs!
- **Biostrings & Motiv** packages have PFM to PWM conversion methods.

BioConductor motif analysis packages

- [rGADEM](#) -motif discovery
- [MotifRG](#) -motif discovery
- [MotIV](#) -map motif to known TFBS, visualize logos
- [motifStack](#) -plot sequence logos
- [MotifDb](#) -motif database
- [PWMenrich](#) -motif enrichment analysis
- [TFBSTools](#) – R interface to the JASPAR database

Position Weight Matrices

a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	G	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

b

B R M C W A W H R W G G B M

Consensus sequence

PWM conversion:

$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)}$$

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

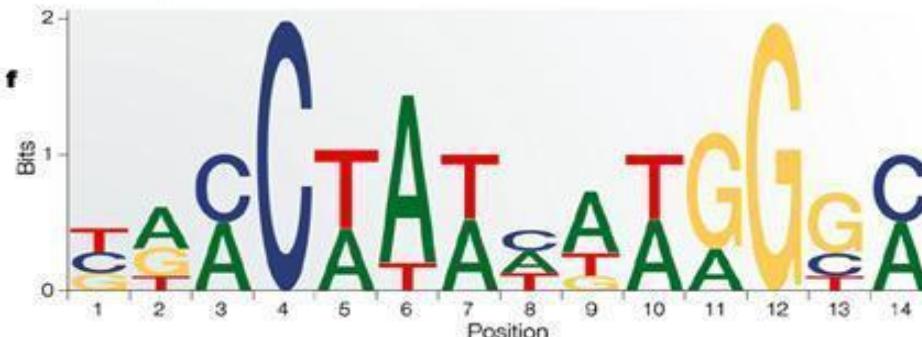
d Position weight matrix (PWM)

A	-1.93	0.79	0.79	-1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	-1.93	-1.93	0.79
C	0.45	-1.93	0.79	1.68	-1.93	-1.93	-1.93	0.45	-1.93	-1.93	-1.93	-1.93	0.00	0.79
G	0.00	0.45	-1.93	-1.93	-1.93	-1.93	-1.93	-1.93	0.66	-1.93	1.30	1.68	1.07	-1.93
T	0.15	0.66	-1.93	-1.93	1.07	0.66	0.79	0.00	0.00	0.79	-1.93	-1.93	-0.66	-1.93

e Site scoring

0.45	-0.66	0.79	1.68	0.45	-0.66	0.79	0.45	-0.66	0.79	0.00	1.68	-0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

$\Sigma = 5.23$, 78% of maximum



TFBS PWM/PFM sources

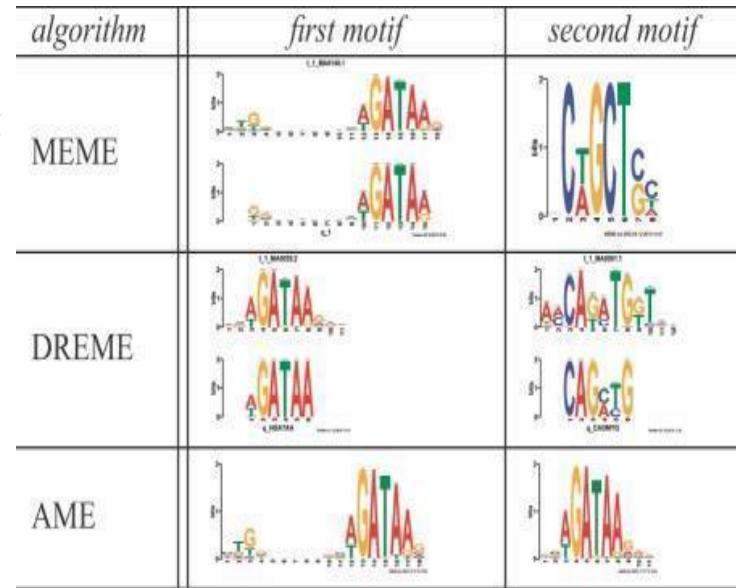
TRANSFAC public	Matys et al., 2006	Multiple species	v7.0 2005, Not been updated for a while!
TRANSFAC professional	Matys et al., 2006	Multiple species	v2017
JASPAR 2014	Mathelier et al., 2014	Multiple species	(656)
ORegAnno		Multiple species	Curated collection from different sources.
hPDI	Xie et al., 2010	Human	(437)
SwissRegulon	Pachkov et al., 2010	mammalian	(190)
HOMER	Heinz et al., 2010	Human	(1865)
UniPROBE	Newburger & Bulyk, 2009	Multiple species	
Dimers	Jonawski et al., 2013	Human	(603) predicted dimers
FactorBook	Wang et al., 2012	Human	(79) ENCODE ChIP-seq motifs
SCPD, YetFasco		Yeast	
Elemento, Redfly FlyFactorSurvey,Tiffin		Drosophila	
		Brachy-1	

Motif Enrichment Analysis

- Identifies over and under-represented known motifs in a set of regions
- The TFs whose DNA binding motifs are enriched in a set of regulatory regions are candidate transcription regulators of that gene/promoter/enhancer set.
- Without ChIP-seq, identifying a co-regulated gene sets is difficult. Use Ontologies, pathways, GSEA etc.
- Picking the right **background model** will determine the success of the motif enrichment analysis:
 - All core-promoters from protein coding or non-coding genes etc.
 - Higher order Markov model based backgrounds
 - A sequence set similar in nucleotide composition, length and number to the test set
 - Open chromatin regions or a shuffled test sequence set.

