Introduction to Epigenomics

Shamith Samarajiwa

University of Cambridge

CRUK Autumn School in Bioinformatics
September 2017

UNIVERSITY OF CAMBRIDGE

MRC Cancer Unit

Introduction to Epigenomics

- Epigenetics and Epigenomic
- Chromosomal Territories
- Chromatin organization
    - Open Chromatin and Transcription
- Histone modifications
    - Impact on gene regulation
- Epigenomic codes
- 3D Architecture of chromatin
    - Long distance chromatin interactions with Hi-C
    - A-B compartments
    - TADs

Overview

- Understand computational biology methods, tools and resources that explore DNA in the context of chromatin
- Detecting open and closed chromatin
- Functional transcriptional regions
- Long Distance interactions
- 3D structure of chromatin

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.
- 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.

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

HEP Human Epigenome Project

Methylomes

Various human, mouse tissues

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.

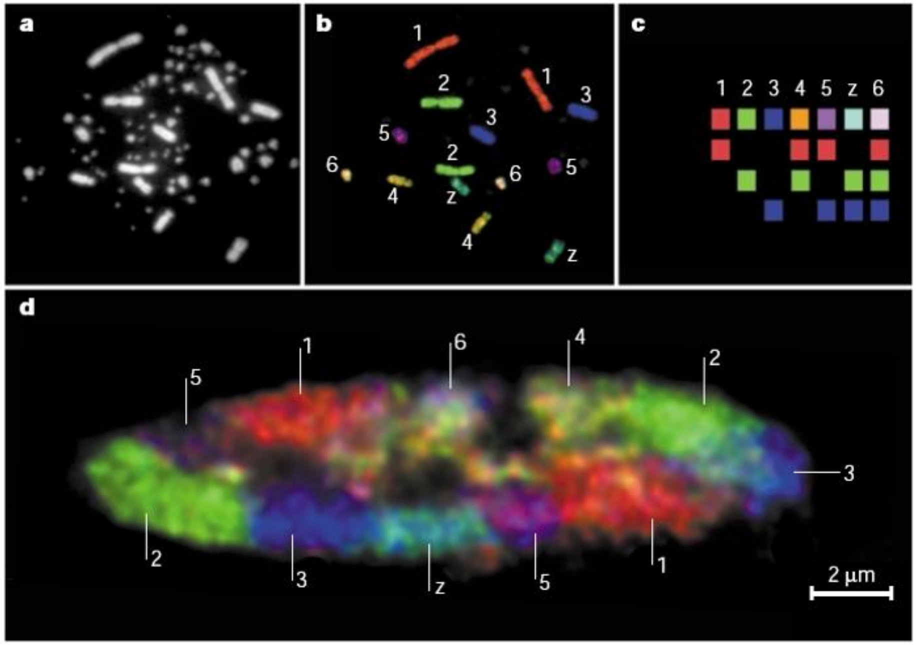
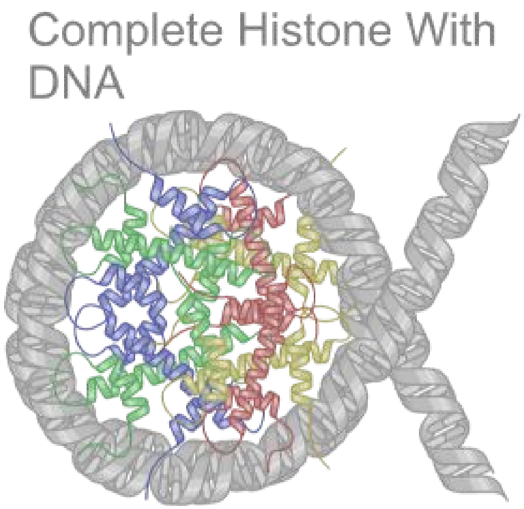
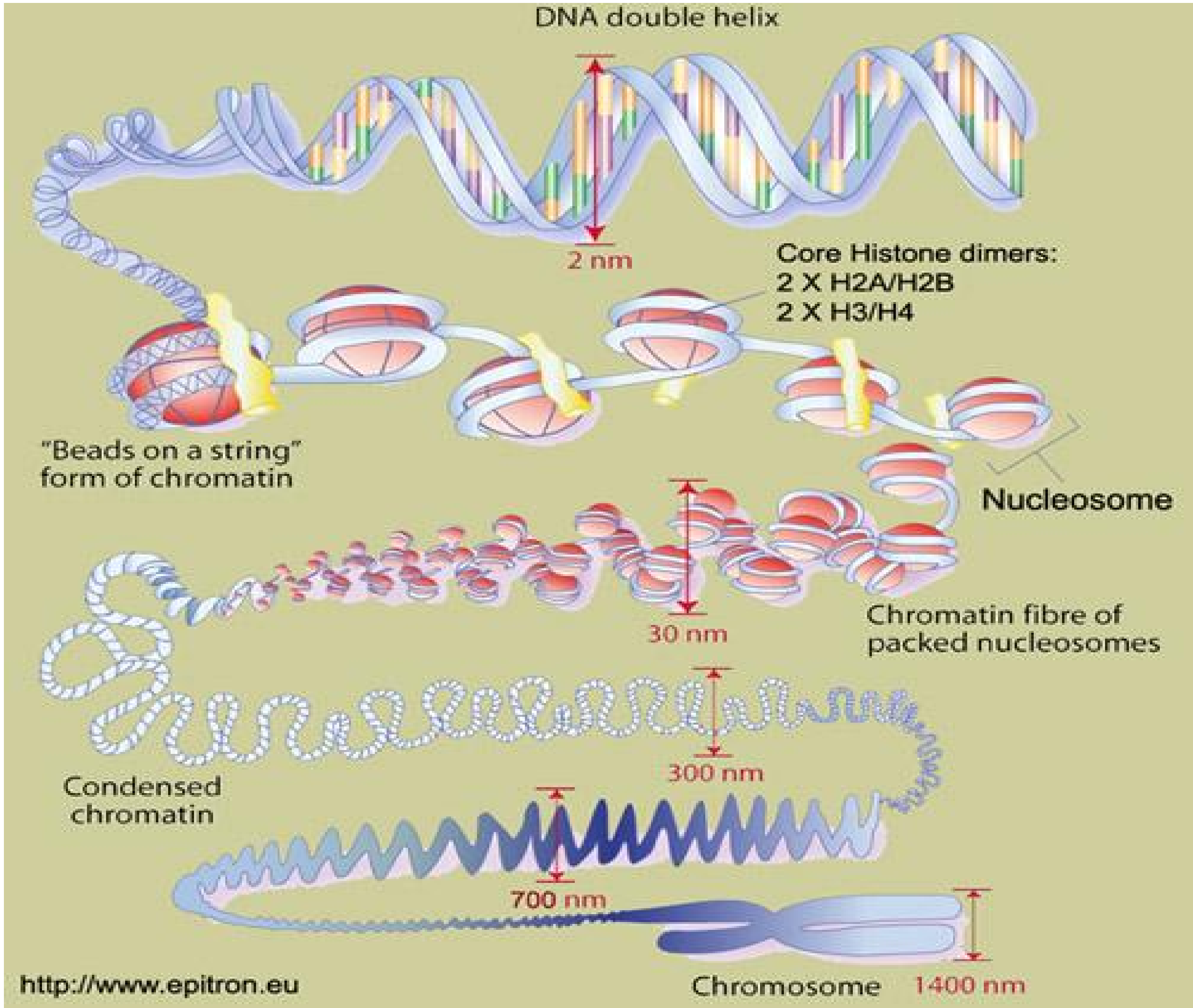


Fig: Multicolour FISH labelled chicken nucleus. (Misteli, T. 2008 Nature Education 1(1):167)

