Regulatory Genomics II: Introduction to Epigenomics

Shamith Samarajiwa

Group Leader (Computational Biology & Data Science)
MRC Cancer Unit,
University of Cambridge.

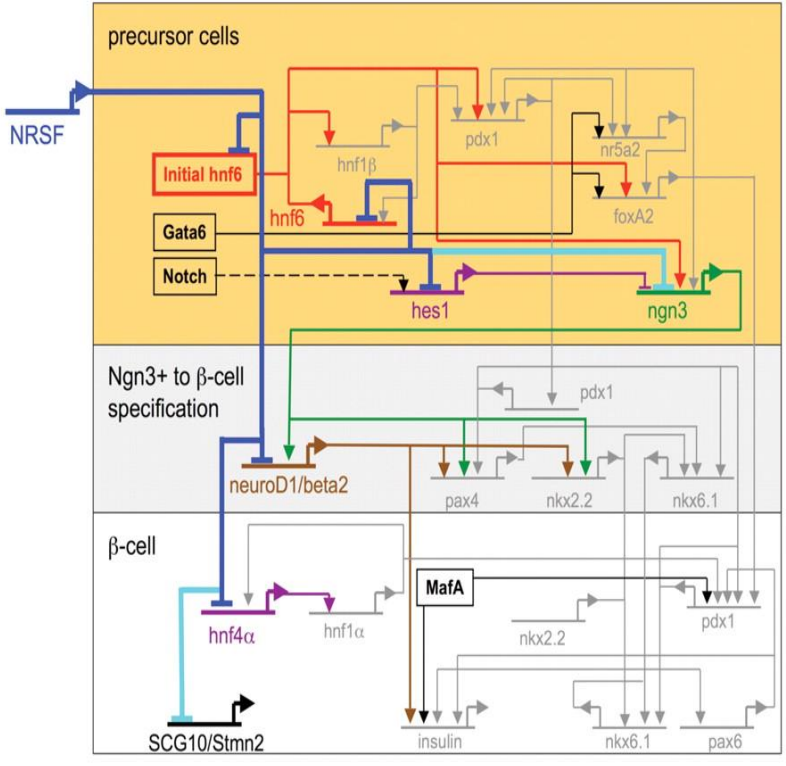
Computational Biology MPhil: Functional Genomics Lectures
February 24th 2021

Overview

- Cellular Regulomes
- Detecting open chromatin
- Epigenomics
- Functional transcriptional regions
- Chromosomal Interactions
- Chromatin Structure

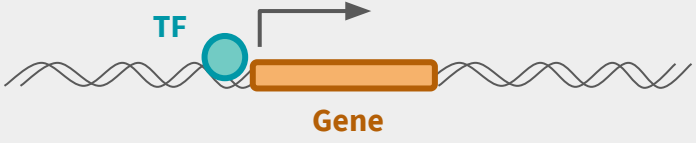
Regulomes: from target genes to networks

- Not all TF binding sites (cistrome) 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 deconvolute the networks underlying phenotypes.
- These networks provide information on connectivity, information flow, and regulatory, signaling and other interactions between cellular components.



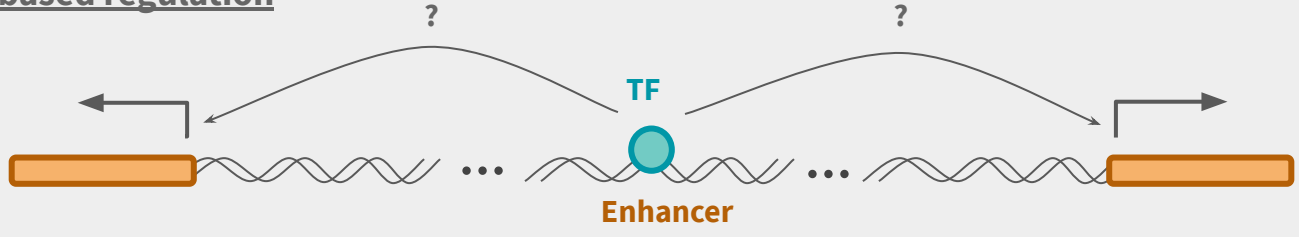
Detection of TF Direct Target

Promoter based regulation



- TF binding with **ChIP-seq**
- Gene expression with **RNA-seq/microarray**
- 3D architecture with **Hi-C**
- Regulatory element activity with **Histone ChIP-seq**

Enhancer based regulation



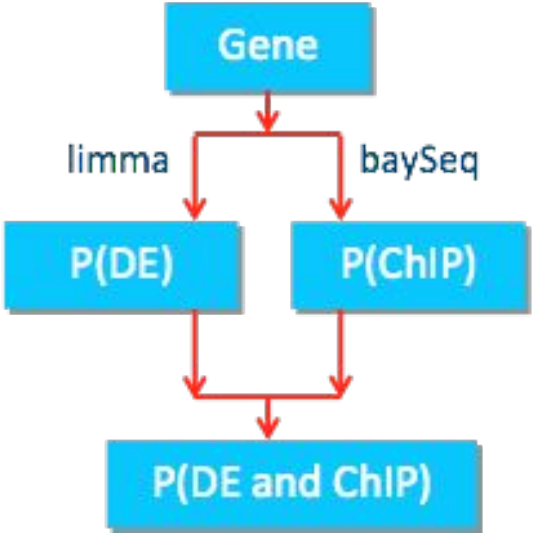
- Rcade (Bioconductor)
- COBRA
- Beta

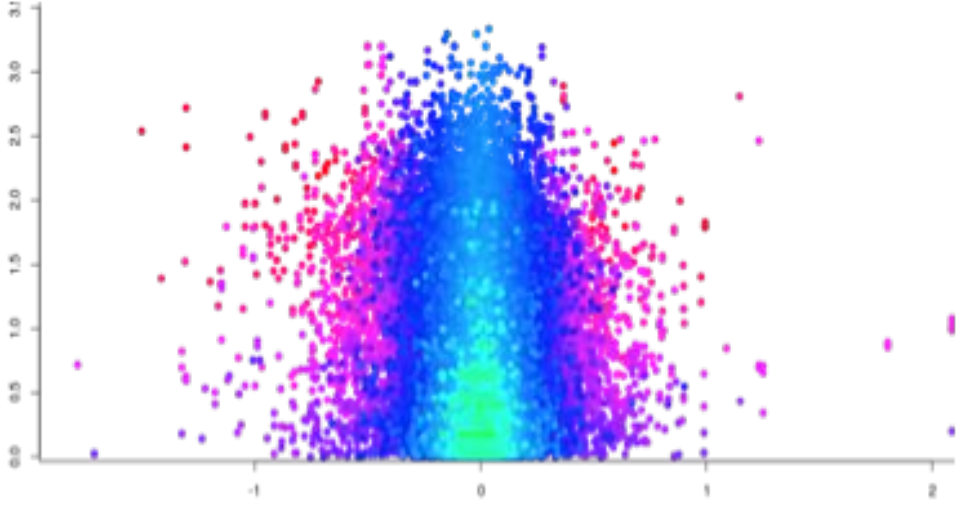
Rcade (R-based analysis of ChIP-seq And Differential Expression)

- 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.
- Rcade can infer 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.





Rcade

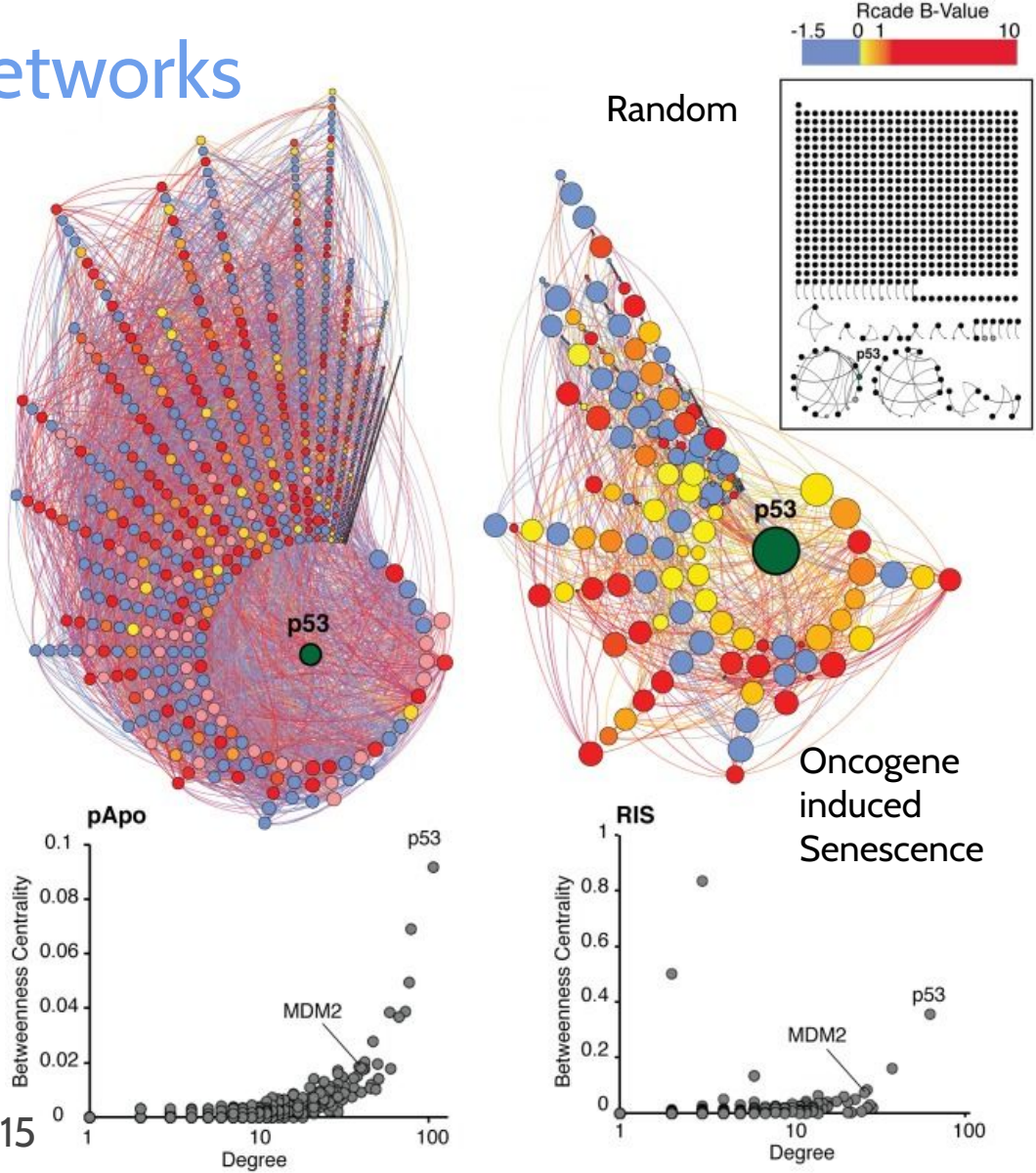
- Rcade integrates posterior probabilities of binding (determined via the baySeq package) with those of differential expression (determined via the limma package).

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

- 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.