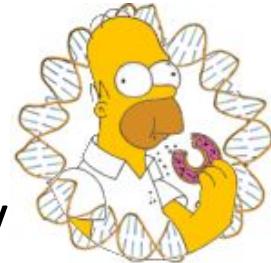
Motif detection and enrichment analysis

- **MEME Suite and MEME-Chip** <http://meme.nbcr.net>
- Given a set of genomic regions, it performs
 - Motif detection ([FIMO](#))
 - *ab initio* motif discovery -novel TF binding sites ([MEME](#), [DREME](#))
 - motif enrichment analysis -known TF enrichment ([Centrimo](#), [AME](#))
 - motif visualization ([MAST](#) and [AMA](#))
 - binding affinity analysis
 - motif identification -compare to known motifs ([TOMTOM](#))
- MEME -expectation maximization (EM) to discover probabilistic models of DNA-binding by single TFs or TF complexes.
- DREME -simpler, non-probabilistic model (regular expressions) to describe the short binding motifs.



Motif detection

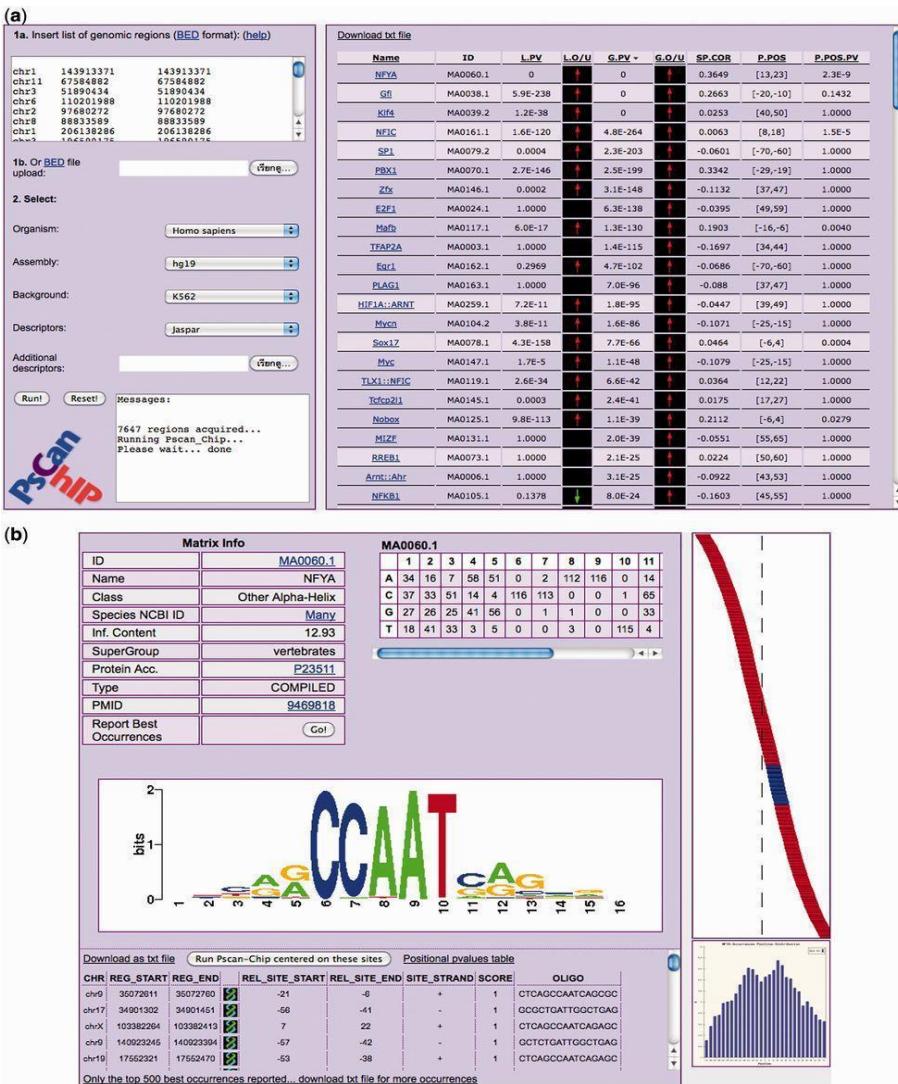
- HOMER v4 <http://homer.salk.edu/homer/index.html>
- Large number of (Perl and C++) tools for ChIP-seq analy
- Provides both *de novo* and PWM scanning based motif identification and enrichment analysis.
- User can specify custom background. (Randomly selected, GC or CGI matched backgrounds.)
- Uses a collection of ChIP-seq derived PWMs or user can specify PWM.
- Can help with Peak annotation, GO enrichment analysis, Extract peak sequences, Visualization.



Motif Enrichment Analysis

Pscan-Chip

- Motif enrichment analysis using PWM databases and user defined background models.
- Optimized for ChIP-seq.
- Ranked lists of enriched motifs.
- Sequence logo's and motif enrichment distribution plots.



Zambelli et al., 2013 Nucleic Acids Res.

Meta-Motif Analyzers

<http://131.174.198.125/bioinfo/gimmemotifs/>

GimmeMotifs: a *de novo* motif prediction pipeline, especially suited for ChIP-seq datasets. It incorporates several existing motif prediction algorithms in an ensemble method to predict motifs and clusters these motifs using the weighted information content (WIC) similarity scoring metric.

BioProspector <http://motif.stanford.edu/distributions/bioprospector/>

GADEM <http://www.niehs.nih.gov/research/resources/software/gadem/index.cfm>

Improbizer <http://users.soe.ucsc.edu/~kent/>

MDmodule (included in the MotifRegressor Package) <http://www.math.umass.edu/~conlon/mr.html>

MEME <http://meme.sdsc.edu/>

MoAn <http://moan.binf.ku.dk/>

MotifSampler <http://homes.esat.kuleuven.be/~sistawww/bioi/thijs/download.html>

Trawler <http://ani.embl.de/trawler/>

Weeder <http://159.149.160.51/modtools/>

L9: Identifying direct targets of TFs

Network Biology: reverse engineer regulatory networks by integrating TF binding and gene expression