Chromatin Organization

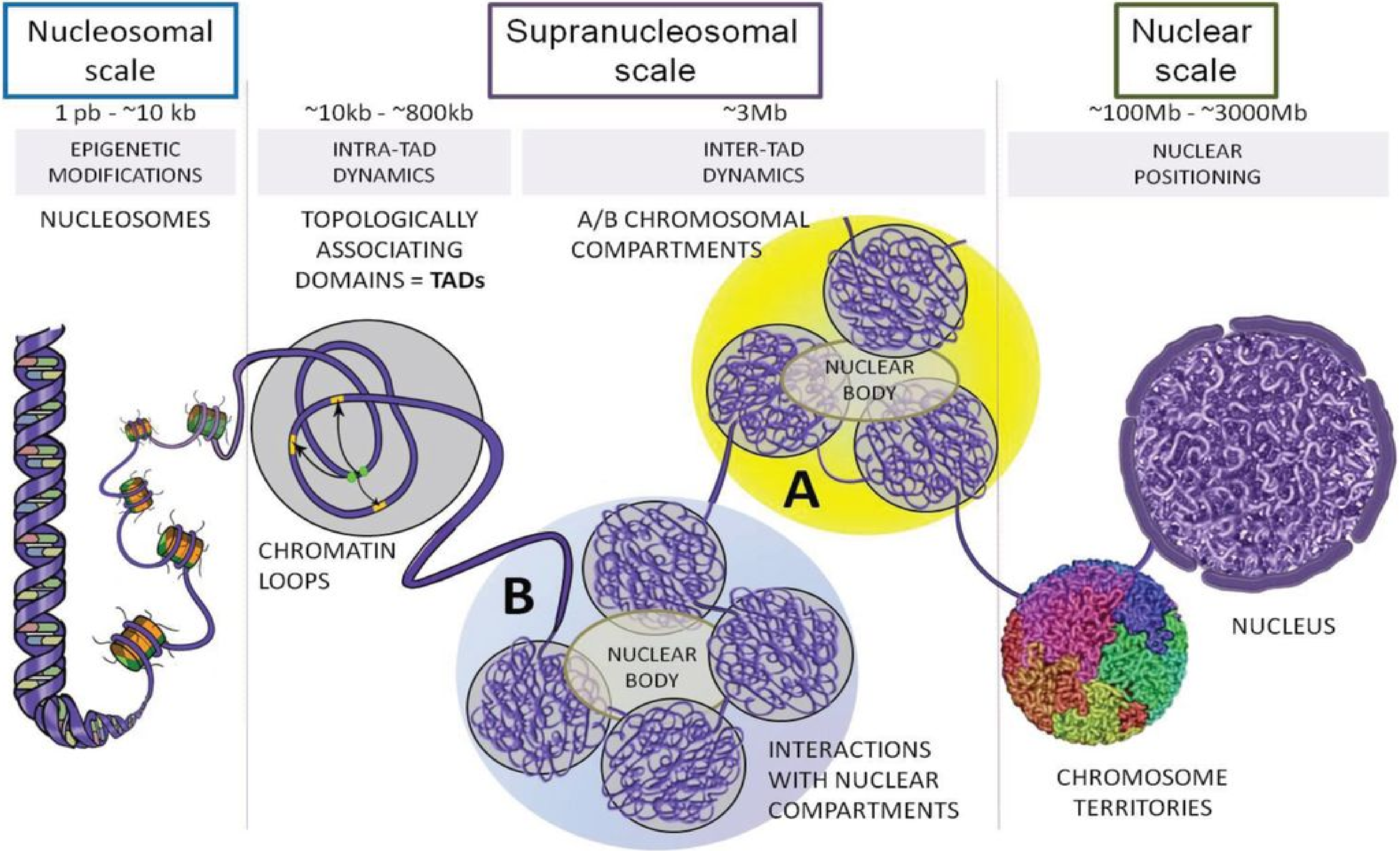
- Histone octamer core wrapped around by 147 bp of DNA and separated by linker DNA.