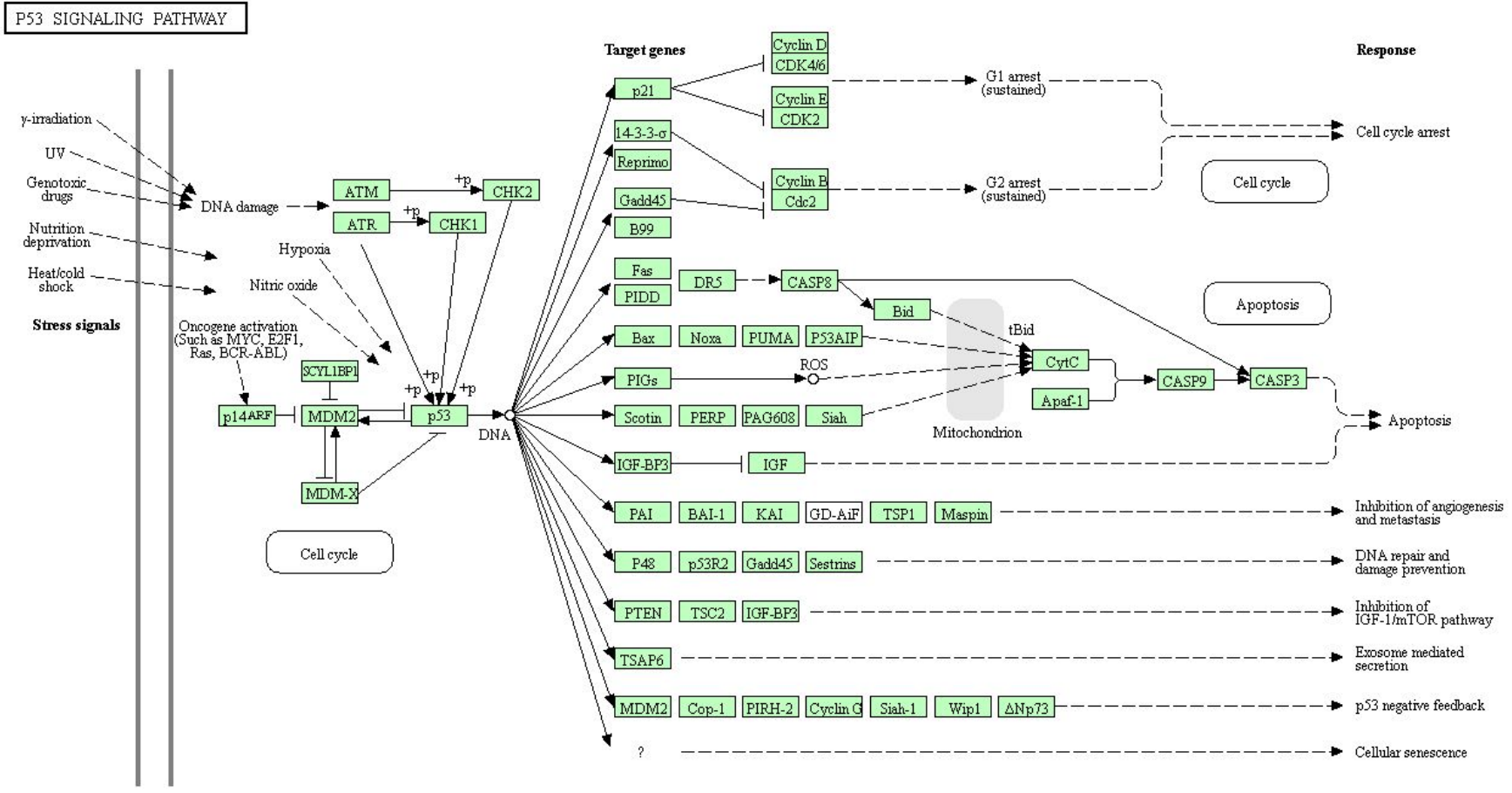
Functional Association Networks

Semi Supervised Network Generation and Topology analysis

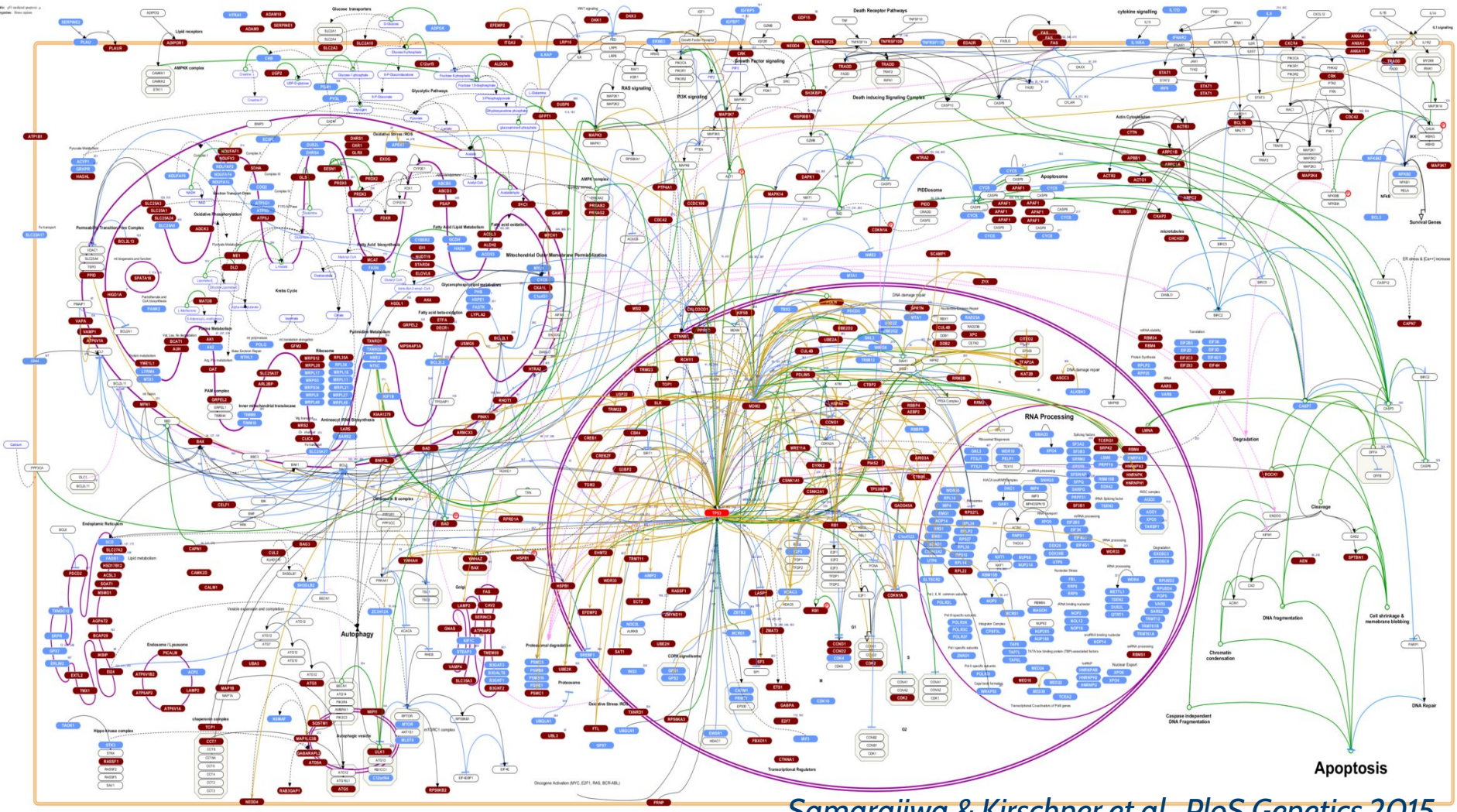


Kirschner & Samarajiwa et al., PloS Genetics 2015

KEGG: p53 signalling pathway

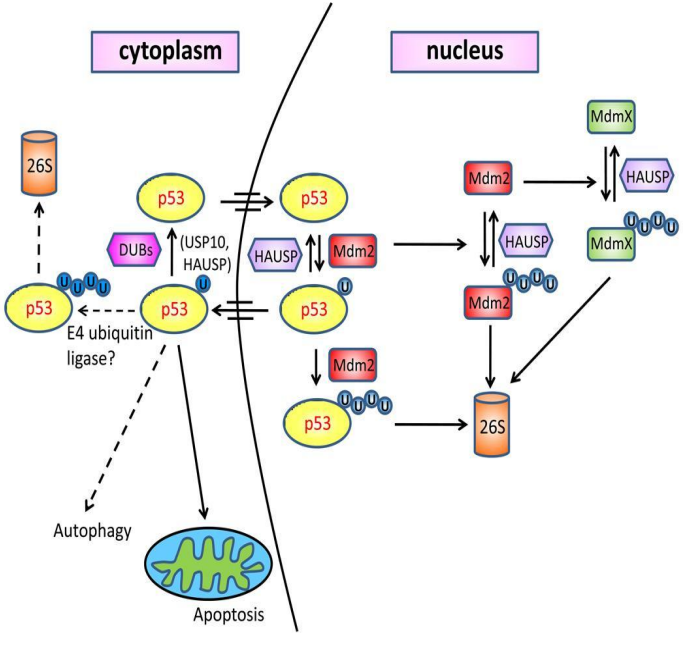


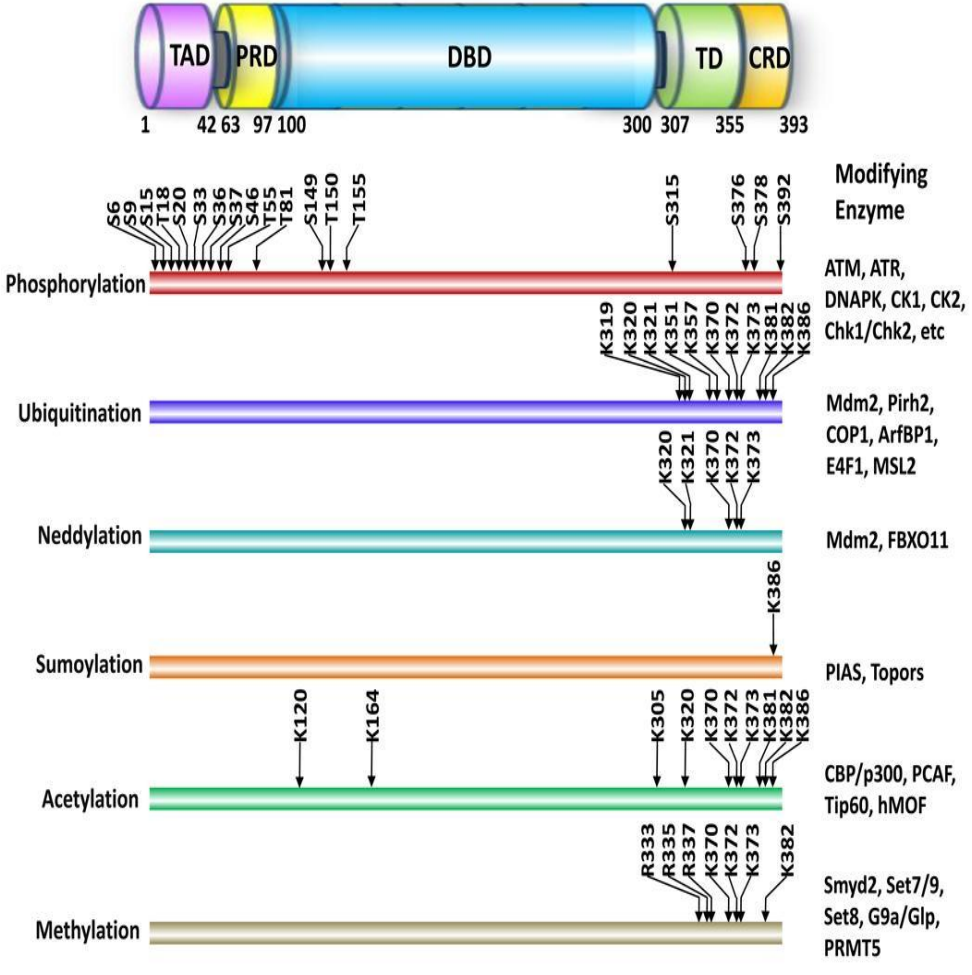
The TP53 Regulome



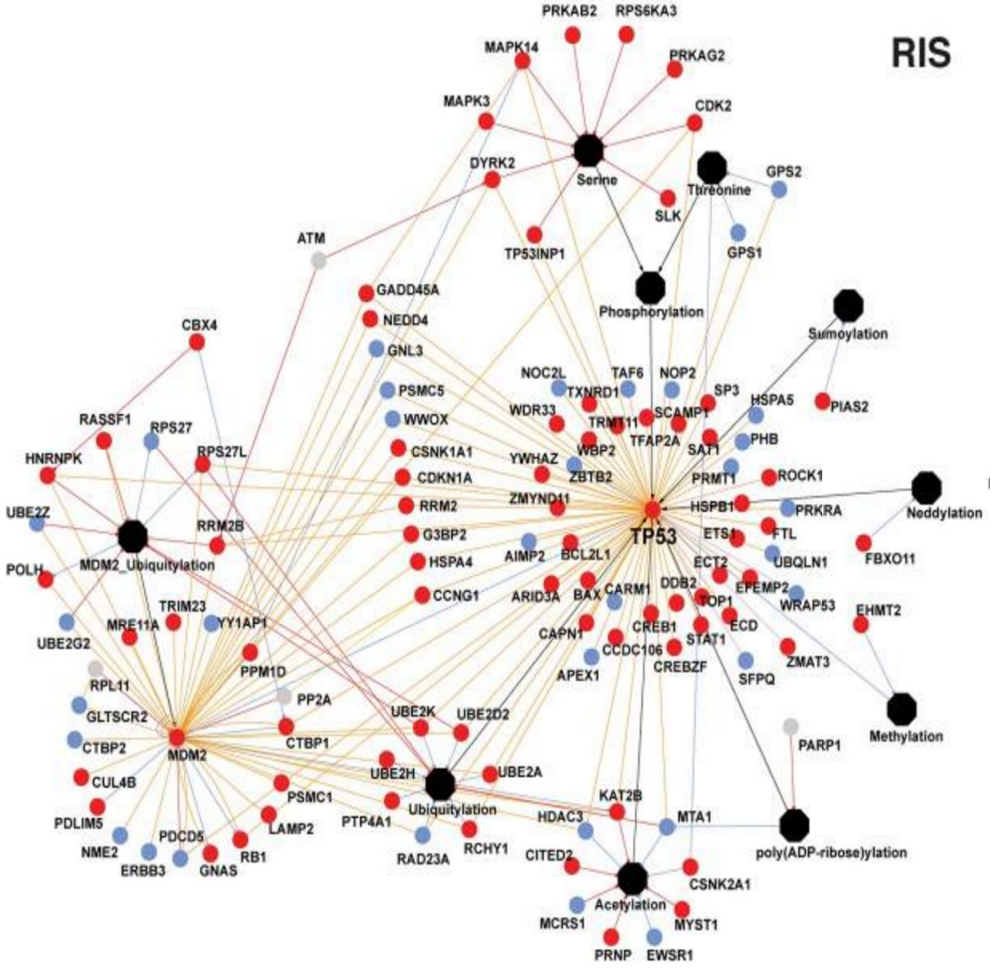
Samarajiwa & Kirschner et al., PloS Genetics 2015