- Not all TF binding sites are transcriptionally active. The collection of transcriptionally active targets of a TF is its **regulome**.
- Regulomes can be used to “explain” the phenotype under consideration and understand aspects of biological systems.
- Regulomes in combination with pathway and network modelling approaches can then be used decipher the networks underlying phenotypes.
- These networks provide information on connectivity, information flow, and regulatory, signaling and other interactions between cellular components.
- **BioNet, GeneNetworkBuilder**

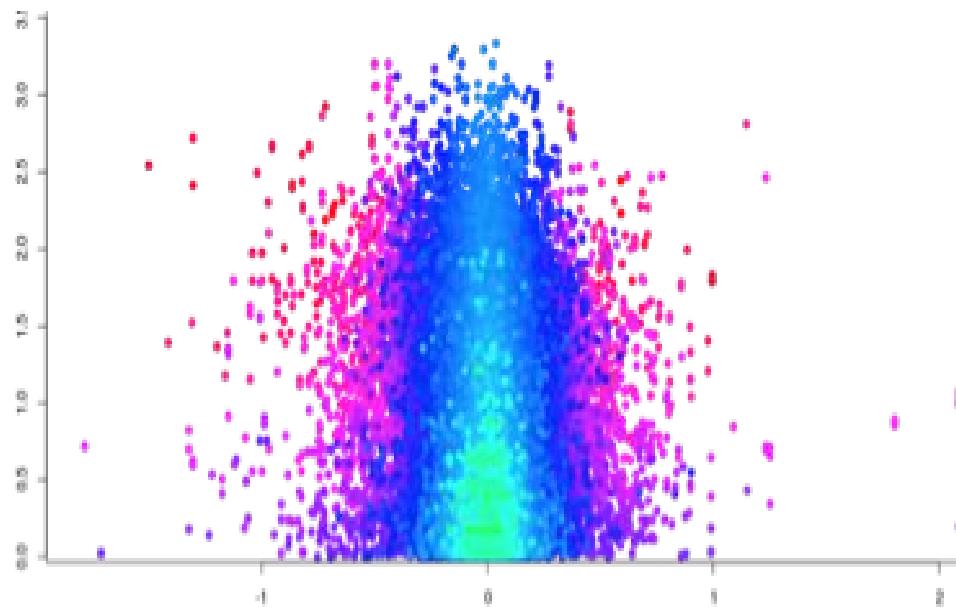
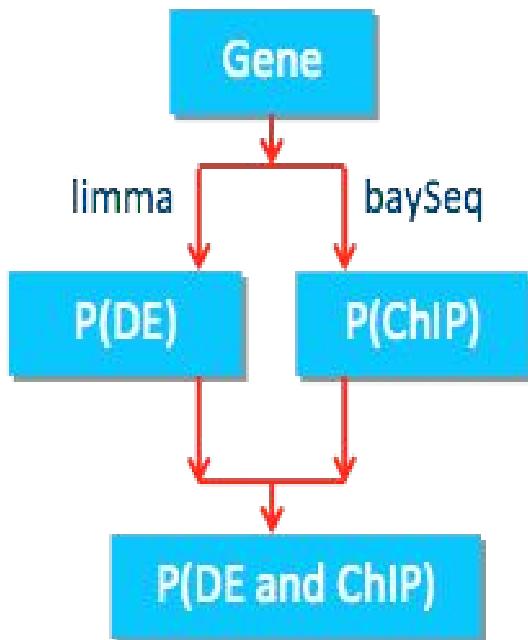
TF Direct Target detection

Rcade (R-based analysis of ChIP-seq And Differential Expression)

- Rcade is a Bioconductor package developed by Cairns *et al.*, that utilizes **Bayesian** methods to integrates ChIP-seq TF binding, with a transcriptomic Differential Expression (DE) analysis.
- The method is read-based and independent of peak-calling, thus avoids problems associated with peak-calling methods.
- A key application of Rcade is in inferring the direct targets of a transcription factor (TF).
- These targets should exhibit TF binding activity, and their expression levels should change in response to a perturbation of the TF.

Rcade

- **Rcade: R based analysis of ChIPseq And Differential Expression**
- Bayesian approach used to integrate ChIP-seq with differential expression to identify direct transcriptional targets of transcription factors.



$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

Rcade

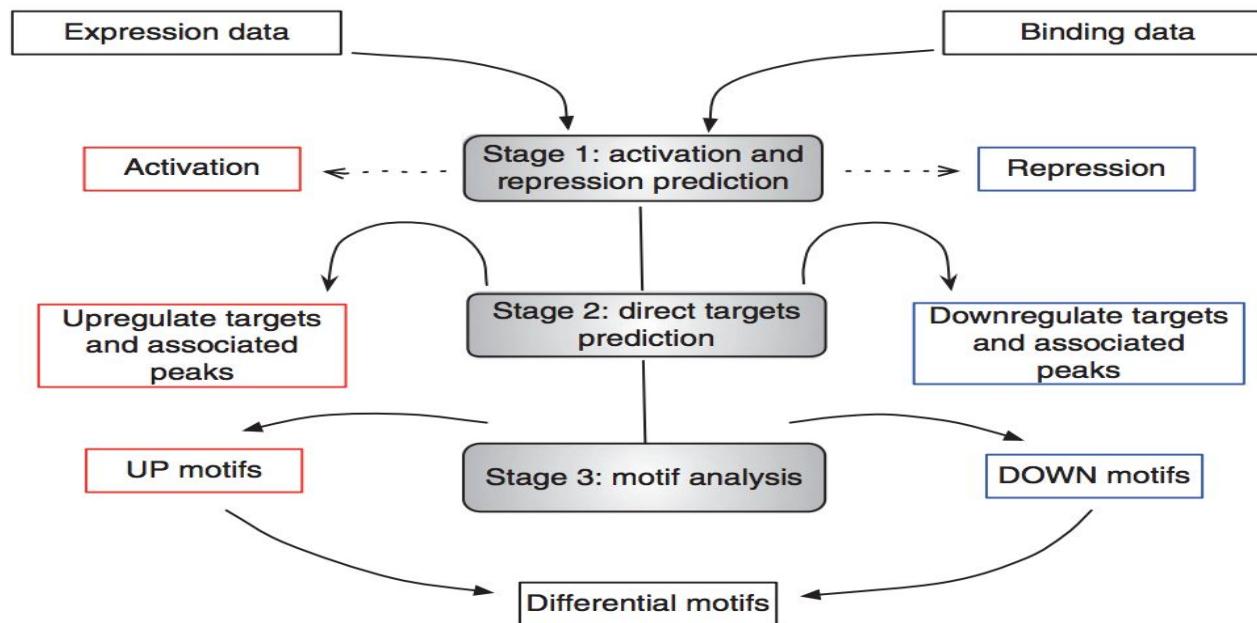
- Rcade integrates posterior probabilities of binding (determined via the `baySeq` package) with those of differential expression (determined via the `limma` package).

