

Introduction to (epi)Genomic data integration

Shamith Samarajiwa

CRUK Bioinformatics Summer School
July 2019



UNIVERSITY OF
CAMBRIDGE

(epi)Genomic Data Integration

- Combining different biological layers to understand phenotypes
- Computational or Statistical approaches

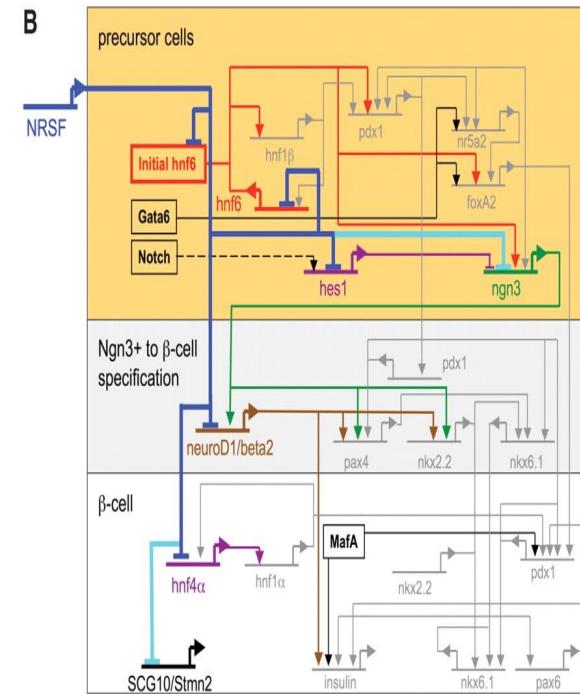
Some examples:

1. TF direct targets: **TF ChIP-seq** + **RNA-seq**
2. Enhancer-Promoter Interactions
 - Chromosomal Conformation Capture (**Hi-C**) + **Histone mark ChIP-seq** + **RNA-seq**
3. Impact of chromatin accessibility on transcription
 - **ATAC-seq** + **RNA-seq** + **DRIP-seq**
4. Translational regulation
 - **iCLIP** + **Ribo-seq** + **RNA-seq**
5. Epigenetic silencing: **DNA methylation** + **RNA-seq**
6. Regulatory elements: multiple **Histone mark ChIP-seq**

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 TF binding sites are called the **Cistrome** and the collection of transcriptionally active targets (**regulons**) 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 reverse engineer the networks underlying phenotypes.
- These networks provide information on **connectivity**, **information flow**, and **signaling**, **regulatory**, **metabolic** and other interactions between cellular components and processes.



TF Direct Target detection

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

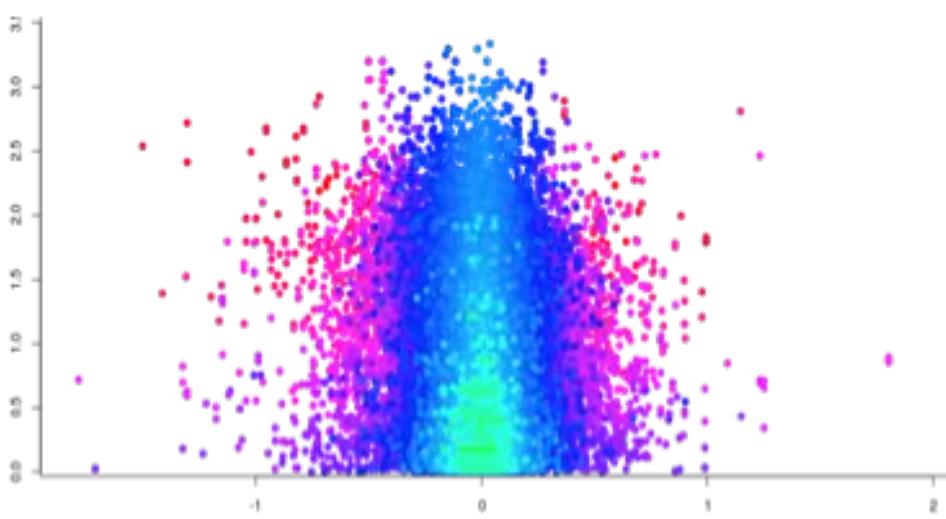
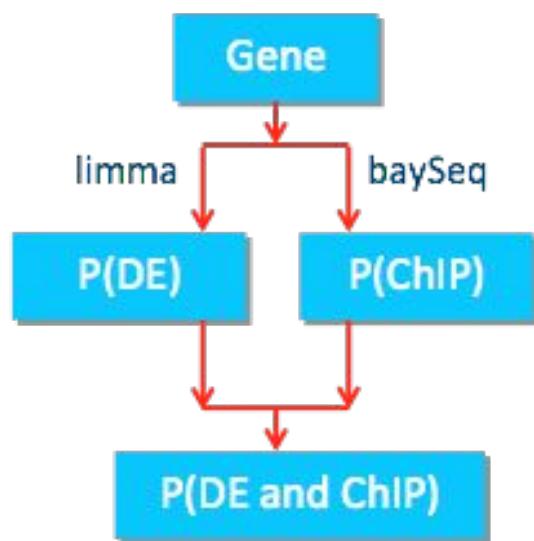
- Rcade is a Bioconductor package we (Cairns *et al.*) developed that utilizes **Bayesian** methods to integrates TF binding ChIP-seq, with transcriptomic Differential Expression.
- The method is “read-based” and independent of peak-calling, thus avoids problems associated with peak-calling methods.
- 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**.

Statistical approaches to data integration

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

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.



Rcade

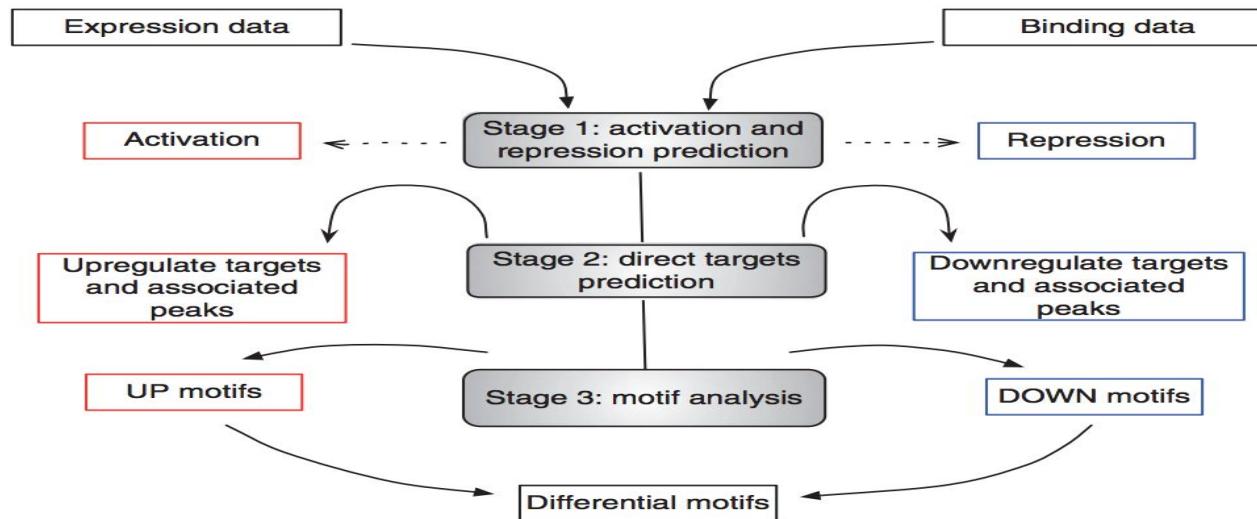
- Rcade integrates posterior probabilities of binding (determined via the `baySeq` package) with those of differential expression (determined via the `limma` package).