Fine tuning regulation: post-translational modifications





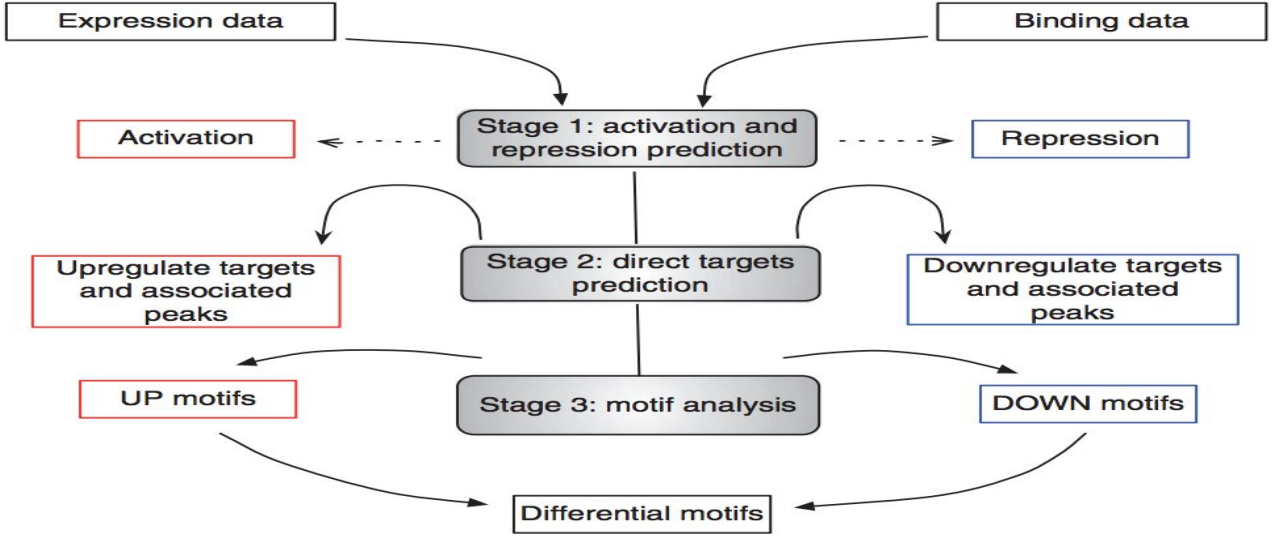
The Self-Regulatory TP53 Network



Samarajiwa & Kirschner et al., 2015, PloS Genetics

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



Wang, 2013 Nat Protoc. 2013

Introduction to Epigenomics

- Epigenetics and Epigenomic
- Open Chromatin with ATAC-seq
- Chromosomal Territories
- Chromatin organization
  - Open Chromatin and Transcription
- Histone modifications
  - Impact on gene regulation
- Epigenomic codes
- 3D Architecture of chromatin
  - Chromatin interactions with Hi-C (Hubs, Loops and EPIs)
  - A-B compartments
  - TADs

Epigenetics and Epigenomics

- Epigenetics encompasse 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 plays a role in gene regulation. Chromatin states can vary between conditions, cells and tissue types and even within a single chromosome.
- 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.

Schulz, PNAS 2010

Large-scale epigenomic studies

ENCODE ENCYCLOPEDIA OF DNA ELEMENTS

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

BLUEPRINT epigenome

Epigenomes of 100 blood cell types

ROADMAP epigenomics PROJECT

Stem cells, fetal tissues, adult tissues

DEEP Deutsches Epigenom Programm

Various human, mouse tissues

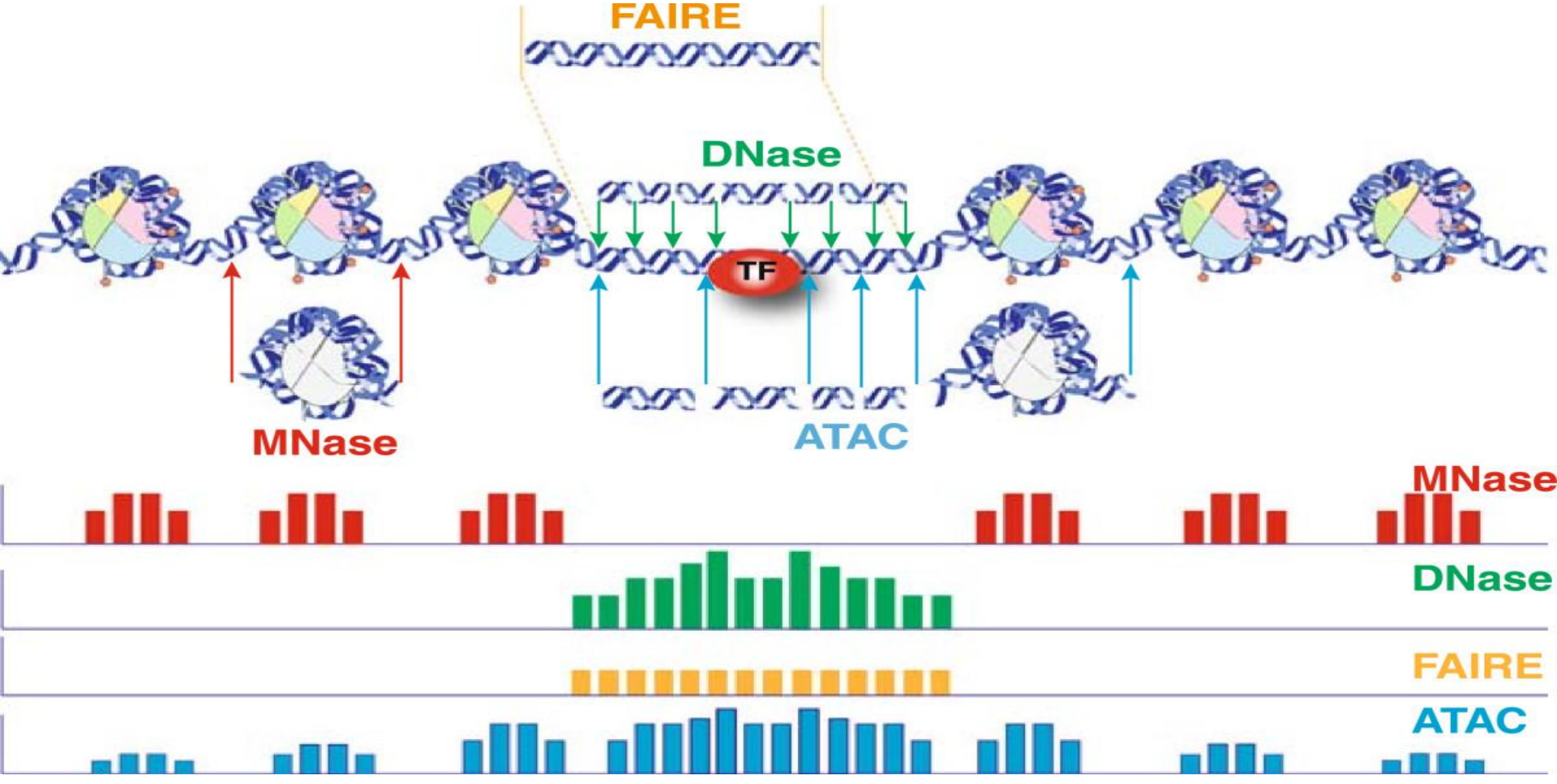
IHEC International Human Epigenome Consortium

~1000 human epigenomes

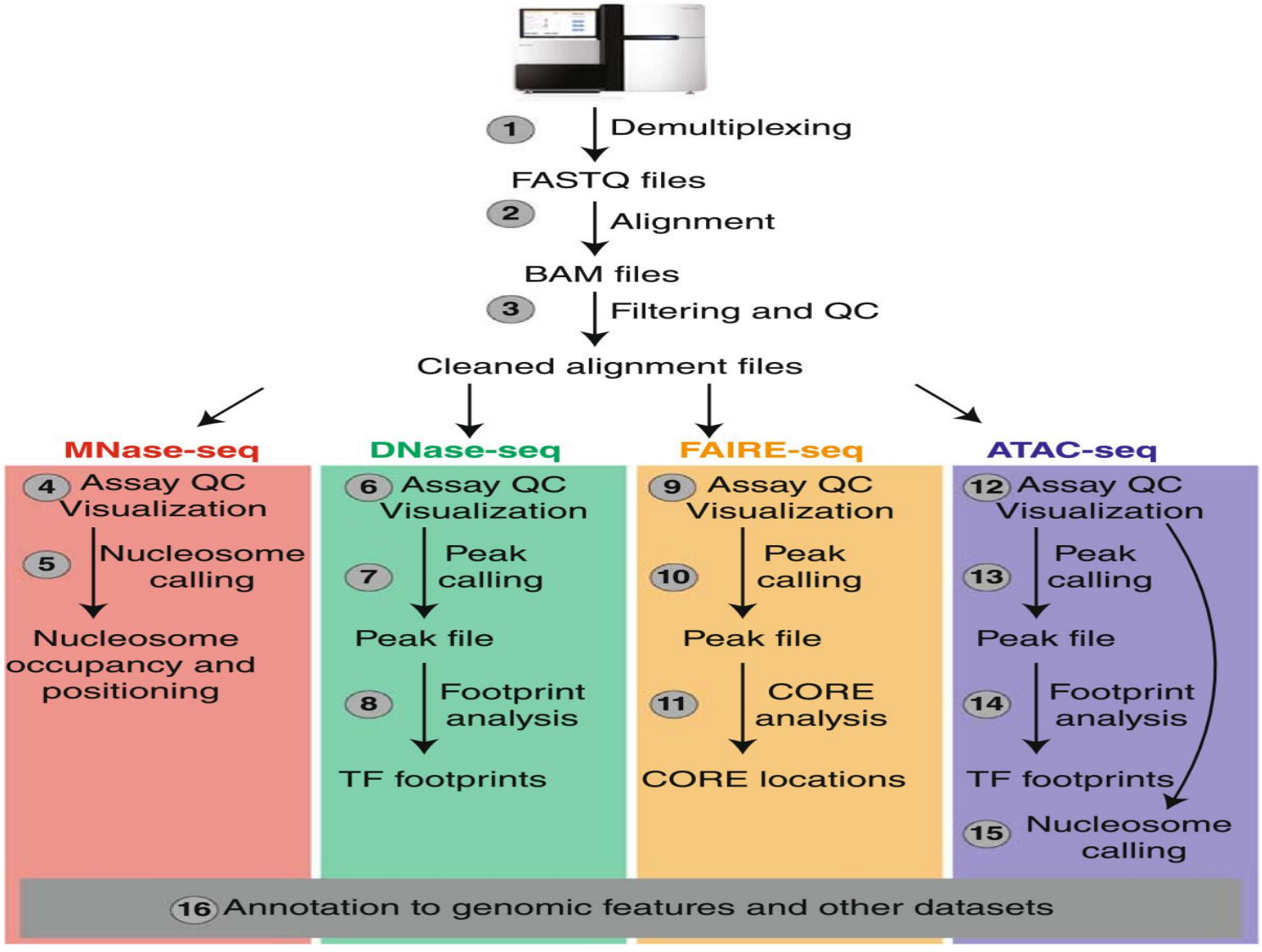
HEP Human Epigenome Project

Methylomes

Detecting Chromatin Accessibility

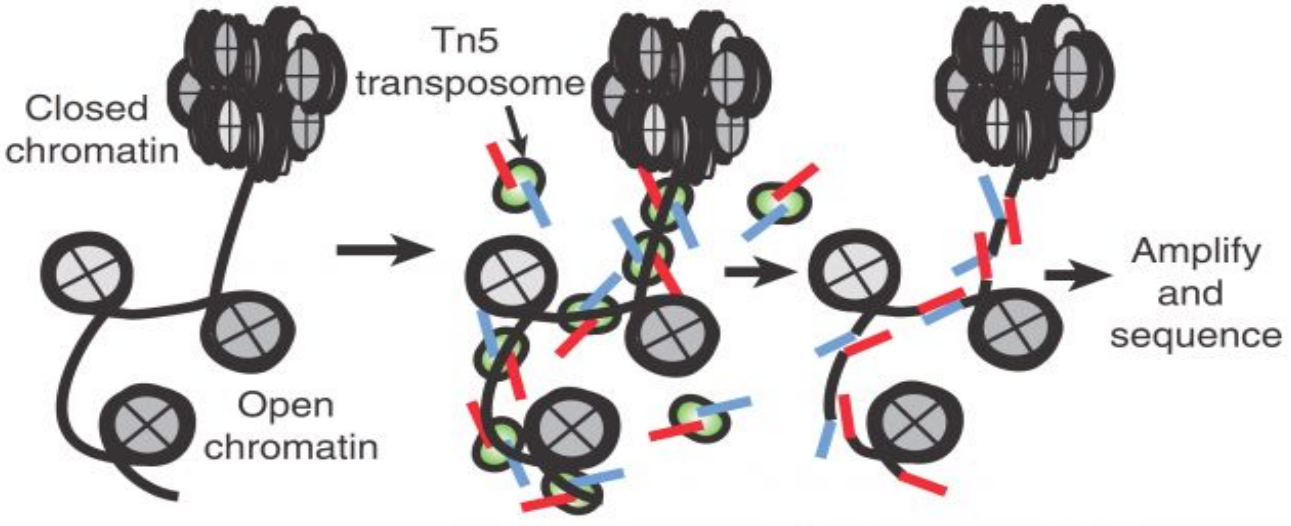


Tsompana and Buck, *Epigenetics & Chromatin* 2014



Tsompana and Buck, *Epigenetics & Chromatin* 2014

ATAC-seq



- Enables measurement of chromatin accessibility and nucleosome free regions genome wide.
- Does not require antibodies or tags that can introduce potential bias.
- Hyperactive Tn5 transposase is used to fragment DNA and integrate into open chromatin regions.
- During ATAC-seq, ~50,000 unfixed nuclei are tagged *in vitro* with sequencing adapters by purified Tn5 transposase.
- Can also detect nucleosome packing, positioning and TF footprints.