$$B = \log\left(\frac{PP}{1 - PP}\right)$$

- Rcade uses a fully Bayesian modelling approach. In particular, it uses log-odds values (a measure of probability), or B-values, in both its input and output. The log-odds value is related to the posterior probability (PP) of an event, as per the formula above.
- Priors need to be defined.
- A number of output files are generated by Rcade. Usually, the file of interest is “DEandChIP.csv”, which contains a list of genes most likely to have both DE and ChIP signals ranked by their B-value.
- More on Rcade @ the practical!

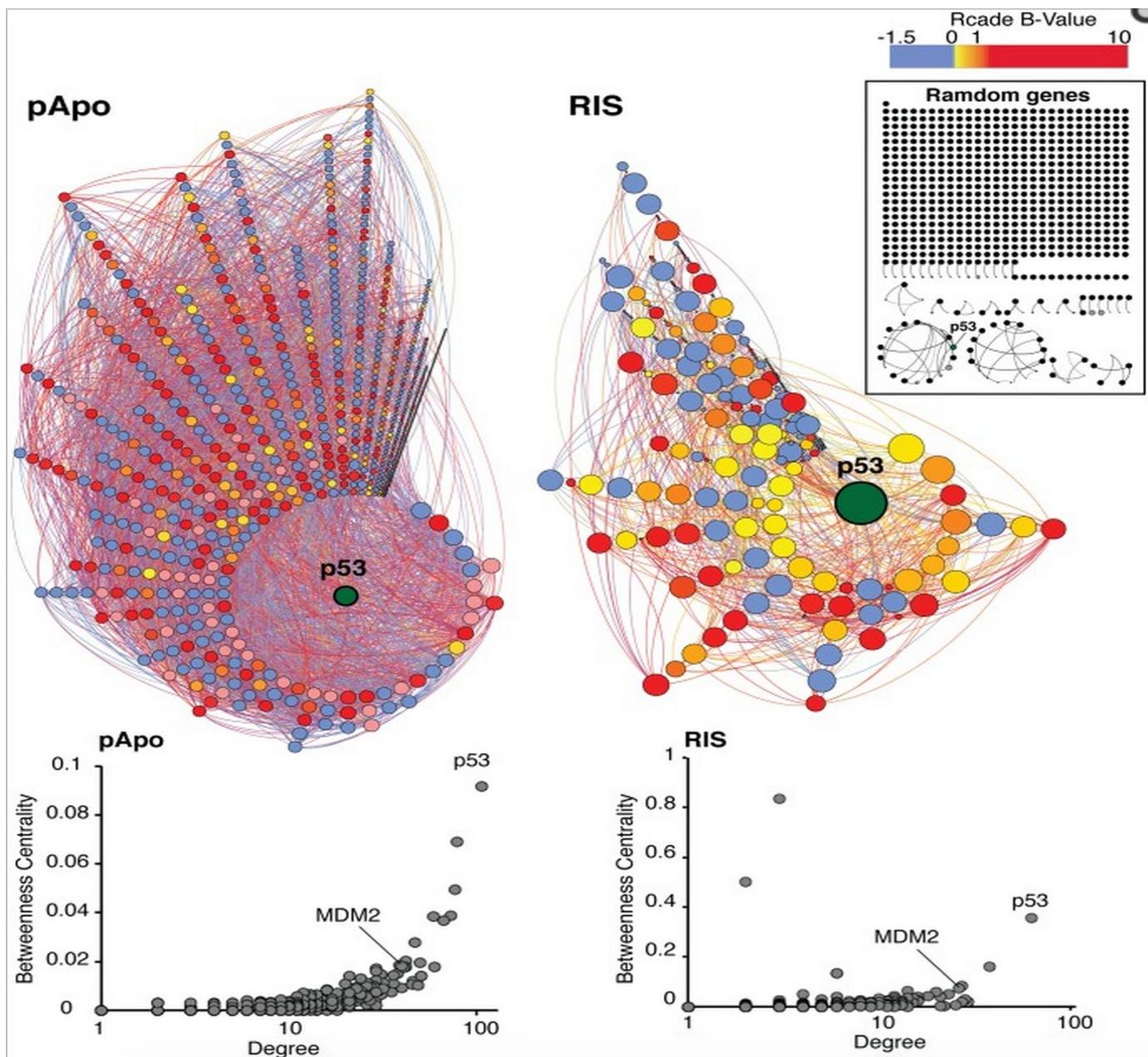
Beta

- Three main functionalities:
 - to predict whether a factor has activating or repressive function
 - to *infer* the factor's target genes
 - to identify the binding motif of the factor and its collaborators