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

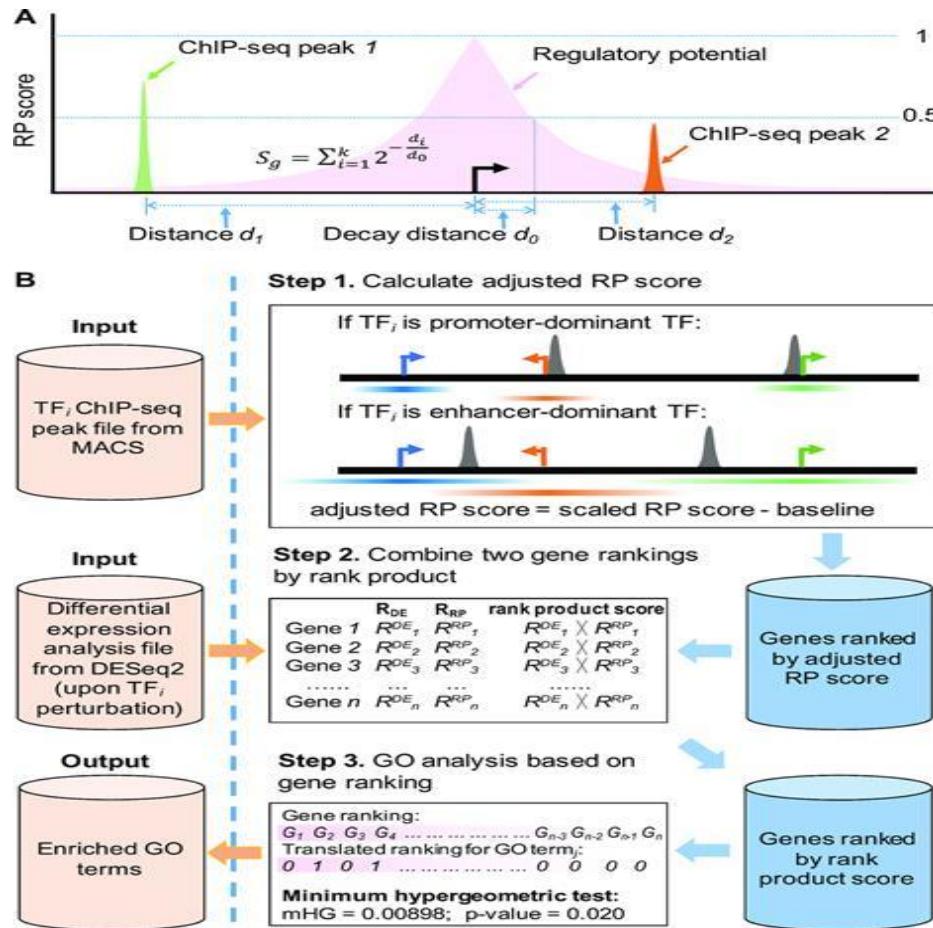
- Rcade uses a fully Bayesian modelling approach. In particular, it uses log-odds values (a measure of probability), or B-values, in both its input and output. The log-odds value is related to the posterior probability (PP) of an event, as per the formula above.
- Priors need to be defined.
- A number 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

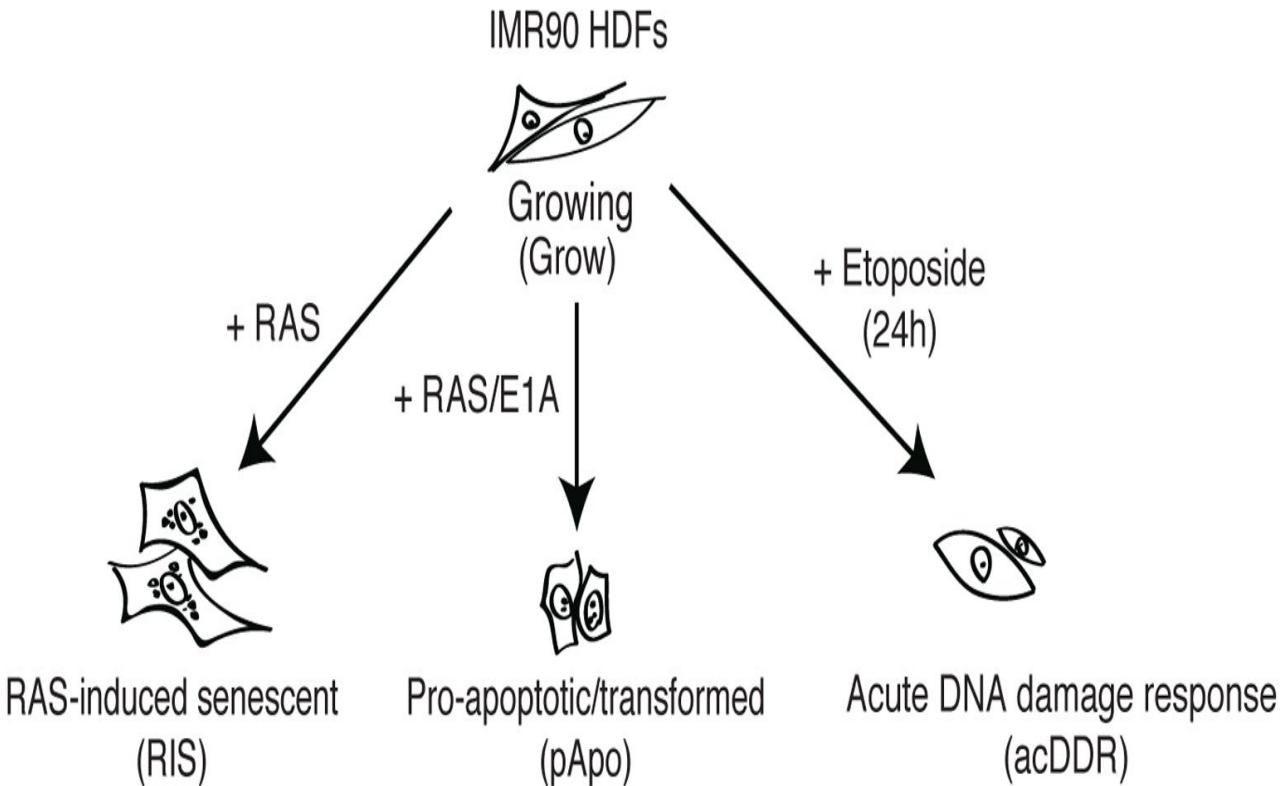
- Three main functionalities:
 - to predict whether a factor has activating or repressive function
 - to *infer* the factor's target genes
 - to identify the binding motif of the factor and its collaborators