JD Buenrostro et al, Nature Methods, 2013.

Differences from ChIP-seq data processing

- Remove mitochondrial reads
  - A large fraction of ATAC-seq reads map to mitochondrial genome (up to 40-60%) that you will want to remove
  - Blacklisted regions contain the mitochondrial genome

- Normalisation across samples might be needed
  - Efficiency of the ATAC-seq protocol in assaying open regions might be different based on how much transposome gets into nuclei
  - For a normalisation solution see: *Denny et al, Cell, 2016.*
  - Alternatively *THOR* can be used for normalization.

ATAC-seq peak calling

- MACS2 was designed for ChIP-seq and not optimized for ATAC-seq.
- Use the fragment length for smoothing when calling peaks with MACS2
  - MACS2 documentation says when using DNase-seq type data:
    - "... all 5' ends of sequenced reads should be extended in both direction to smooth the pileup signals. If the smoothing window is 200bps, then use '--nomodel --shift -100 --extsize 200"
  - --nomodel: don't build shifting model
  - --shift: when this value is negative, ends will be moved toward 3'->5' direction
  - --extsize: extend reads in 5'->3' direction to fix-sized fragments
  - Use the fragment size for smoothing - you can calculate it with ChIPQC
  - If you have paired end reads use BAMPE function in MACS2 for peak calling.

- HMMRATAC is a custom peak caller for ATAC-seq

Chromosome territories

- **"Chromosomal Territory"** model - ***Theodor Boveri*** in 1885 and Carl Rabl in 1909 (Cremer T, Cremer M. 2010. Cold Spring Harb. Perspect Biol 2:1–22)
- These early observations were superseded when electron microscopy showed evidence of chromosome intermingling during interphase.
- **"Spaghetti"** model of the interphase nucleus — chromatin fibers from different chromosomes are interwoven.
- More recently methods such as

1. Fluorescent in Situ hybridization (FISH)
2. Chromosomal Conformation Capture (3C)

demonstrated genome compartmentalization of chromosomes.

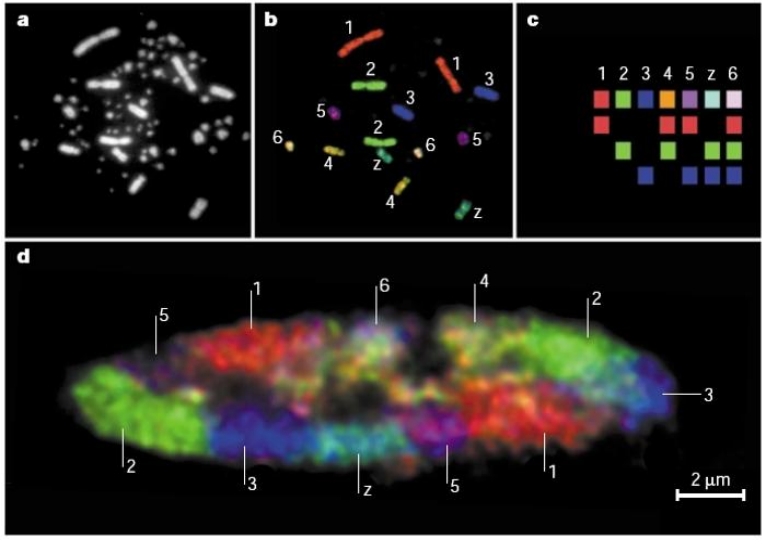
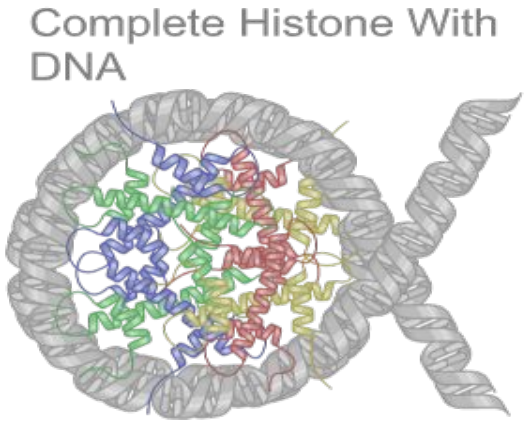
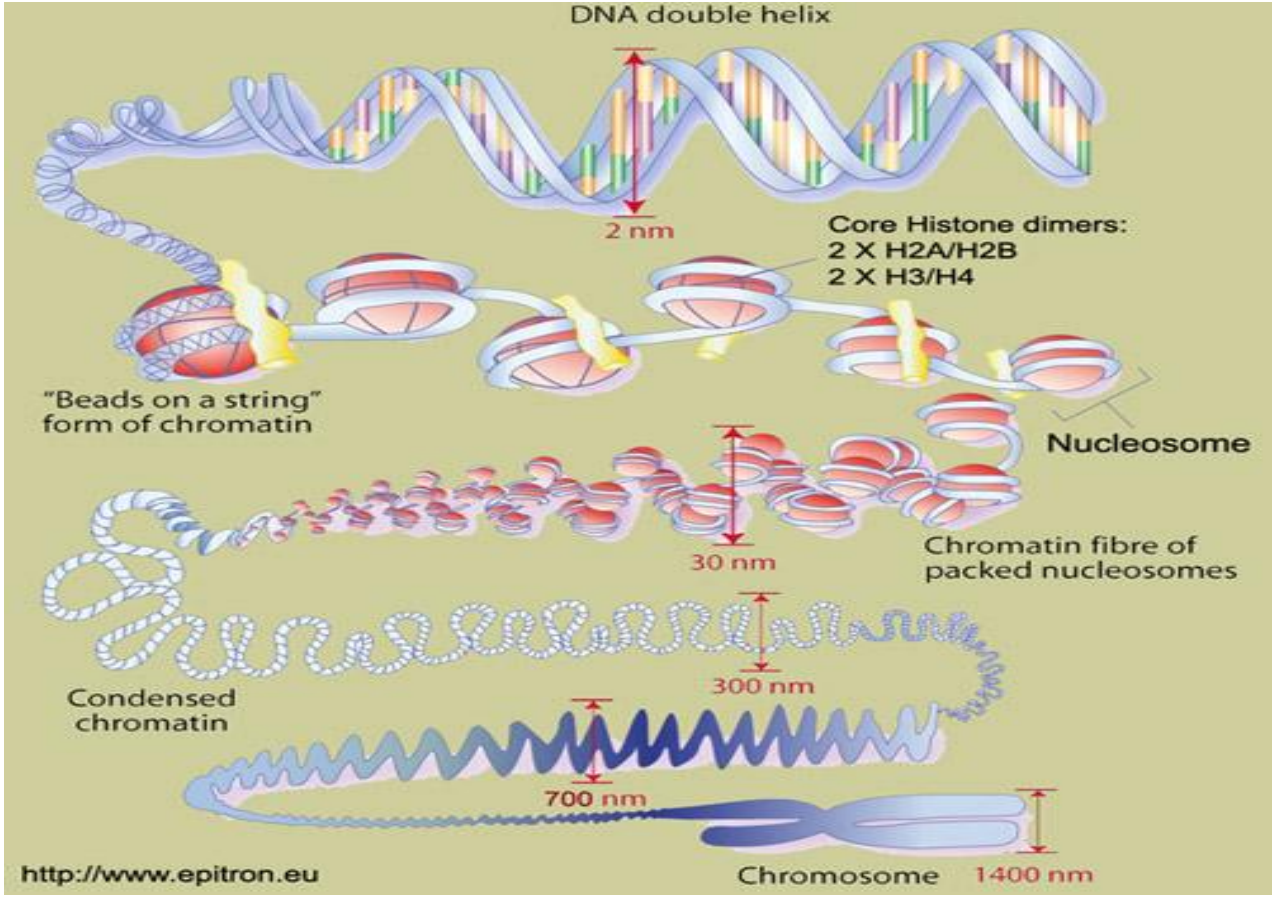


Fig: Multicolour FISH labelled chicken nucleus. (Misteli, T. 2008 Nature Education 1(1):167)

Chromatin Organization

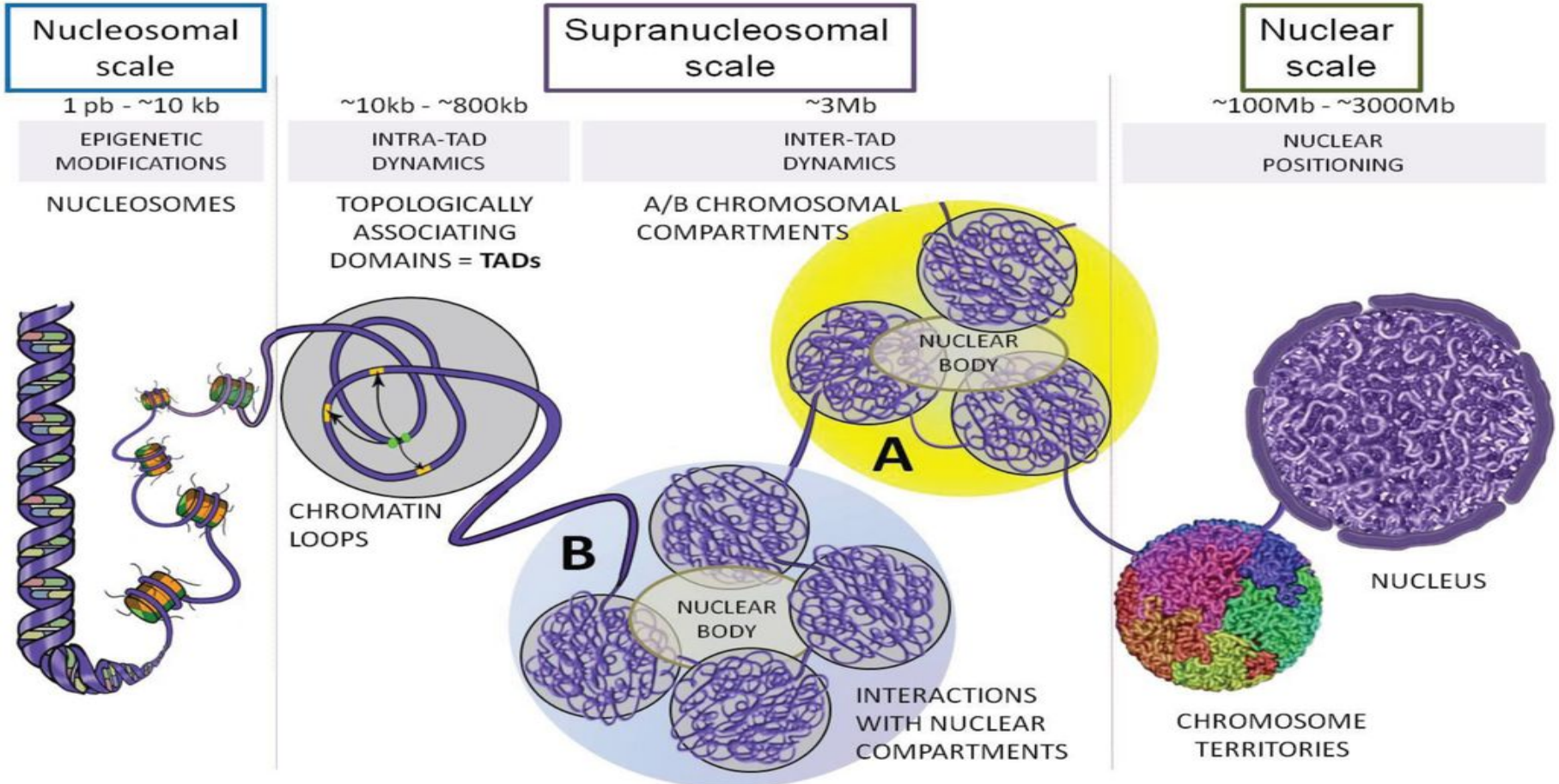
- Histone octamer core wrapped around by 147 bp of DNA and separated by linker DNA.





The DNA double helix wrapped around four histone proteins, in a structure called a nucleosome. By Richard Wheeler (Zephyris) [CC-BY-SA-3.0]]/Wikimedia Commons

Current model of chromatin organization



Ea et al., Genes 2015

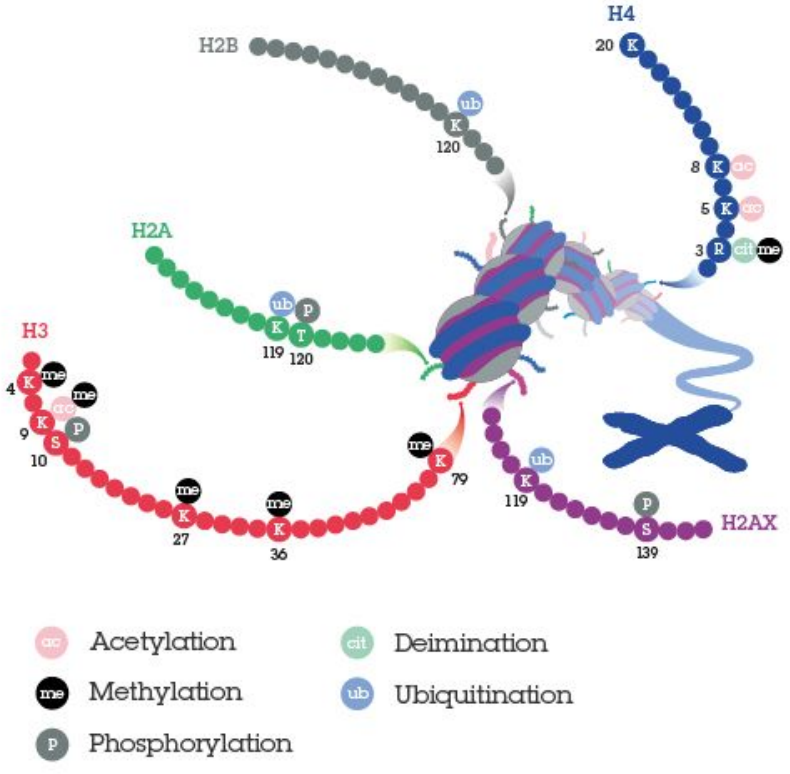
Histone Modifications

- Nucleosomes consist of 2x H2A/H2B and 2x H3/H4 histones.