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

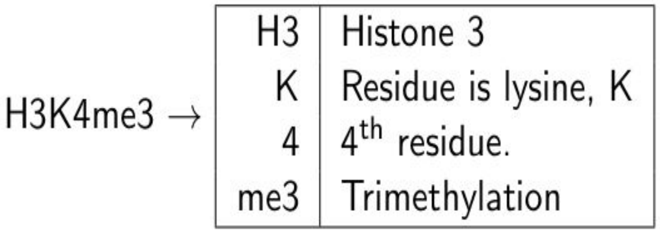


Ea et al., Genes 2015

Histone Modifications

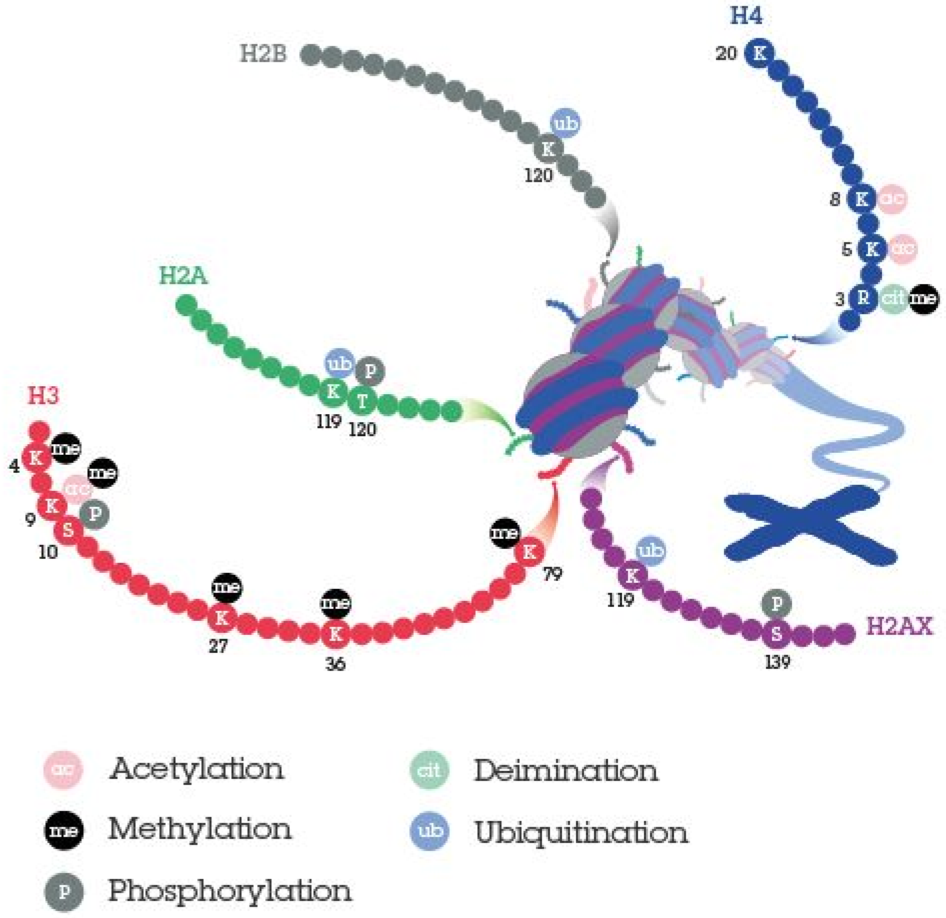
- Nucleosomes consist of 2x H2A/H2B and 2x H3/H4 histones.