Cistrome GO

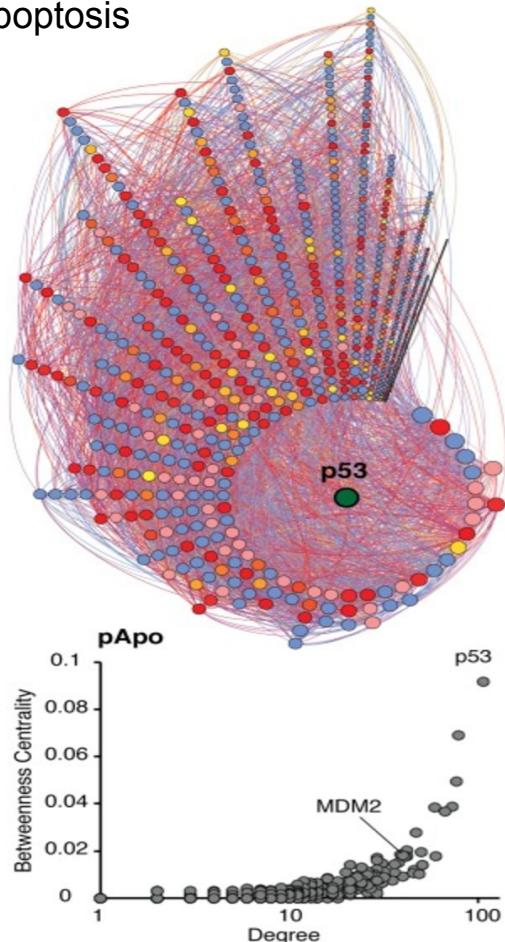


TP53 direct targets

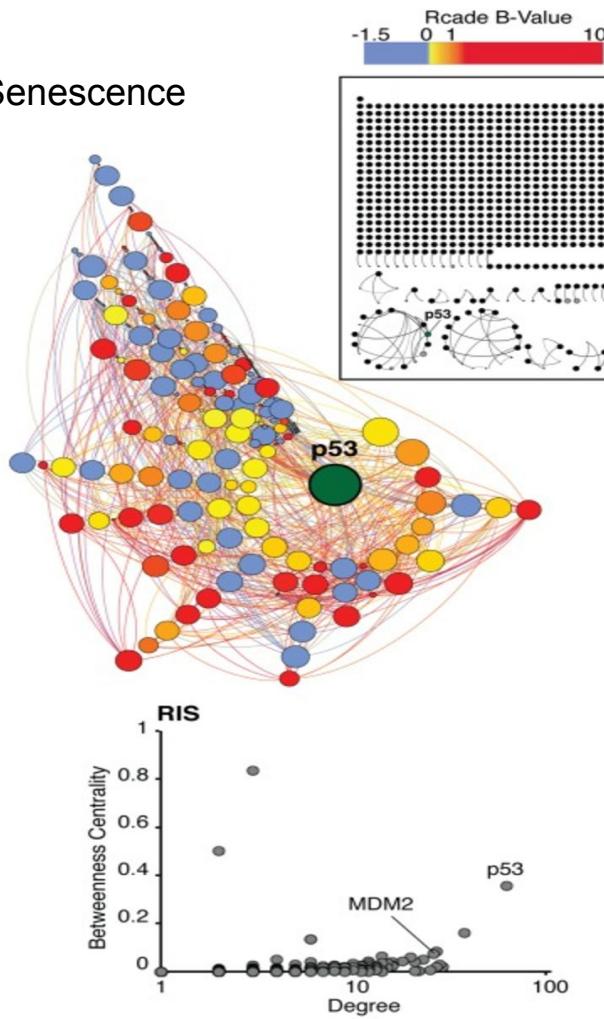


Identifying TP53 direct targets

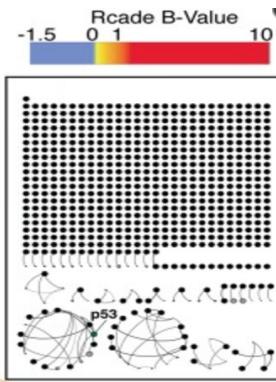
Apoptosis



Senescence



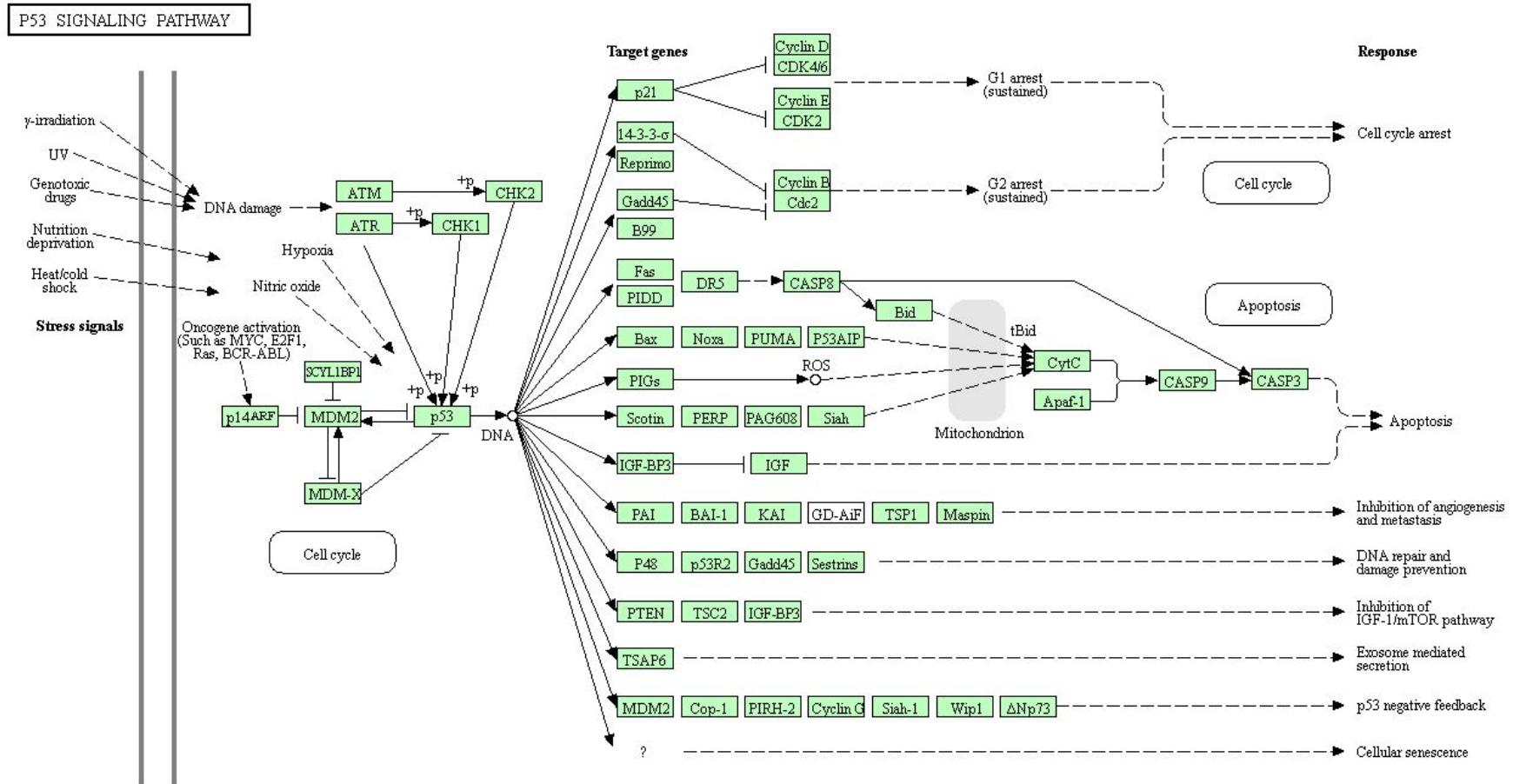
Random Genes



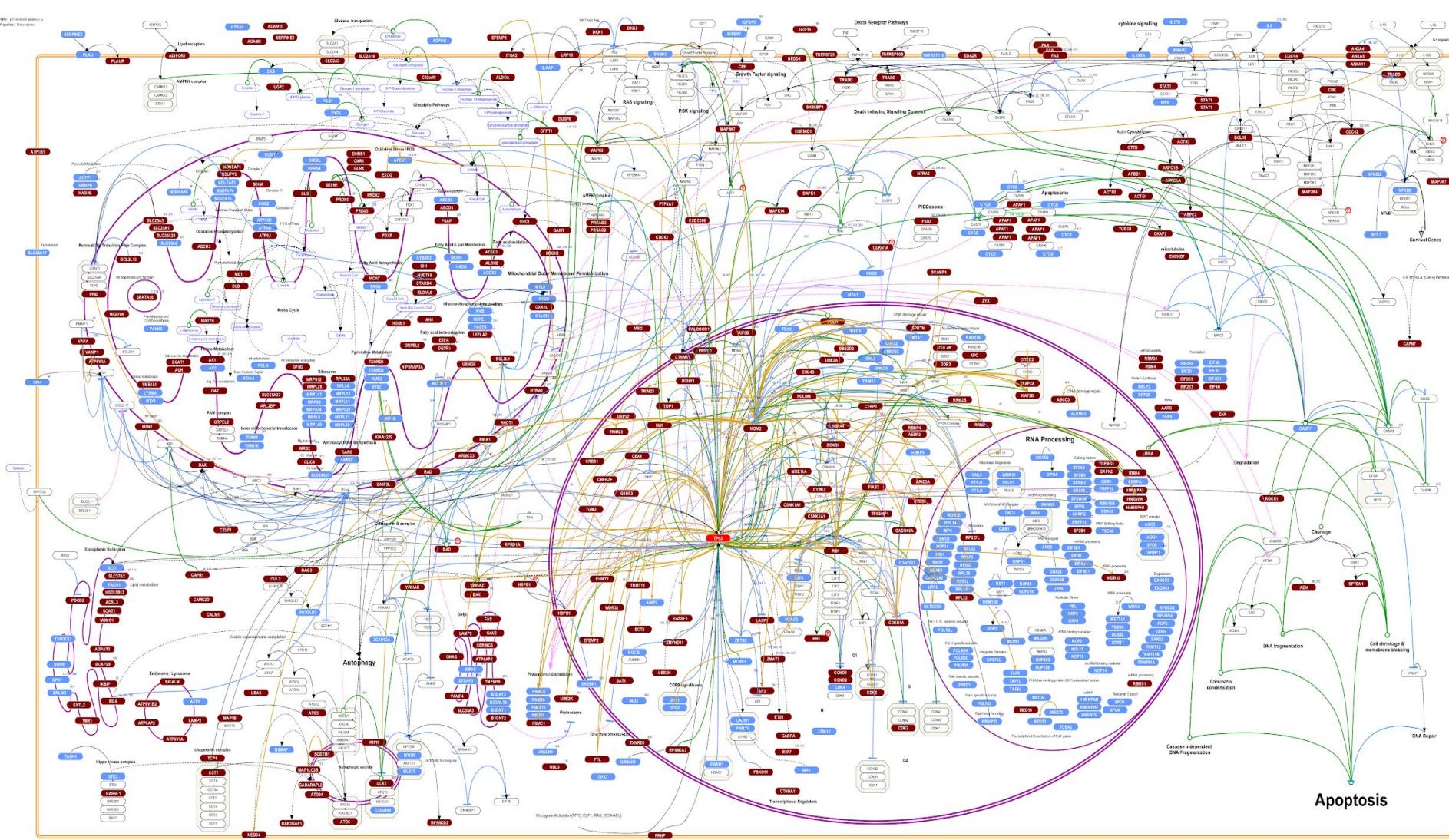
TF Direct Targets:

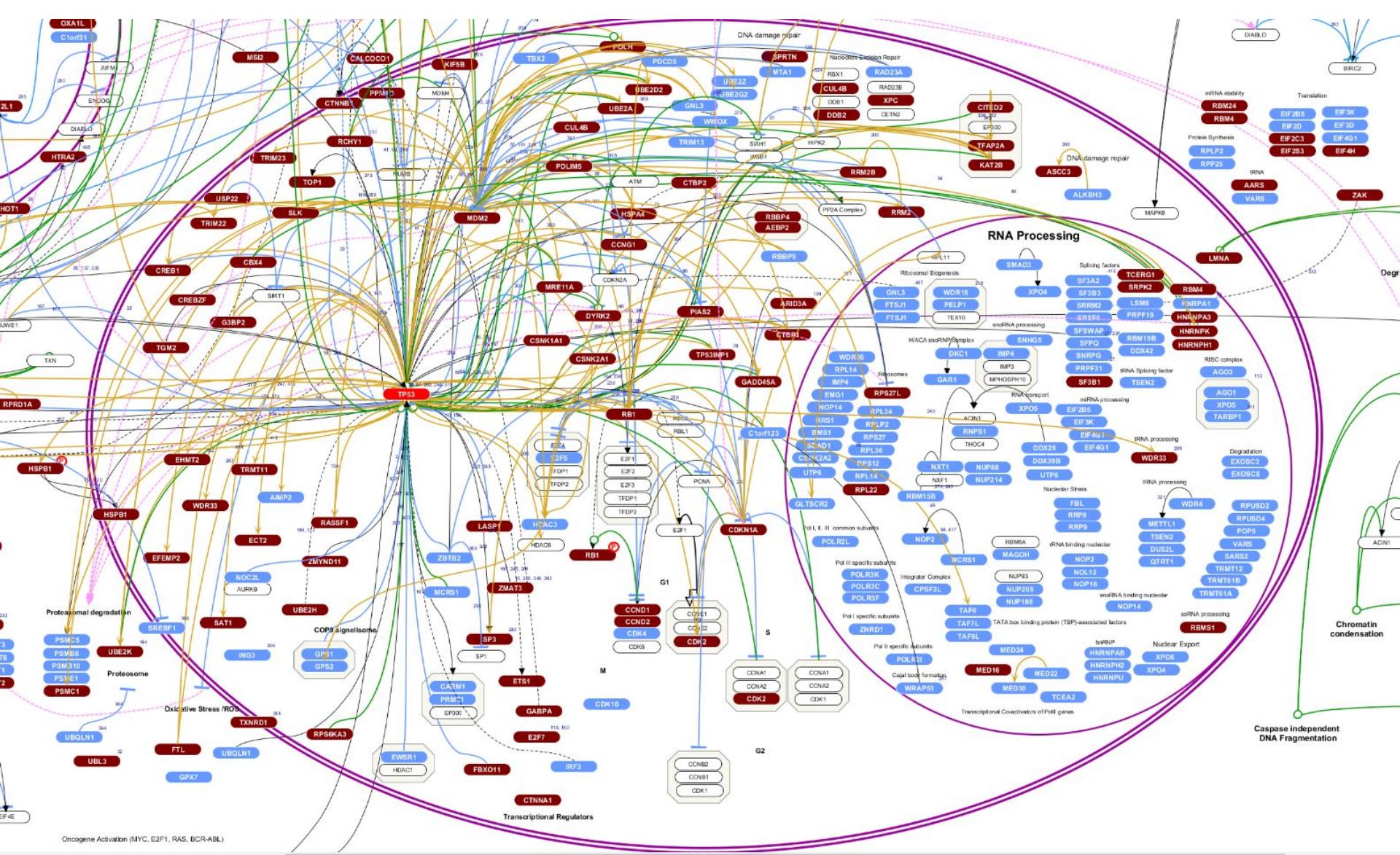
- TF binding
- Differential gene activation or inhibition
- Presence of a TF specific motif
- Literature evidence
- Enhancer - Promoter interactions