- 80 known covalent modifications

H3K4me3 →

H3	Histone 3
K	Residue is lysine, K
4	4th residue.
me3	Trimethylation



The most common histone modifications

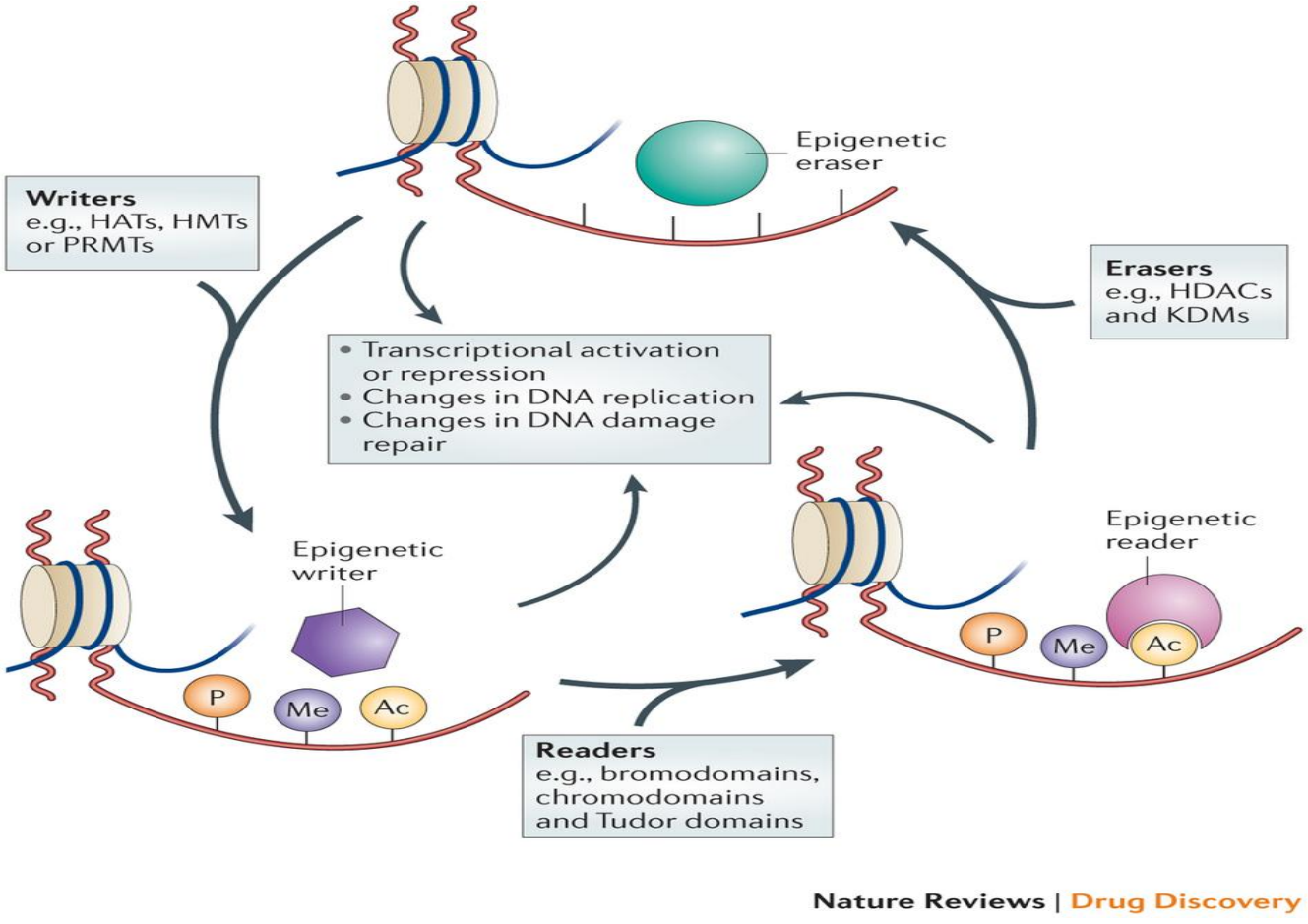
Histone Modifications

Some examples:

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

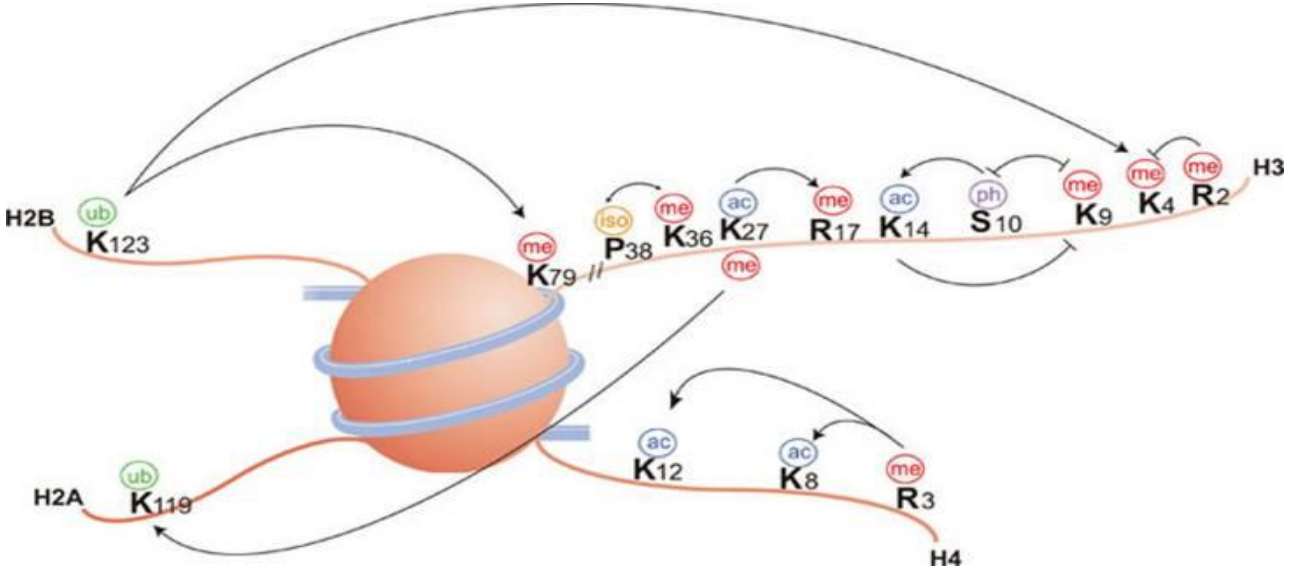
More information at:
http://epigenie.com/key-epigenetic-players/histone-proteins-and-modifications

Epigenetic Readers, Writers and Erasers



Falkenberg and Jonstone, Nature Reviews Drug Discovery, 2014

Combinations of marks can have different effects



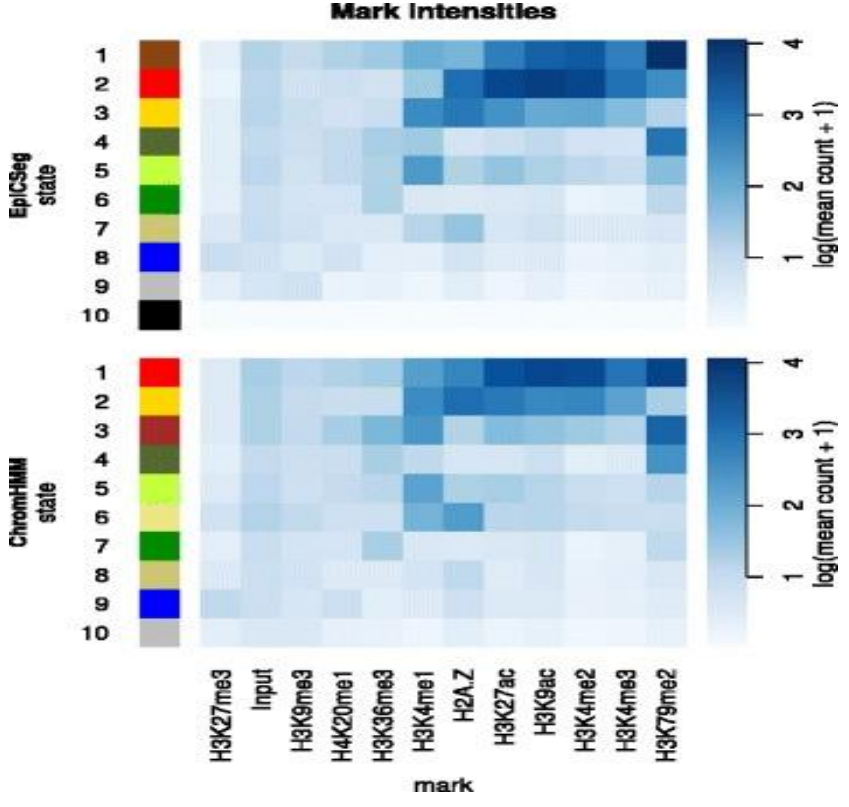
Bannister and Kouzarides (Cell Res. 2011)

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

Simplifying histone marks

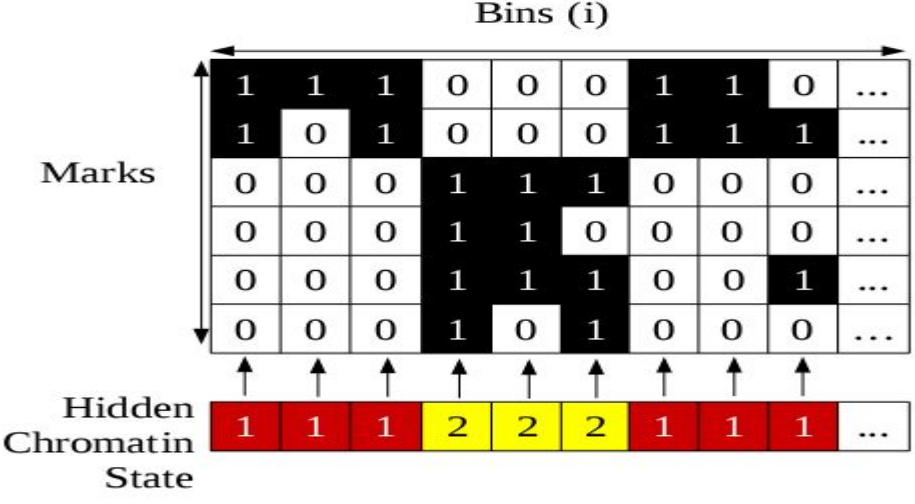
Unsupervised learning methods for *segmentation*;

- 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

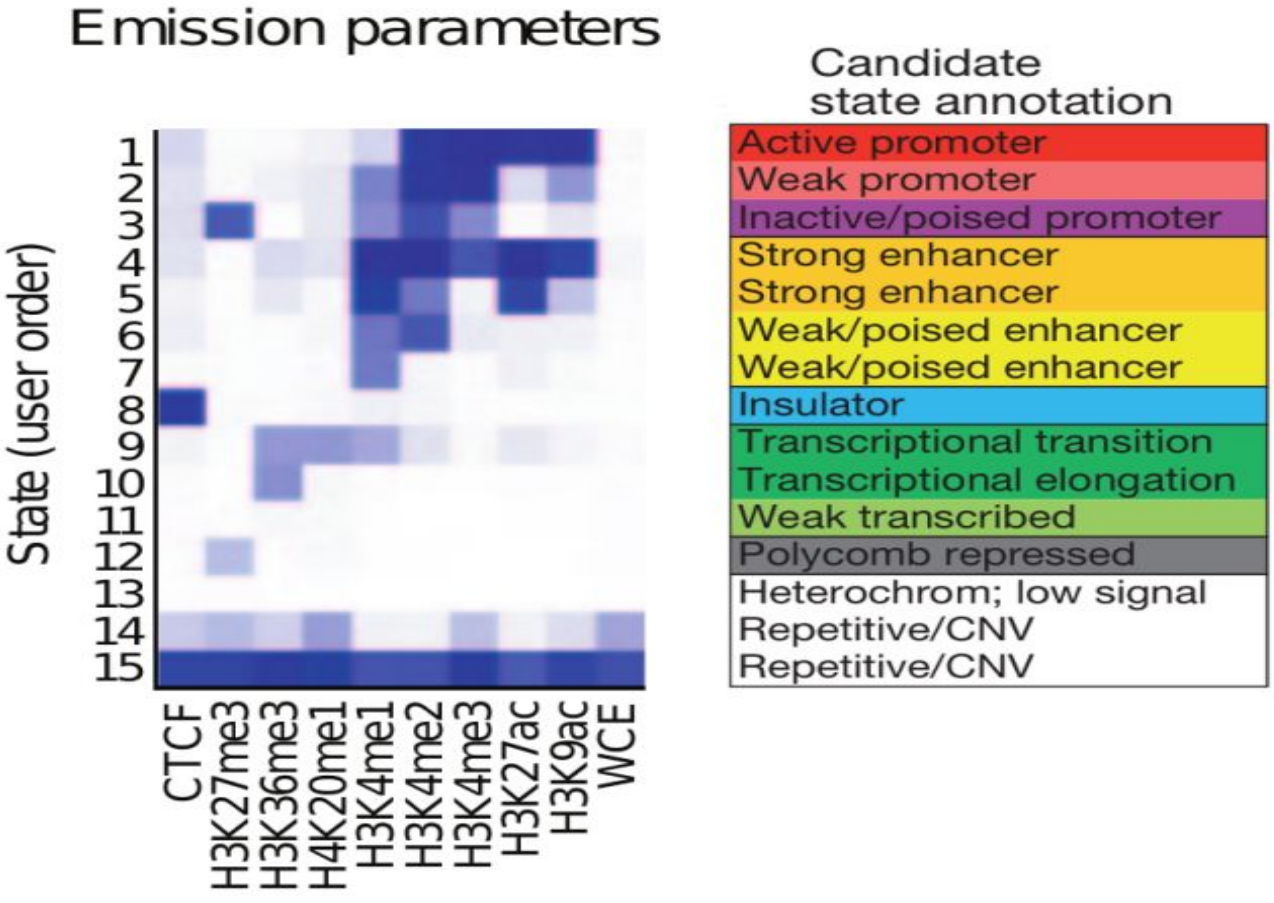
Advantages:

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

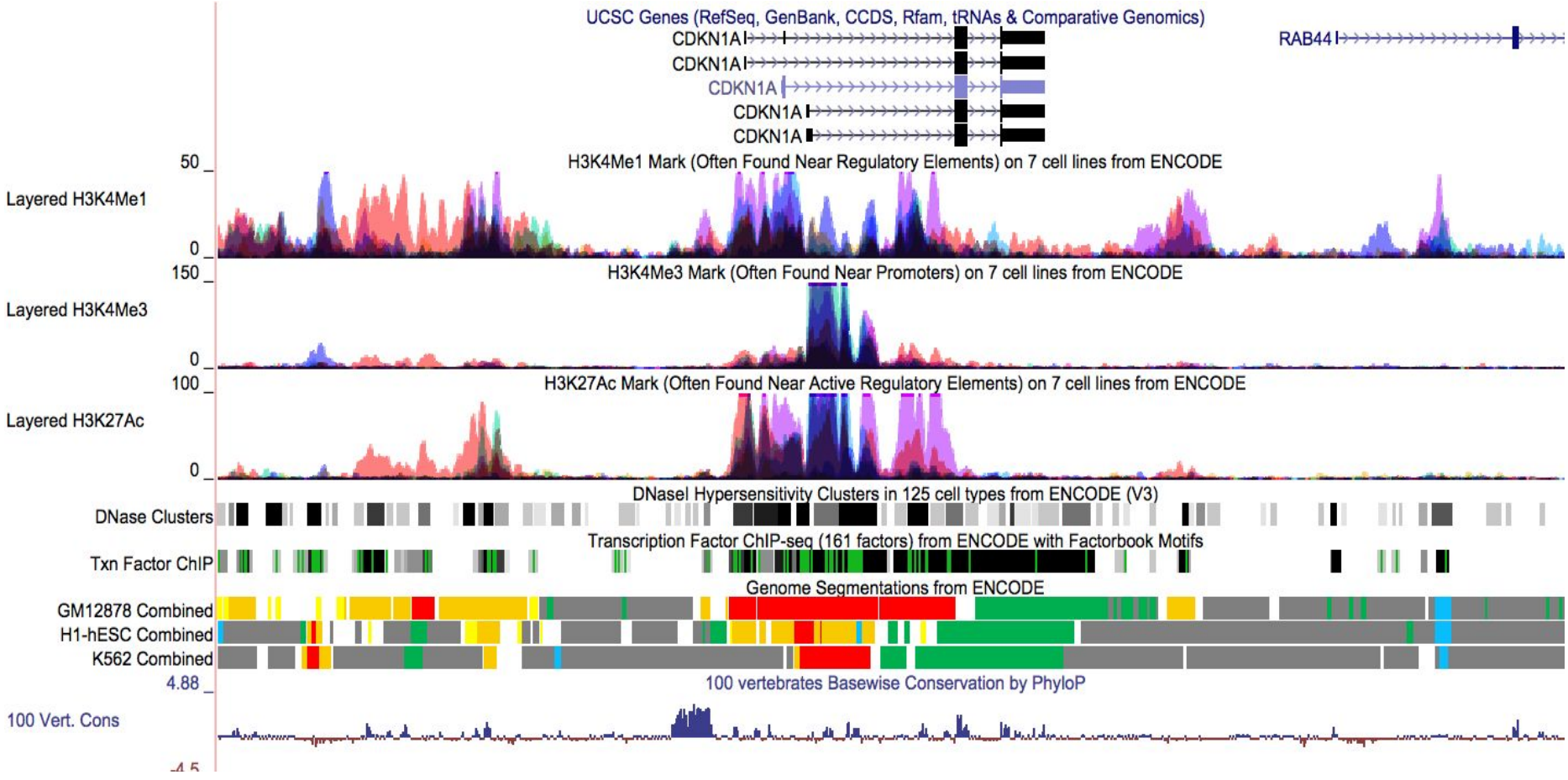
Disadvantages:

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

Chromatin Colours



Visualizing Chromatin Marks



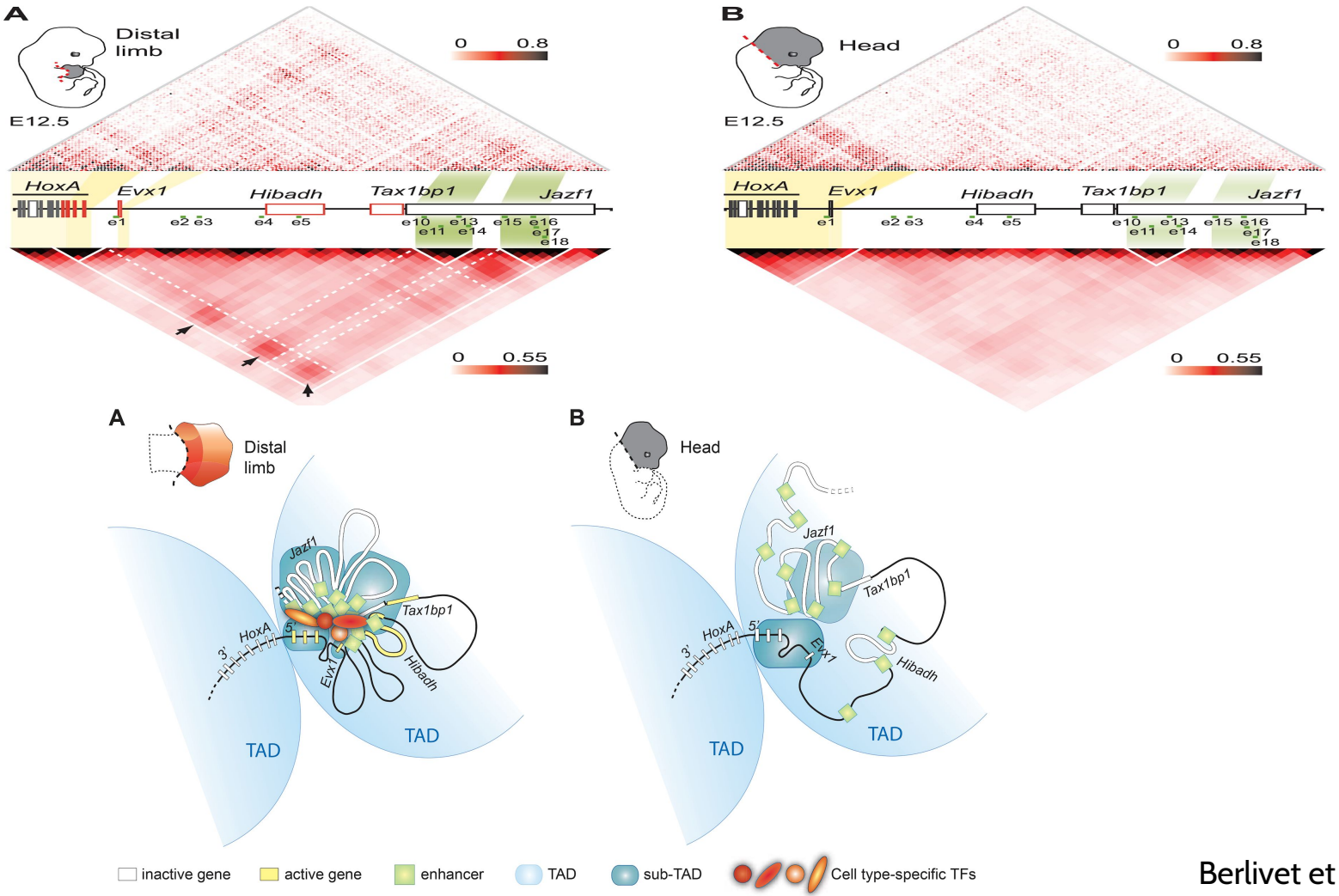
Chromosomal Conformation Capture

- 3C methods identify all possible chromatin interactions between two distinct genomic regions such as promoter-enhancer interactions.

Table 1. Advantages and limits of 3C-derived methods.

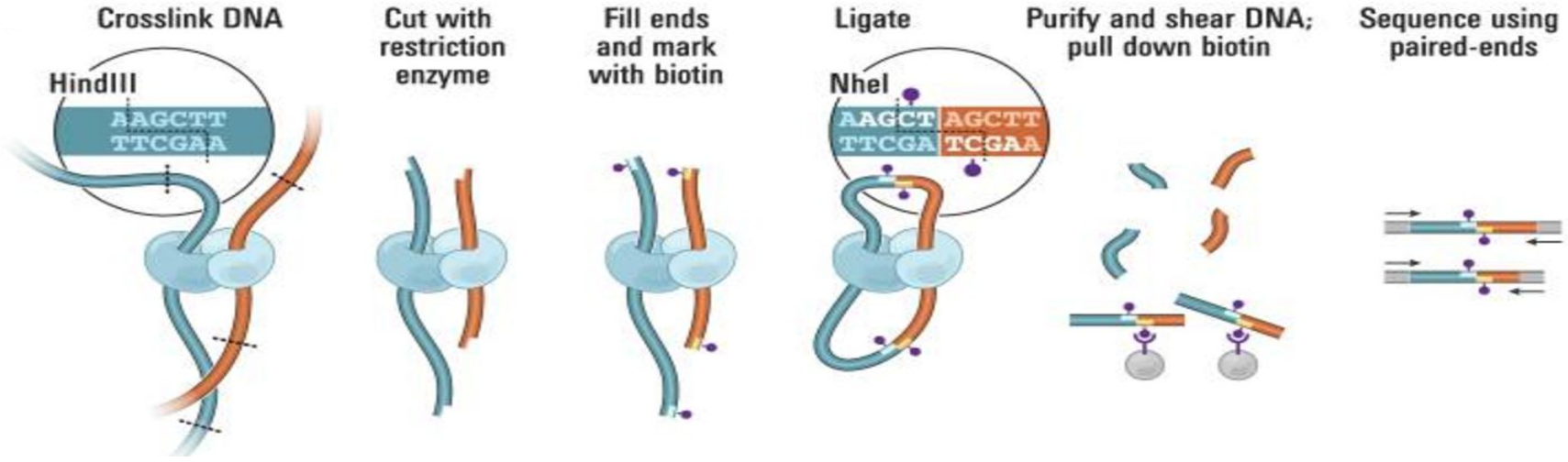
Method	Genomic Scale Investigated	Advantages	Limits
3C-qPCR	~250 kilobases	Very high dynamic range (highly quantitative), easy data analysis	Very low throughput: limited to few viewpoints in a selected region
4C	Complete genome	Good sensitivity at large separation distances	Genome-wide contact map limited to a unique viewpoint (few viewpoints if multiplex sequencing is used)
5C	Few megabases	Good dynamic range, complete contact map (all possible viewpoints) of a specific locus	The contact map obtained is limited to a selected region
Hi-C	Complete genome	Very high throughput (complete contact map)	Poor dynamic range, complex data processing

Tissue-specific regulation of *HoxA* genes

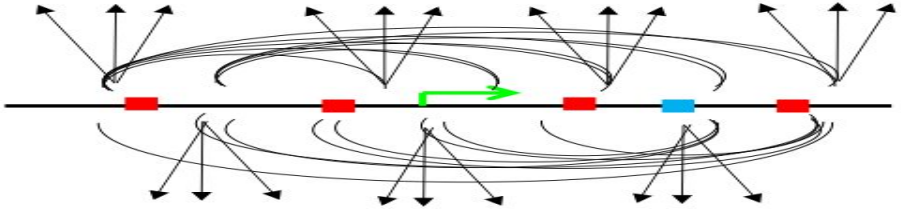


Berlivet et al., 2013

Hi-C: long distance interactions



Lieberman-Aiden, Science 2009



Principle	All against all
Coverage	Genome-wide *
Detection	Paired end HT-sequencing
Resolution	Low *
Limitations	
Examples	All intra- and inter- chromosomal associations