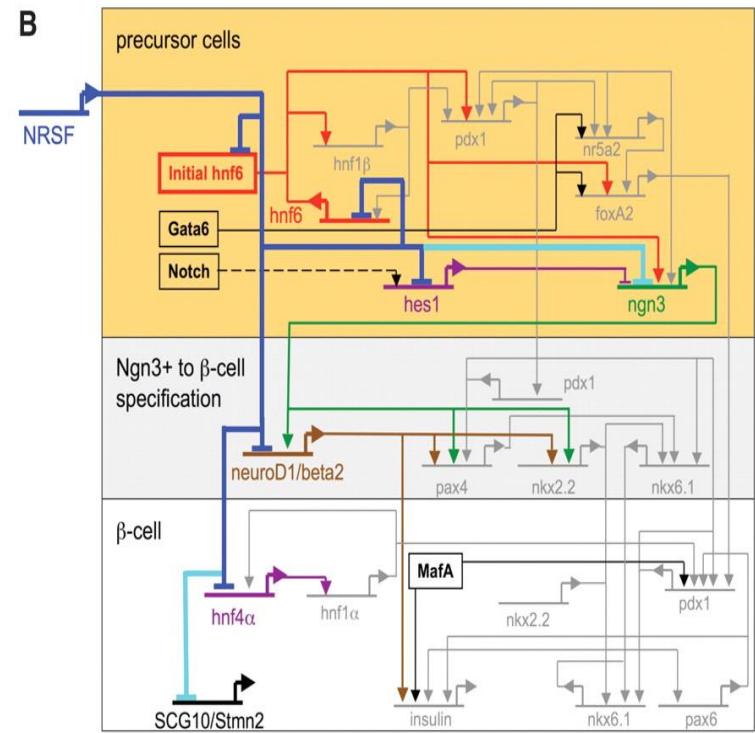
Functional Association Networks

Network Topology Analysis

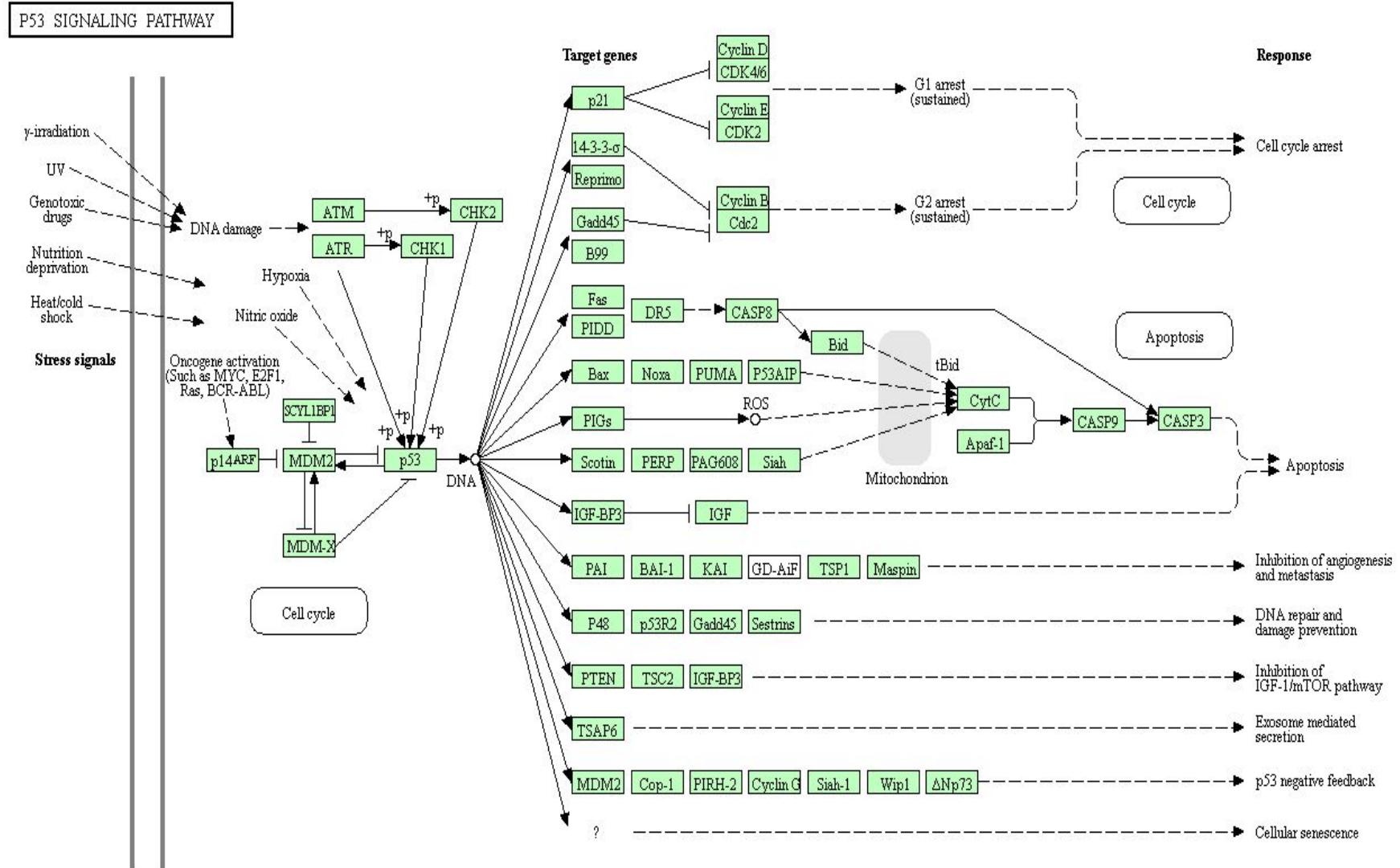


Regulomes: from active regulatory elements to networks

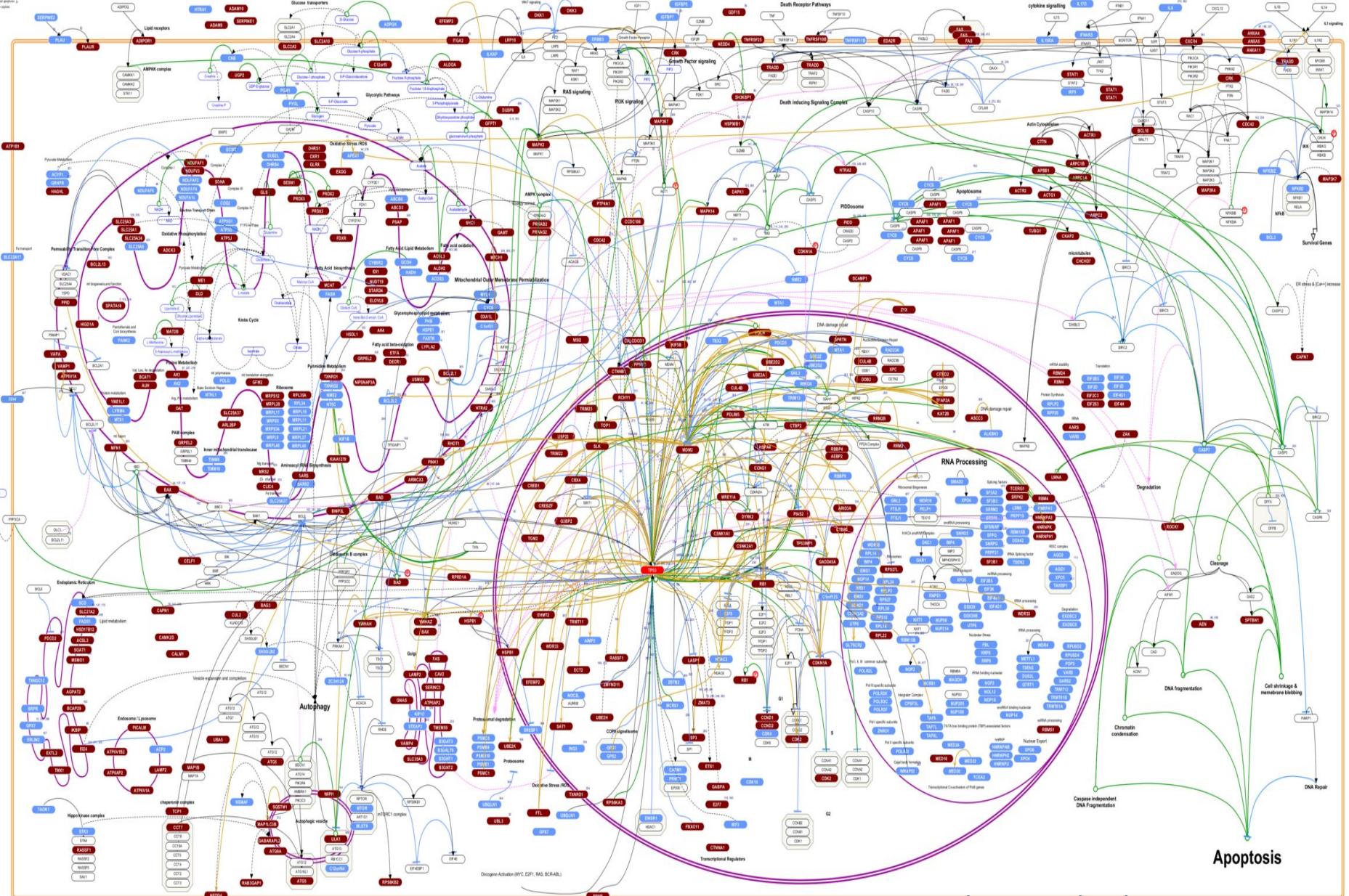
- Not all TF binding sites are transcriptionally active. The collection of transcriptionally active targets of a TF is it's **regulome**.
- Regulomes can be used to “explain” the phenotype under consideration and understand aspects of biological systems.
- Regulomes in combination with pathway and network modelling approaches can then be used decipher the networks underlying phenotypes.
- These networks provide information on connectivity, information flow, and regulatory, signaling and other interactions between cellular components.