KEGG: p53 signalling pathway



Molecular Cartography of TF regulation





Interaction

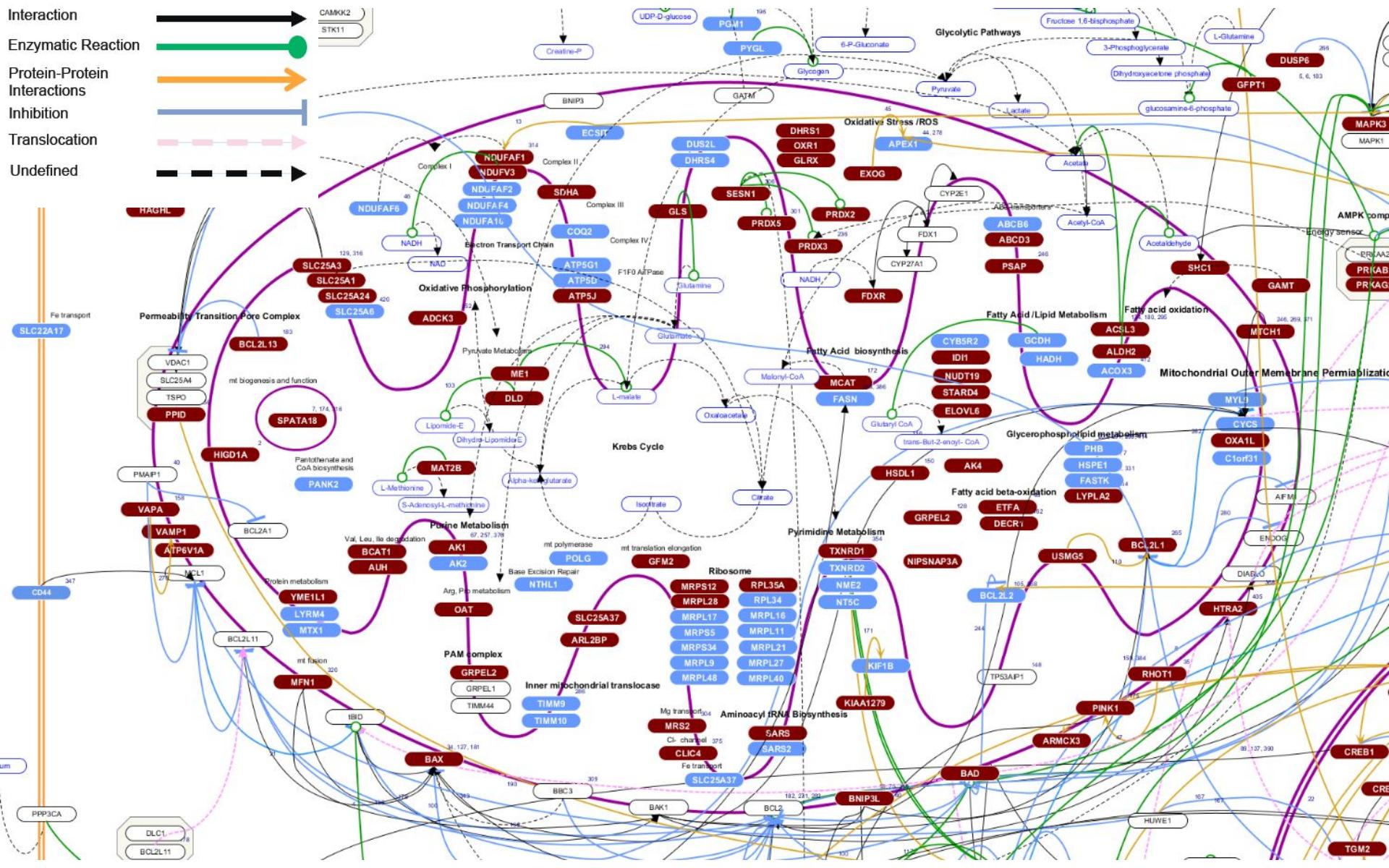
Enzymatic Reaction

Protein-Protein

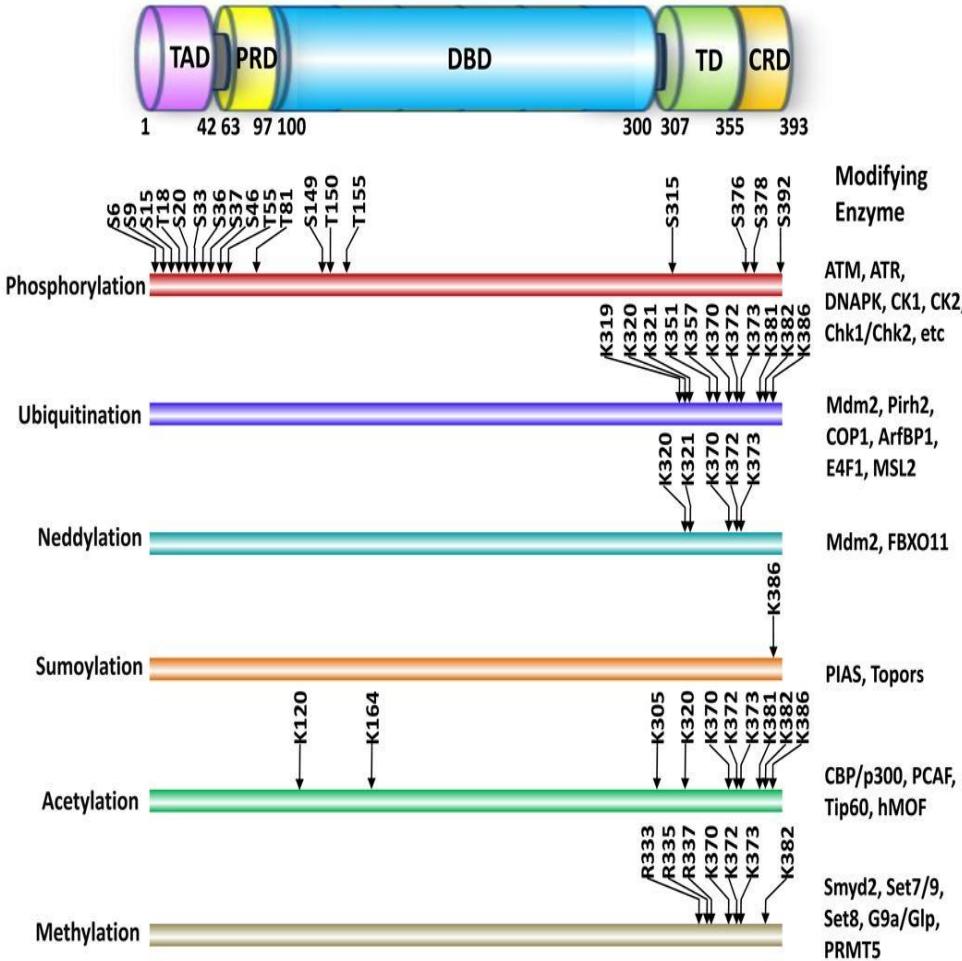
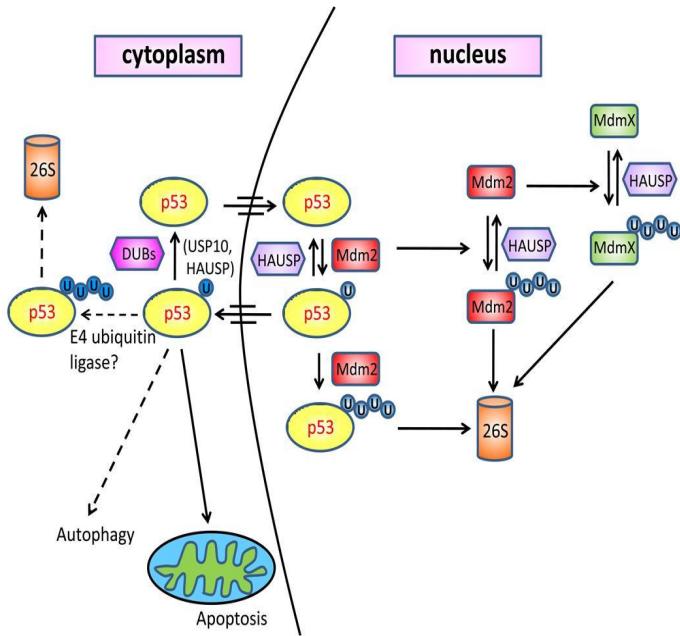
Interactions

Inhibition

Undefined

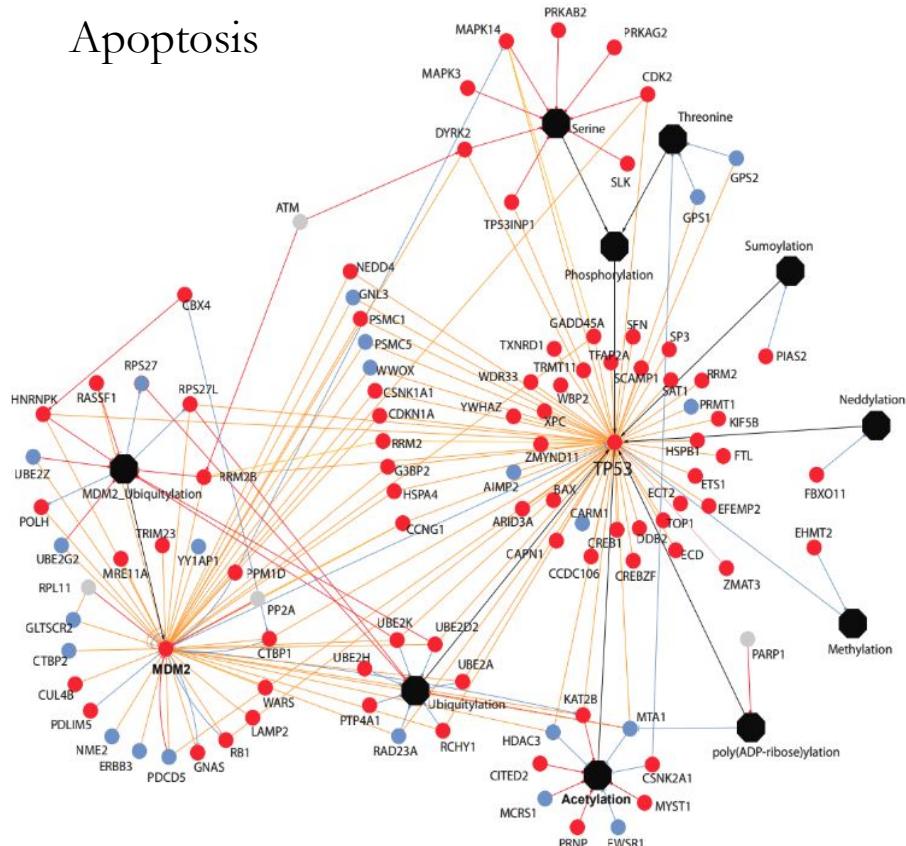


Fine tuning regulation: post-translational modifications

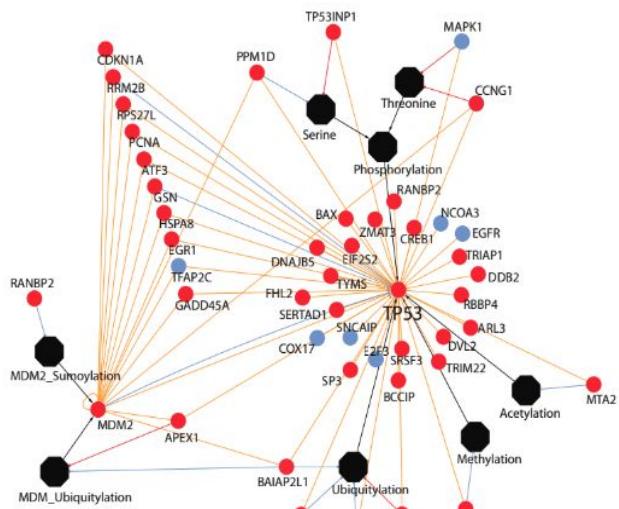


The Self-Regulatory TP53 Network

Apoptosis



Senescence



- p53 repressed hub genes
- p53 activated hub genes
- post- translational modification hubs
- not p53 regulated
- post translational modifiers regulating p53
- physical interactions
- activation
- repression