Hi-C: technical and biological biases

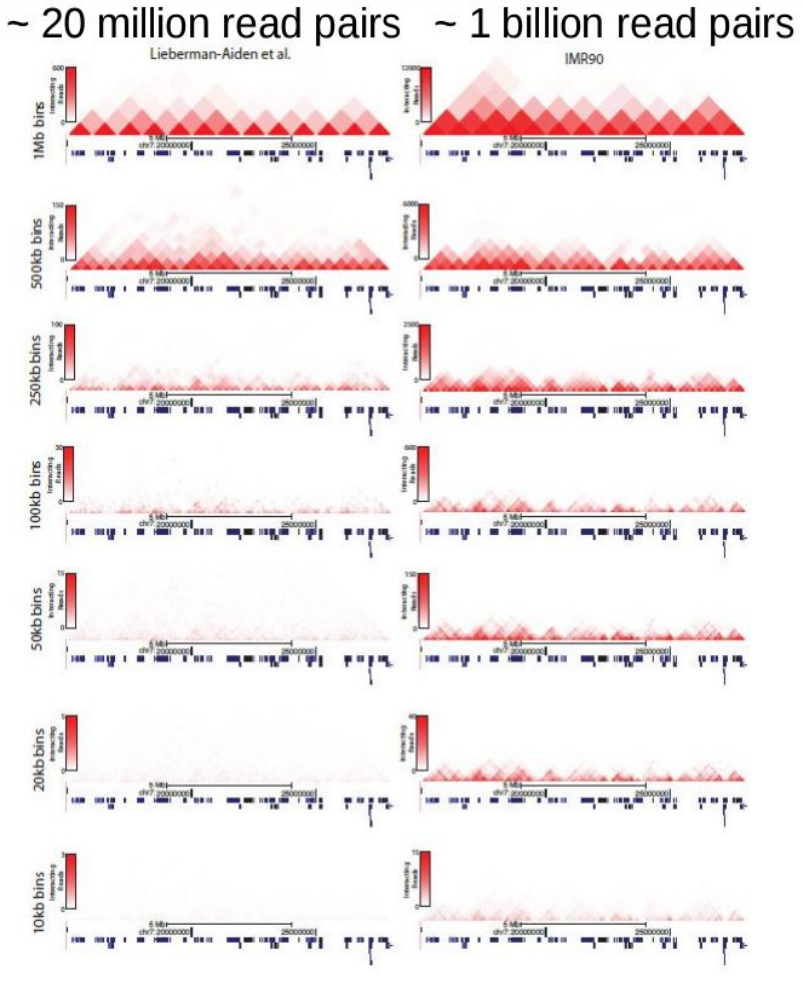
- Hi-C experiments are designed to measure the contact probability between different chromosomal loci on a genome-wide scale.
- This is done by cleaving fixed chromosomes into restriction fragments using six-cutter restriction enzymes and ligating fragment ends to form ligation junctions connecting two loci that are nearby in three-dimensional space.

Biases:

- Hi-C sequence pairs that represent ligation products between nonspecific cleavage sites rather than restriction fragment ends.
- The length of restriction fragments (in other words, the distance between adjacent cutter sites).
- GC bias.
- Mappability (genomic Uniqueness)
- Sequence depth & Coverage.

Yaffe and Tanay, Nat. Genet. 2011

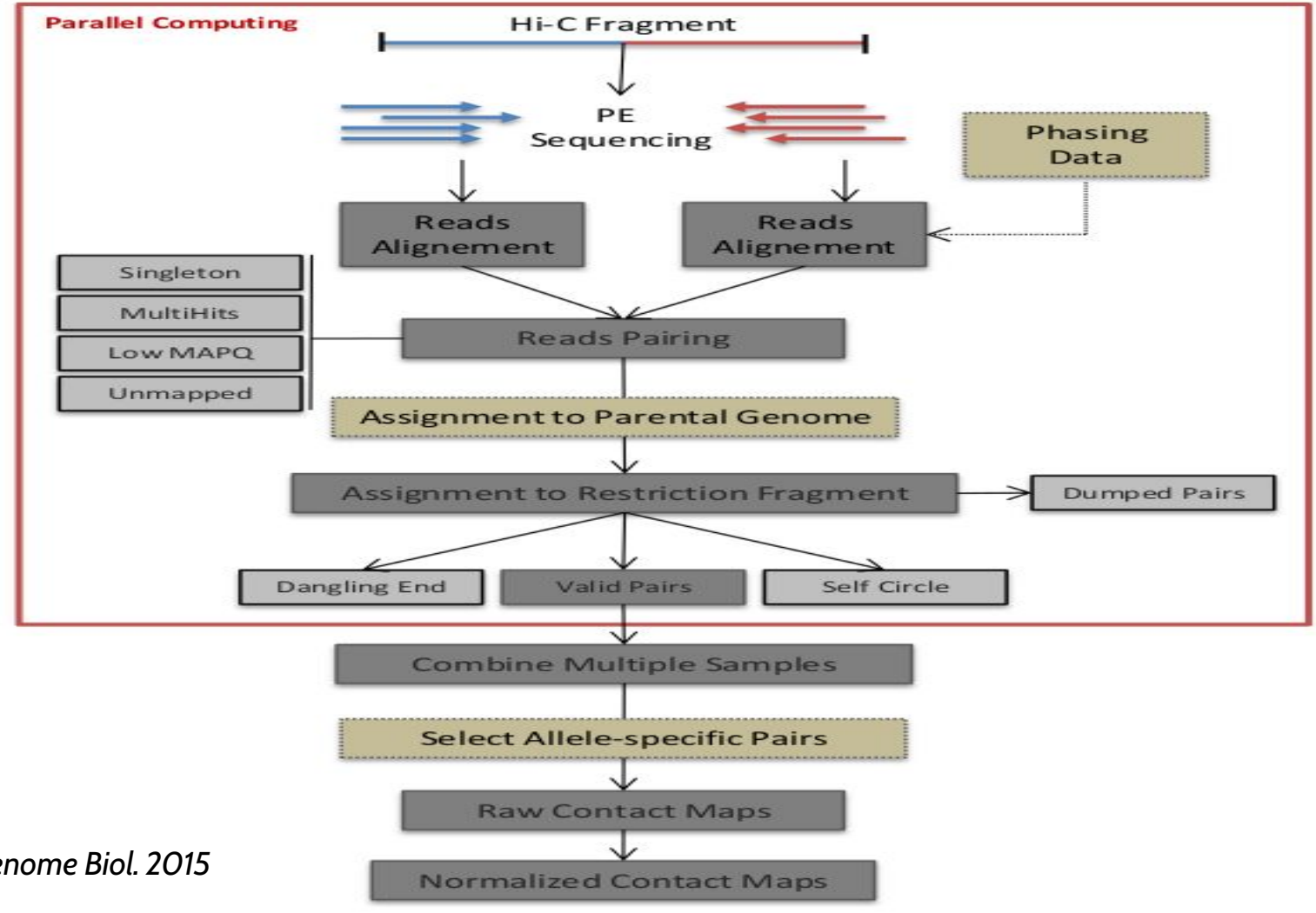
Read depth and bin size



Liebermann-Aiden et al., 2009

Dixon et al., 2012

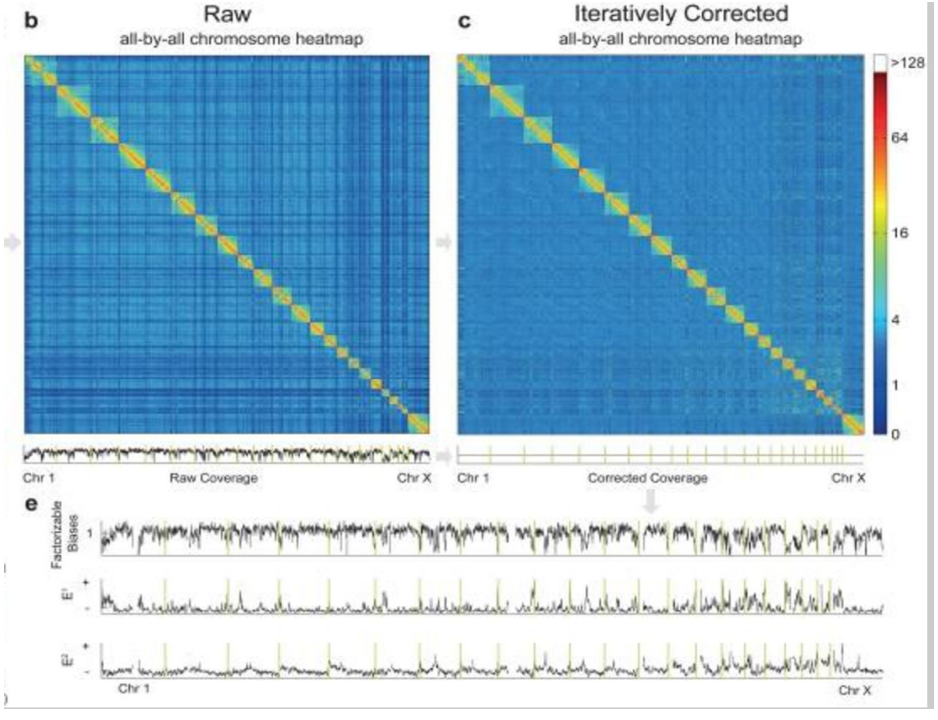
HiC-Pro workflow



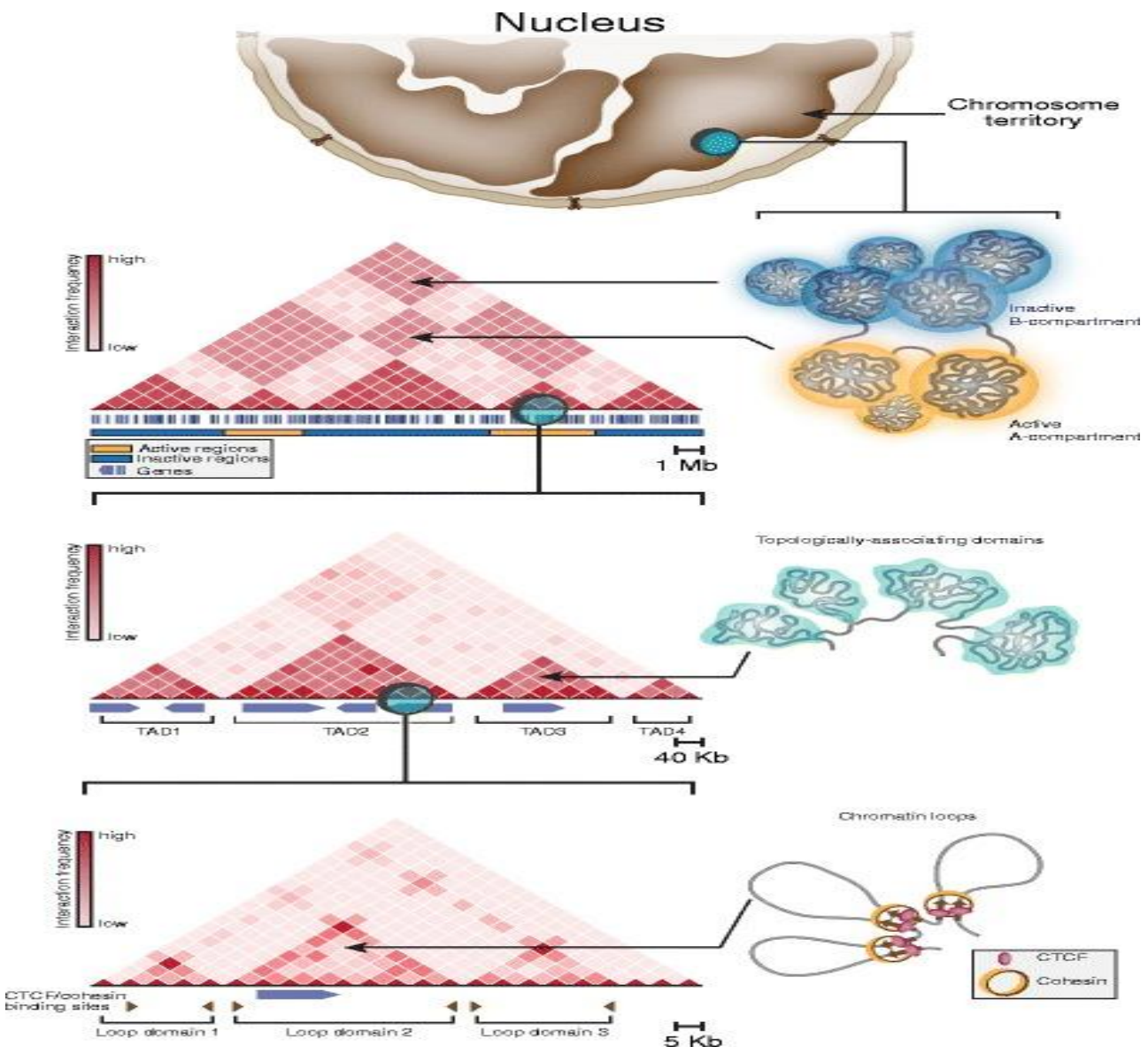
Servant, Genome Biol. 2015

Normalization and bias correction

- **Iterative Correction and Eigenvector decomposition (ICE)**
- A method of iterative correction, which eliminates biases and is based on the assumption that all loci should have equal visibility.
- Iterative correction leverages the unique pairwise and genome-wide structure of Hi-C data to decompose contact matrices into a set of biases and a map of relative contact probabilities between any two genomic loci.
- The obtained corrected interaction maps can then be further decomposed into a set of genome-wide tracks (eigenvectors) describing several levels of higher-order chromatin organization