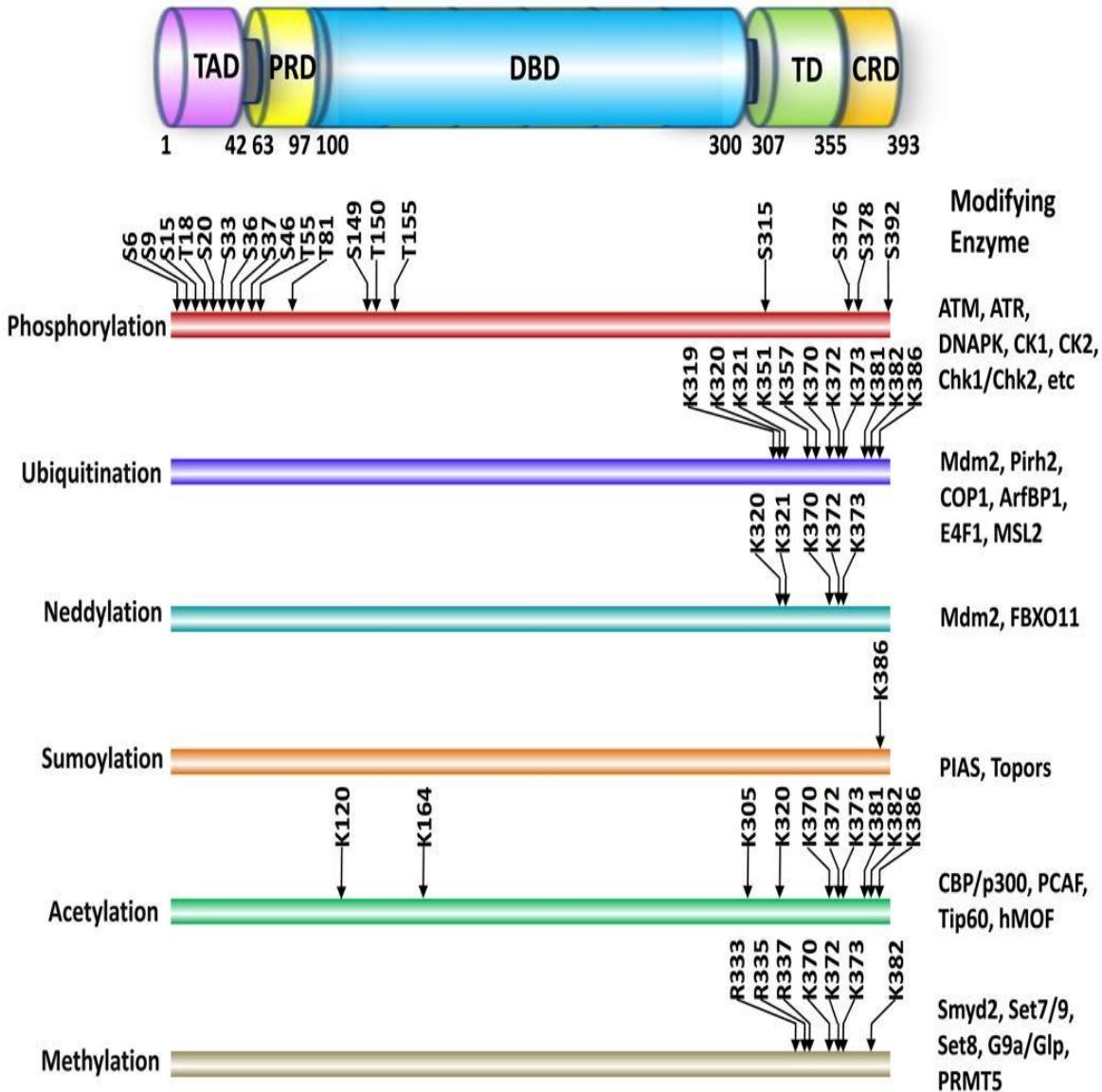
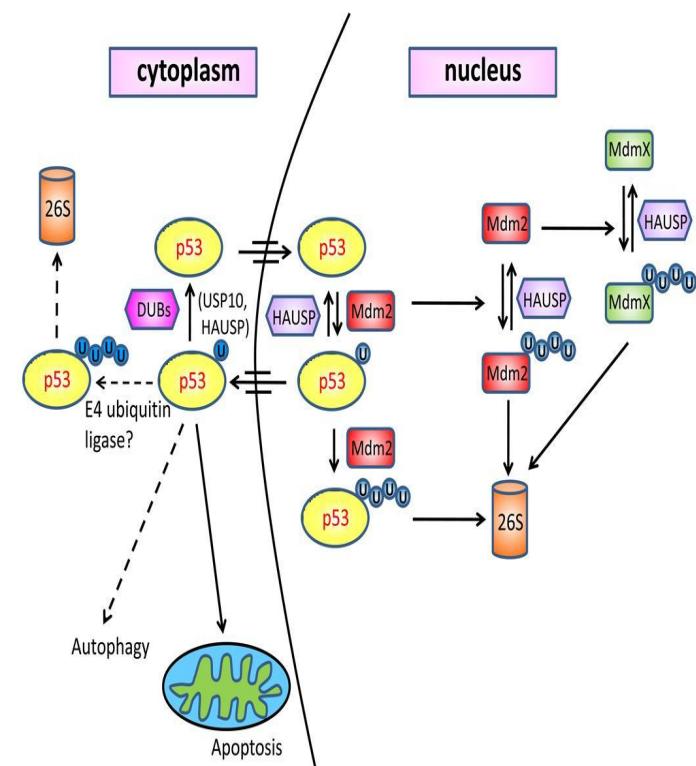
KEGG: p53 signalling pathway