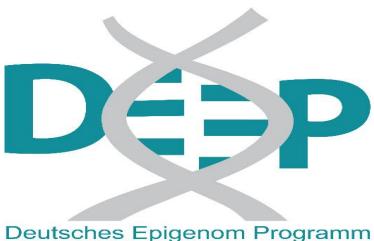
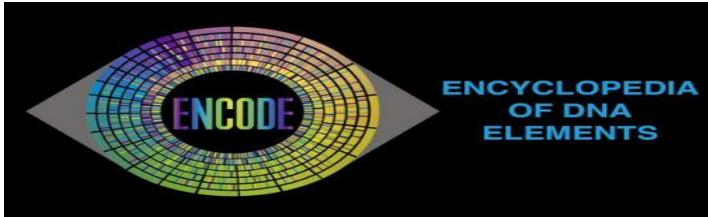
Epigenetics and Epigenomics

- **Epigenetics** encompasses processes that lead to heritable change in gene expression without changes to the DNA itself.
- DNA is packaged into chromatin. This nucleoprotein structure is highly dynamic and important for gene regulation. Chromatin states can vary between conditions, cells and tissue types and even within a single chromosome.
- The **Epigenome** refers to these chromatin states at a whole genome level. A multicellular organism has a single genome but many epigenomes.
 - **Paradox:** Although overall rates of cardiovascular disease increase with rising national prosperity, the least prosperous residents of a wealthy nation suffer the highest rates.

Developmental origin of health and disease

- The dutch famine (“Hongerwinter”) 1944-45 in German occupied Netherlands towards the end of the WWII affected 4.5 million people and led to ~22000 deaths.
- *“People ate grass and tulip bulbs, and burned every scrap of furniture they could get their hands on, in a desperate effort to stay alive.”*
- The Dutch Hunger Winter study, from which results were first published in 1976, provides an almost perfectly designed, although tragic, human experiment in the effects of intrauterine deprivation on subsequent adult health.
- Critical windows during development where epigenetic modification will affect adult health.
- Those exposed during **early gestation** experienced **elevated rates of obesity, altered lipid profiles, and cardiovascular disease**. In contrast, markers of **reduced renal function** were specific to those exposed in **mid-pregnancy**. Those who were exposed to the famine only during **late gestation** were **born small and continued to be small throughout their lives, with lower rates of obesity** as adults than in those born before and after the famine.

Large-scale epigenomic studies



Various human, mouse tissues

Histone and TF ChIP-seq,
Transcriptomics, Hi-C

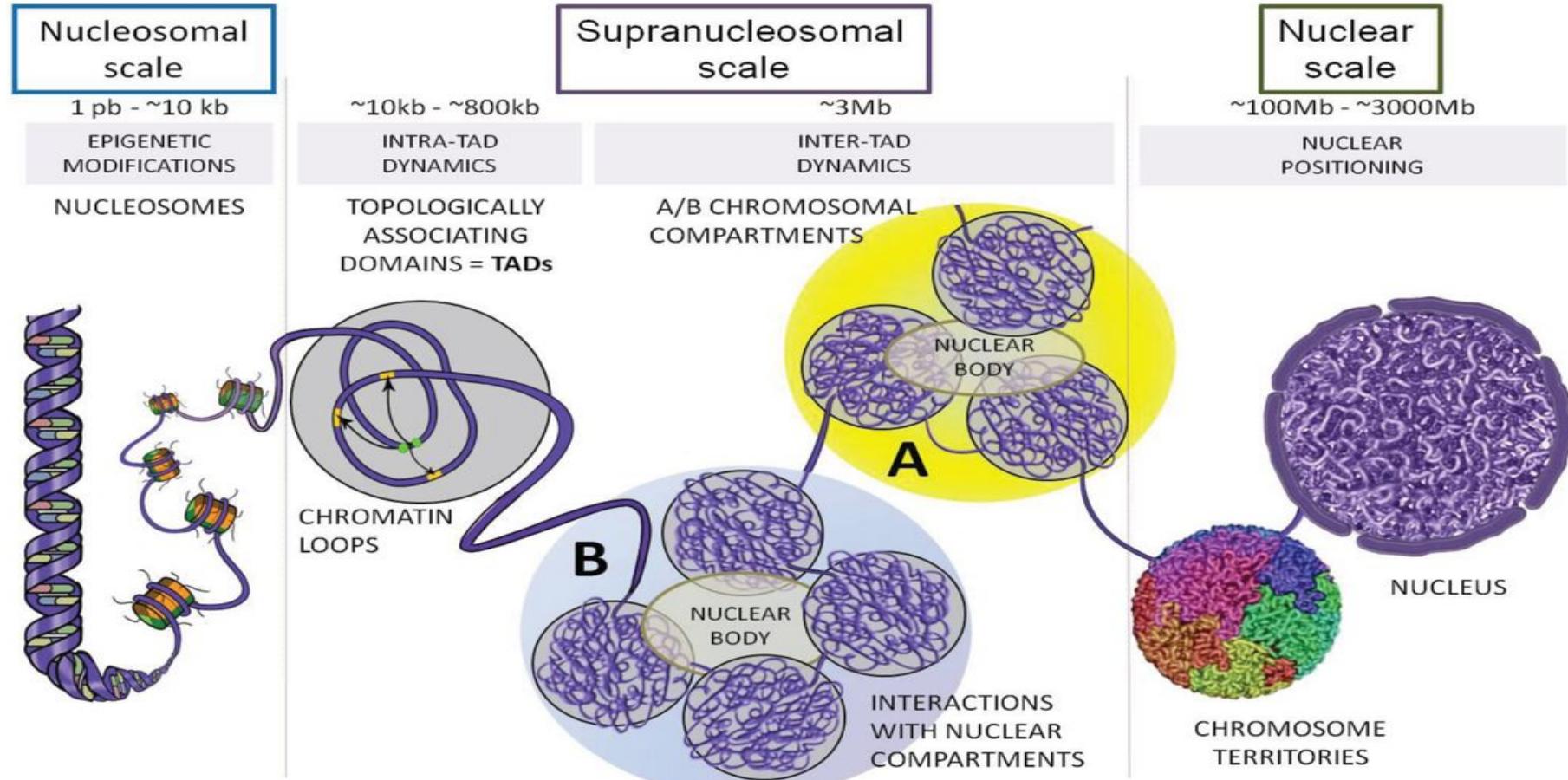
Epigenomes of 100 blood cell
types

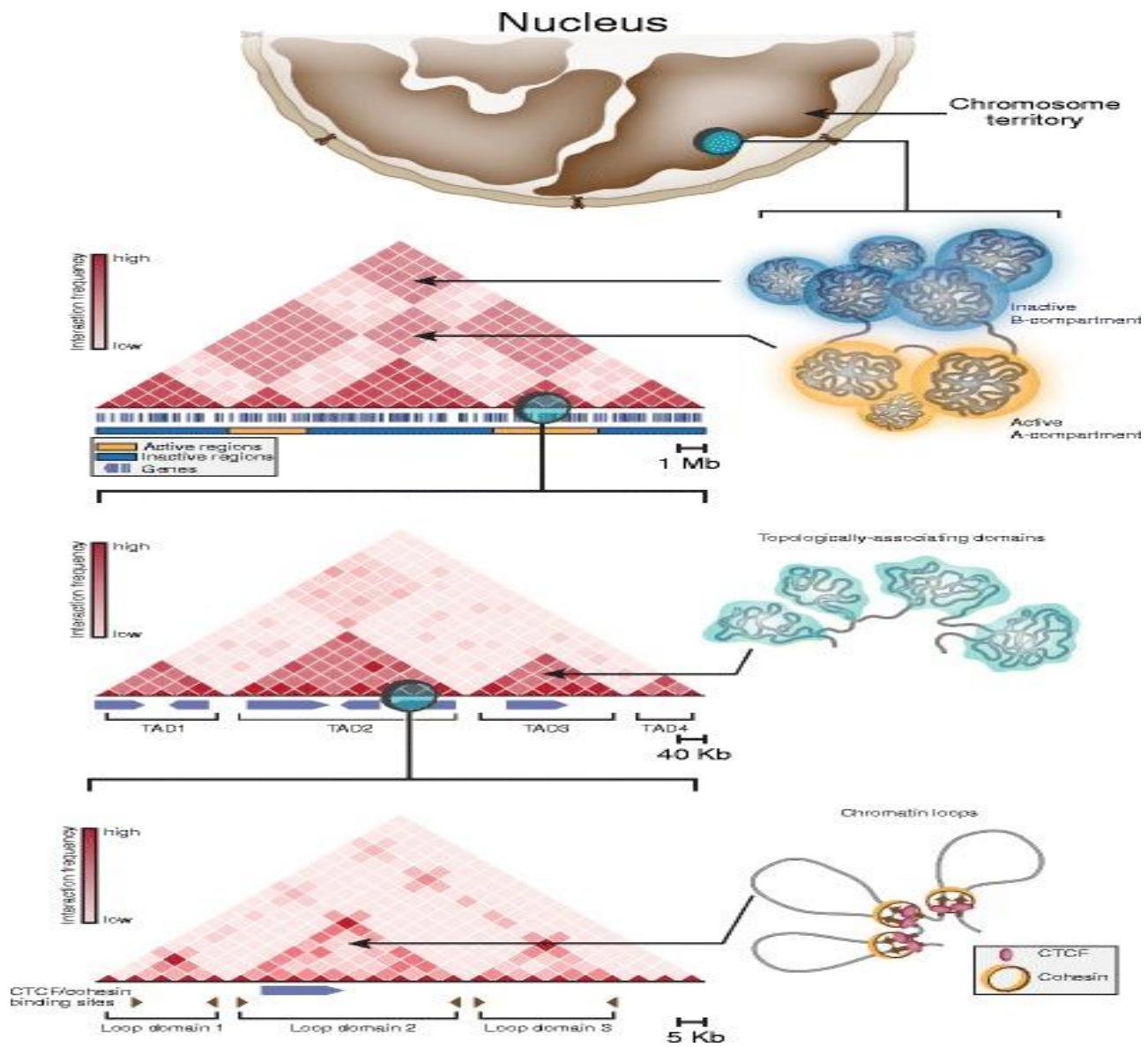
Stem cells, fetal tissues, adult
tissues