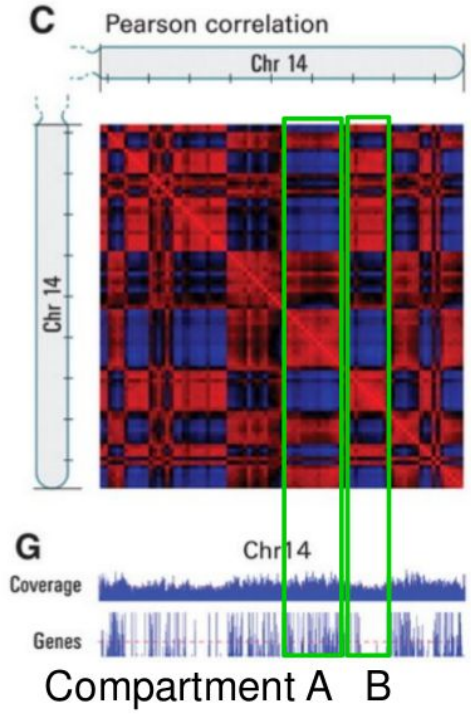
Imakaev et al. 2012


Nucleus
Chromosome territory
Interaction frequency
high
low
Inactive B-compartment
Active A-compartment
Active regions
Inactive regions
Genes
1 Mb
Topologically-associating domains
TAD1
TAD2
TAD3
TAD4
40 Kb
Chromatin loops
CTCF
Cohesin
CTCF/cohesin binding sites
Loop domain 1
Loop domain 2
Loop domain 3
5 Kb

Chromatin: A-B compartments

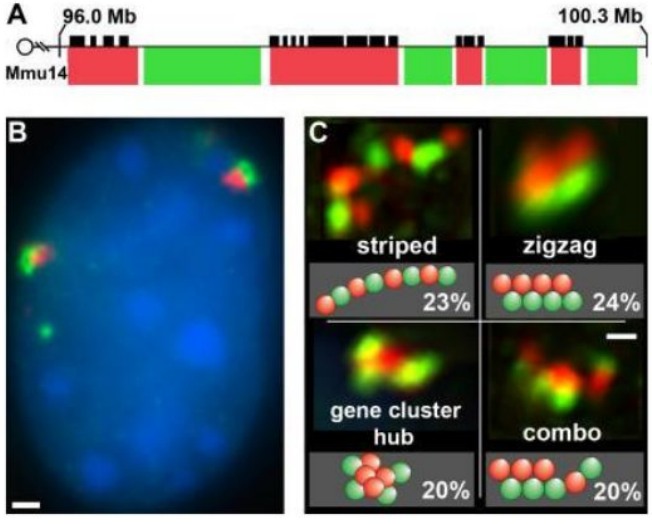
- The genome (at the Mb scale) can be divided into cell type or condition specific A/B compartments that are associated with open and closed chromatin.
- The A compartment is associated with gene rich, transcriptionally active, open chromatin state regions.
- Interactions are more likely between A-A or B-B regions and not A-B regions.
- A and B regions can change into the other.

A-B compartments



Compartment A is gene rich

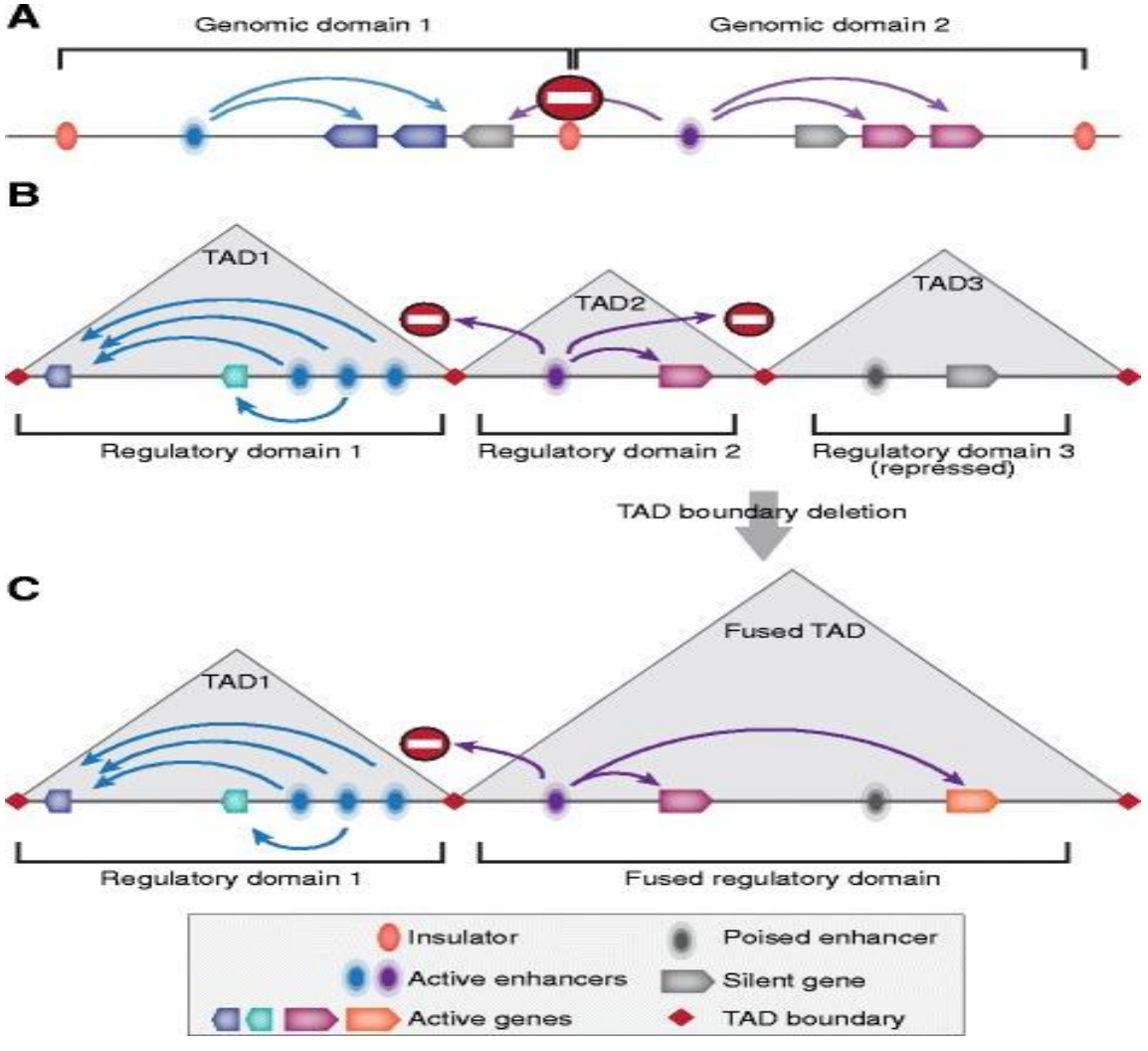
Liebermann-Aiden et al., 2009

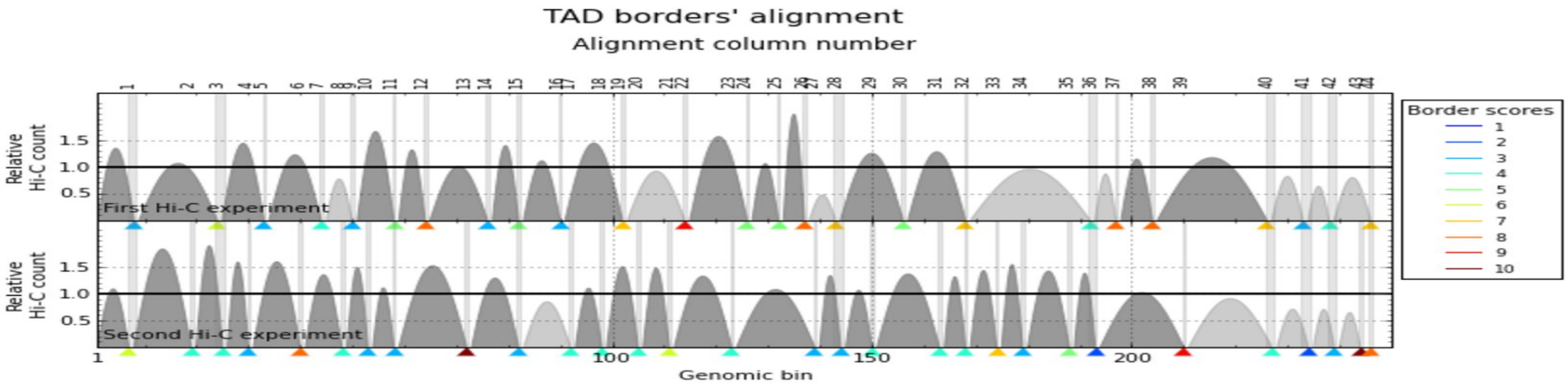


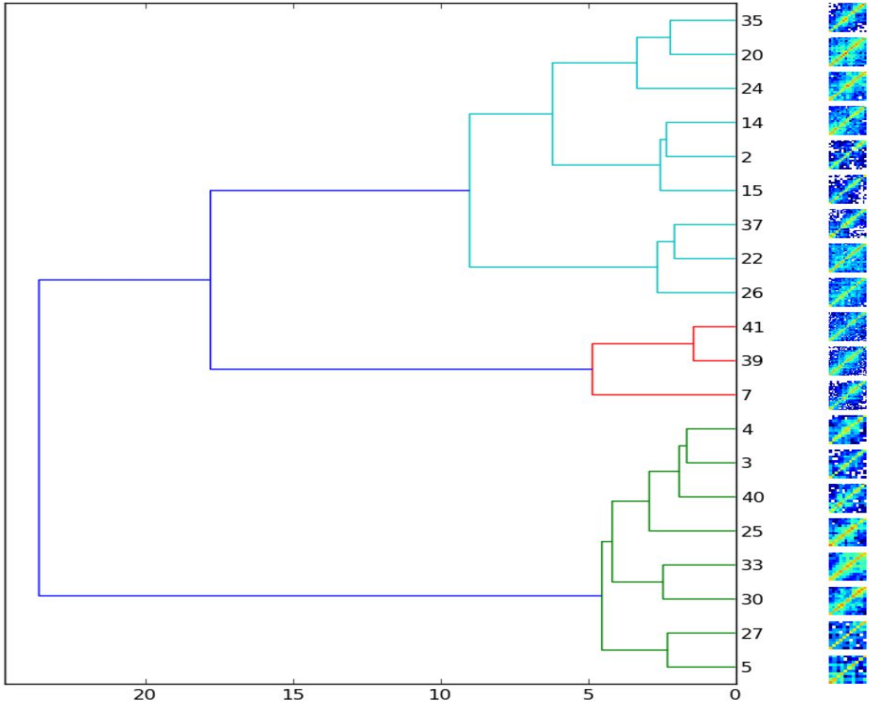
Compartment A
Genes Spearman's ρ = 0.431
Expression Spearman's ρ = 0.476
Accessible chromatin, Spearman's ρ = 0.651
H3K36 trimethylation, Spearman's ρ = 0.601 (active)
H3K27 trimethylation, Spearman's ρ = 0.282 (repressive)
A is more closely associated with open, accessible, actively transcribed chromatin.

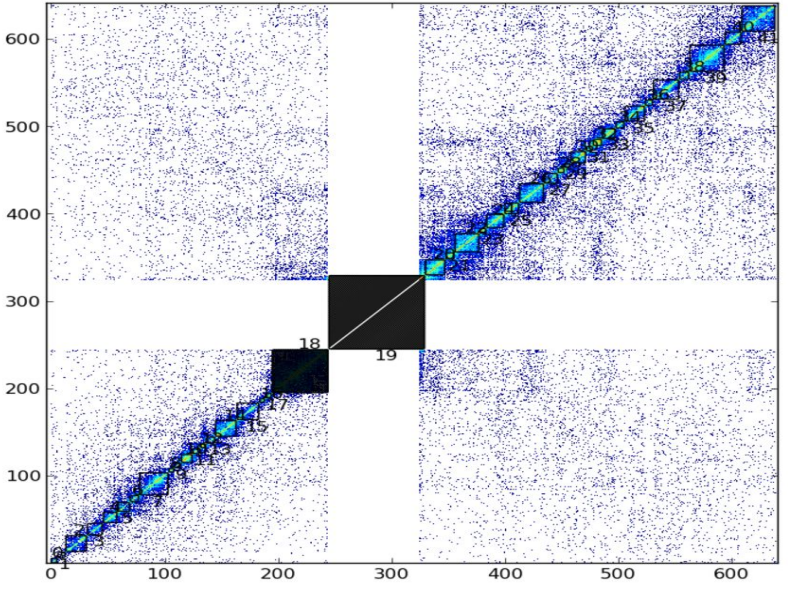
Shopland, et al 2006

Topologically Associated Domains




TAD borders' alignment
Alignment column number
Relative Hi-C count
First Hi-C experiment
Second Hi-C experiment
Genomic bin
Border scores





Deep Learning Chromatin Architecture

Identifying regulatory and spatial genomic architectural elements using cell type independent machine and deep learning models

Laura D. Martens, Oisín Faust, Liviu Pirvan, Dóra Bihary, Shamith A. Samarajiwa

doi: https://doi.org/10.1101/2020.04.19.049585

- Predict chromatin interactions and TAD boundaries using DNN.
- Models trained on Hi-C and Histone modifications, then given just Histone modification ChIP-seq predict Highly Interacting Chromatin Regions (HICRs) and TAD boundaries with high accuracy.
- HICRs turned out to be Enhancer-Promoter interactions
- Models are cell type independent.

Martens et al., 2020 bioRxiv