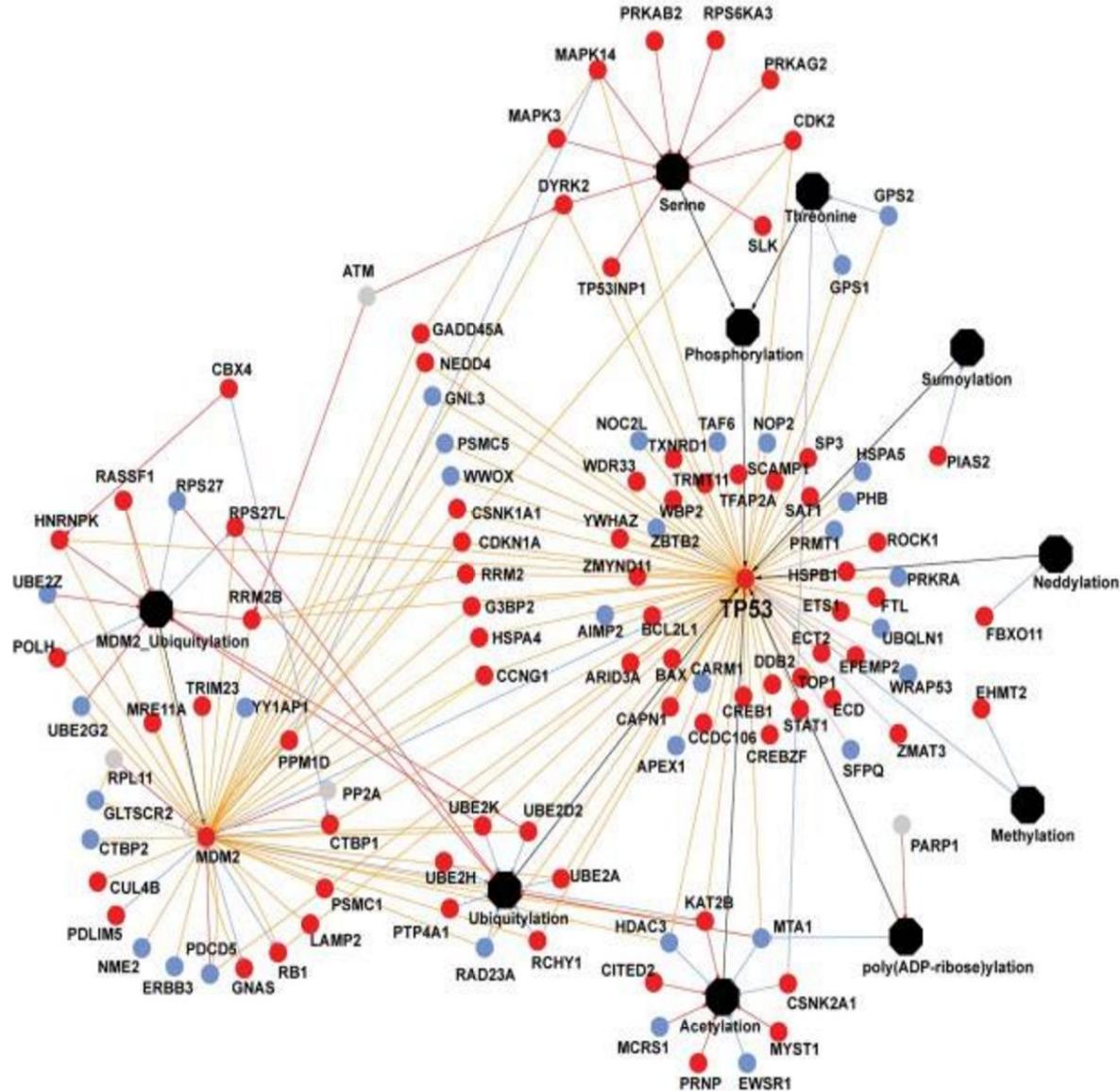
The TP53 Regulome



Fine tuning regulation: post-translational modifications



The Self-Regulatory TP53 Network



Differential binding analysis 1

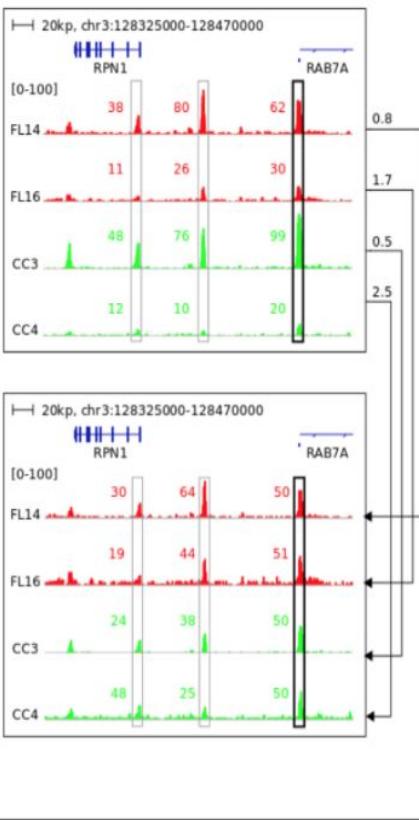
- **THOR** is an HMM-based approach to detect and analyze differential peaks in two sets of ChIP-seq data from distinct biological conditions with replicates.
- Performs genomic signal processing and normalization, peak calling and p-value calculation in an integrated framework.

A - THOR

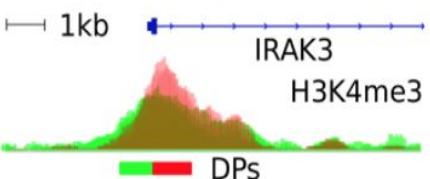
1 - preprocessing

- fragment size estimation
- GC-content normalization
- input-DNA normalization
- input-DNA subtraction

2 - signal normalization



5 - DP estimate example

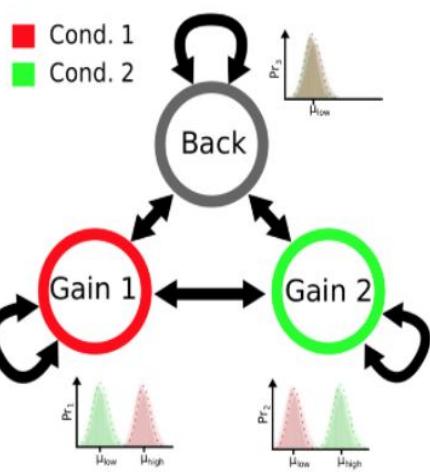


4 - postprocessing

- P-value estimate
- strand lag filter

3 - HMM

- Cond. 1 (red)
- Cond. 2 (green)



B - Competing Methods

One-Stage DPC

- PePr
- DiffReps
- csaw

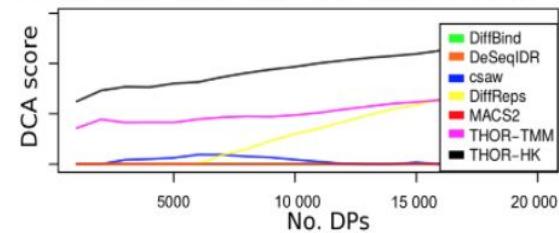
Two-State DPC

- MACS2
- DESeq-IDR
- DESeq-JAMM
- DiffBind

C - Evaluation

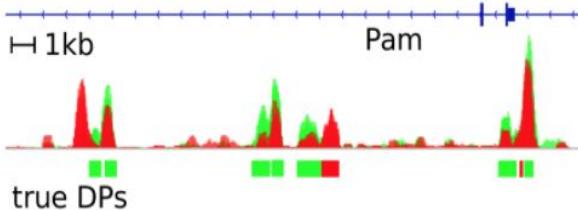
1 - biological data

- 4 studies and 13 DPC problems
- evaluation with expression/histones (DCA)



2 - simulated data

- 12 scenarios: no. of replicates, within condition variance, ...



Differential binding analysis 2

- **Diffbind** is a Bioconductor package by Stark *et al.*, for identifying sites that are differentially bound between two sample groups.
 - It includes functions to support the processing of peak sets, overlapping and merging peak sets, counting sequencing reads overlapping intervals in peak sets, and identifying statistically significantly differentially bound sites based on evidence of binding affinity (measured by differences in read densities).
 - More on DiffBind @ the practical!