Methylomes

Current model of chromatin organization

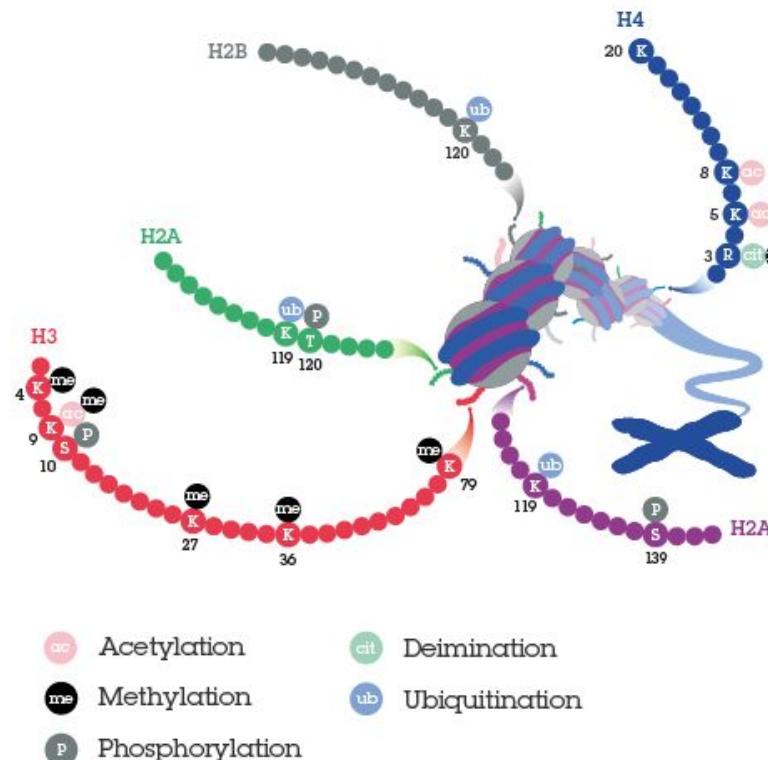




Histone Modifications

- Nucleosomes consist of 2x H2A/H2B and 2x H3/H4 histones.
- **80** known covalent modifications

H3K4me3 →	Histone 3
	K Residue is lysine, K
	4 ^{4th} residue.
	me3 Trimethylation



The most common histone modifications

Histone Modifications

Some examples:

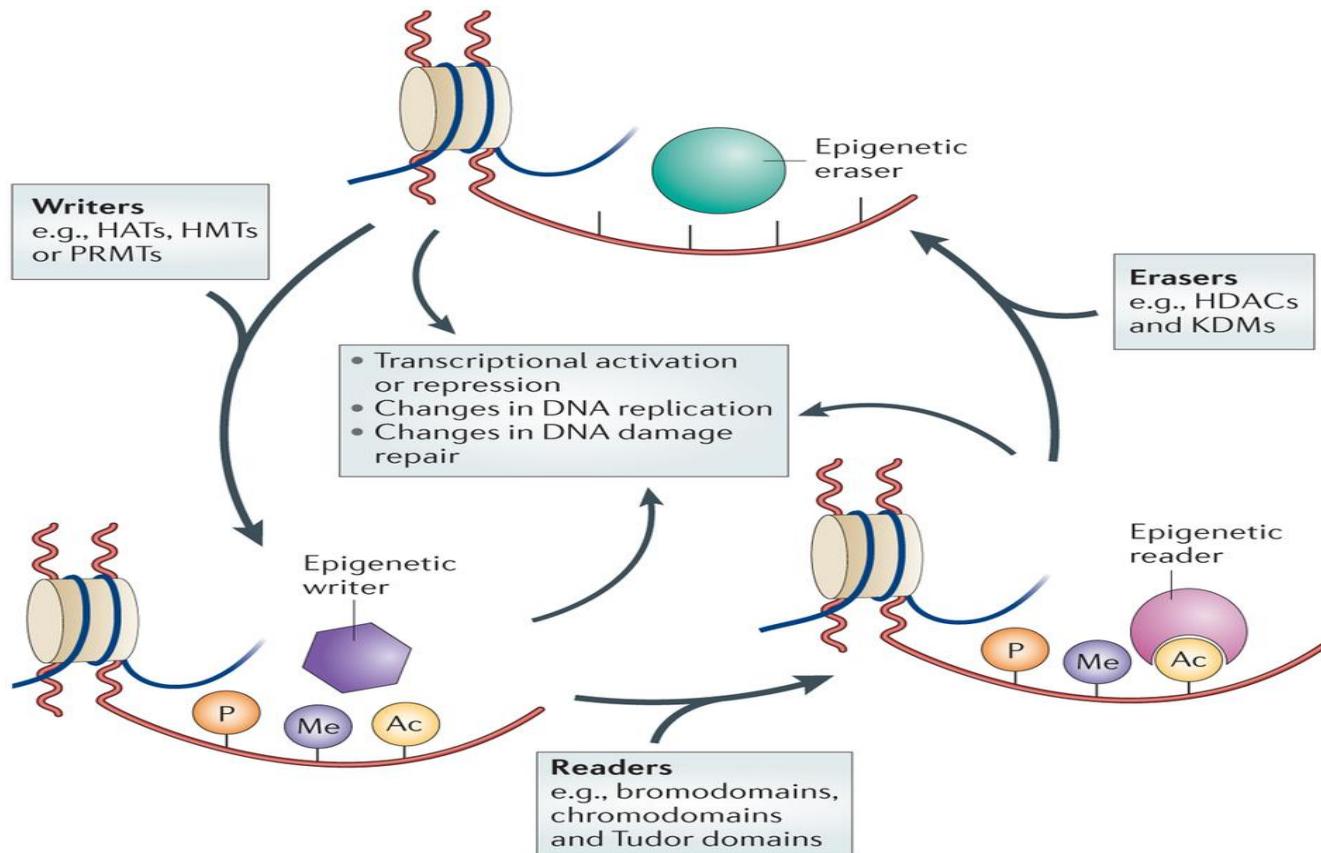
- H3K4me3 - active promoters
- High H3K4me1 and H3k27Ac, low H3K4me3 - active enhancers
- H3K27me3 -repression at promoters
- H3K9me3 - Heterochromatin (inactive, condensed chromatin)

More information at:

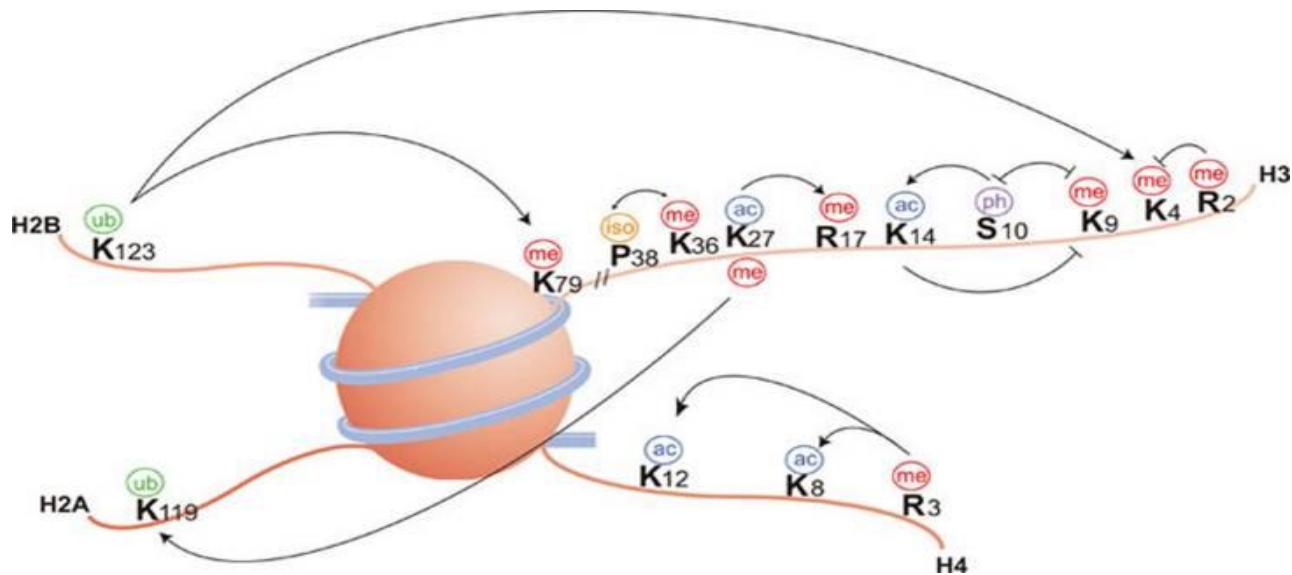
<http://epigenie.com/key-epigenetic-players/histone-proteins-and-modifications>

L

Epigenetic Readers, Writers and Erasers



Combinations of marks can have different effects



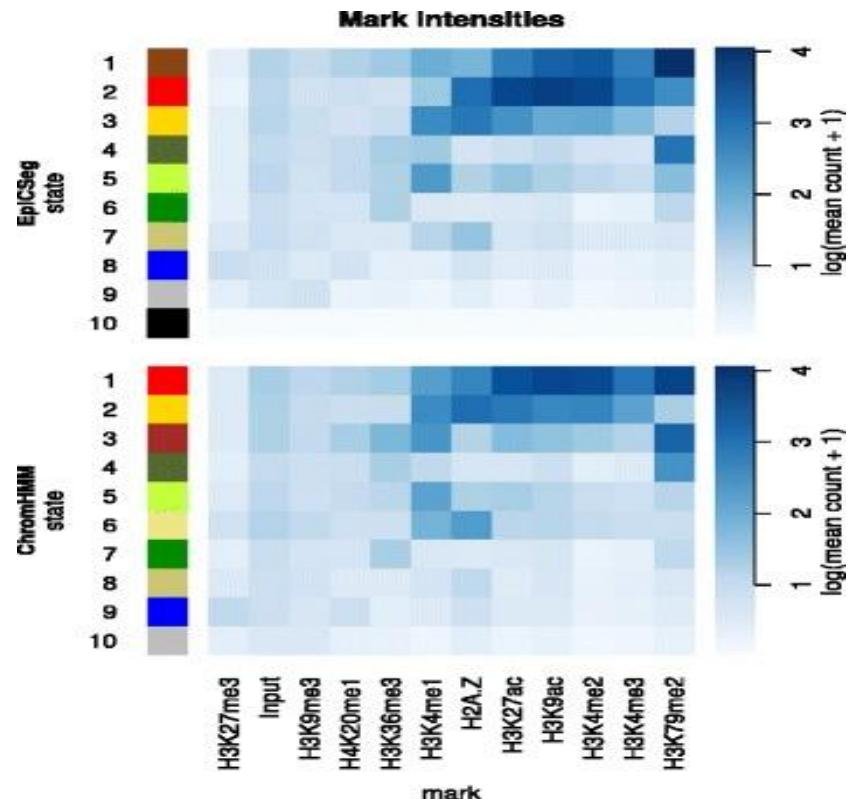
Bannister and Kouzarides (Cell Res. 2011)

To understand the entire code need to ChIP-seq each mark. This information has to be integrated and simplified.

Simplifying histone marks

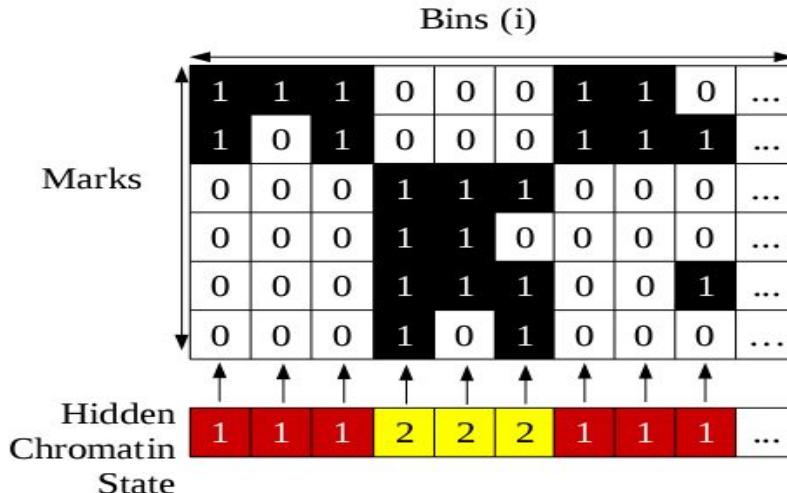
Unsupervised learning methods for *segmentation*:

- ChromHMM (Ernst et al., 2011)
- Segway (Hoffman et al., 2012)
- EpiCSeg (Mammana and Chung, 2015)
- GenoSTAN (Zacher et al., 2016)



Chromatin Segmentation Algorithms

- Genome divided into 200bp bins
- Adjust read position (shift 5' of each read 5'->3' by 0.5 the fragment length)
- Count reads in each bin for each mark and generate count matrix
- HMM with specified states is used to model the count matrices and derive segmentation



Chromatin Segmentation

Advantages:

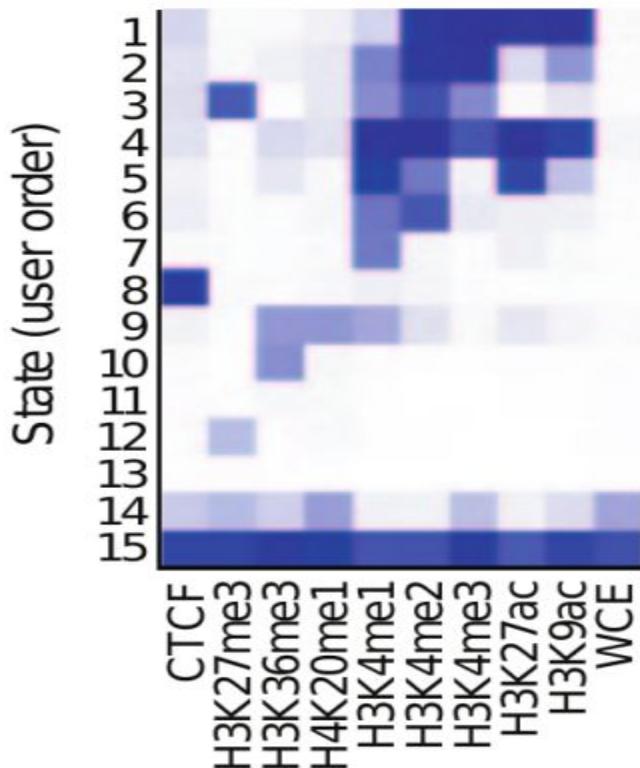
- Derived states, not vectors of chromatin marks -easier to determine genome wide properties.
- Can train on one set and apply to another.

Disadvantages:

- How many states?
- Histone states binary -lose information (except in EpiCSeq)
- Causality unknown

Chromatin Colours

Emission parameters



Candidate state annotation

Active promoter
Weak promoter
Inactive/poised promoter
Strong enhancer
Strong enhancer
Weak/poised enhancer
Weak/poised enhancer
Insulator
Transcriptional transition
Transcriptional elongation
Weak transcribed
Polycomb repressed
Heterochrom; low signal
Repetitive/CNV
Repetitive/CNV

Visualizing Chromatin Marks