- 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

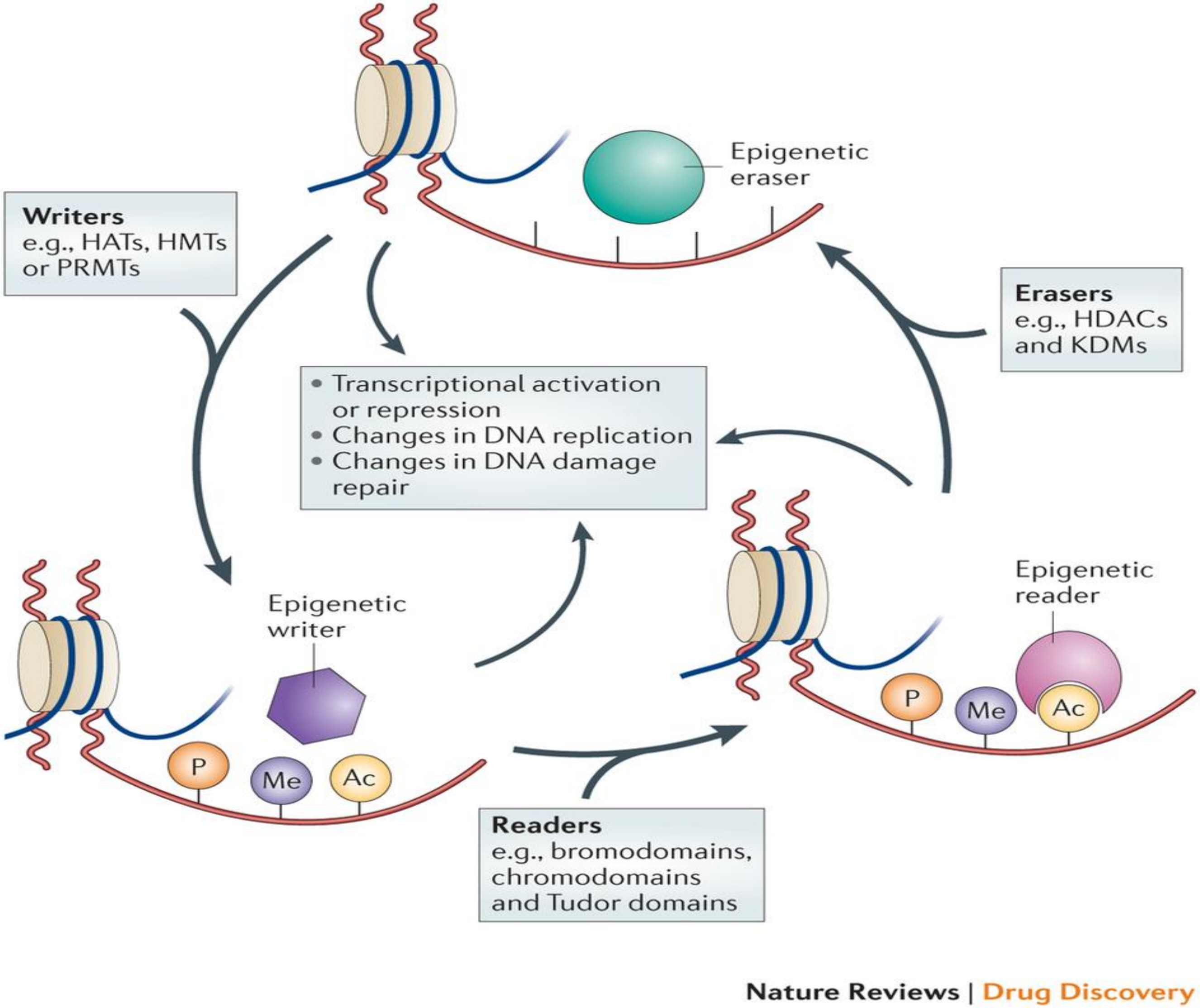
Some examples:

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

More information at:

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

Epigenetic Readers, Writers and Erasers



Writers
e.g., HATs, HMTs or PRMTs

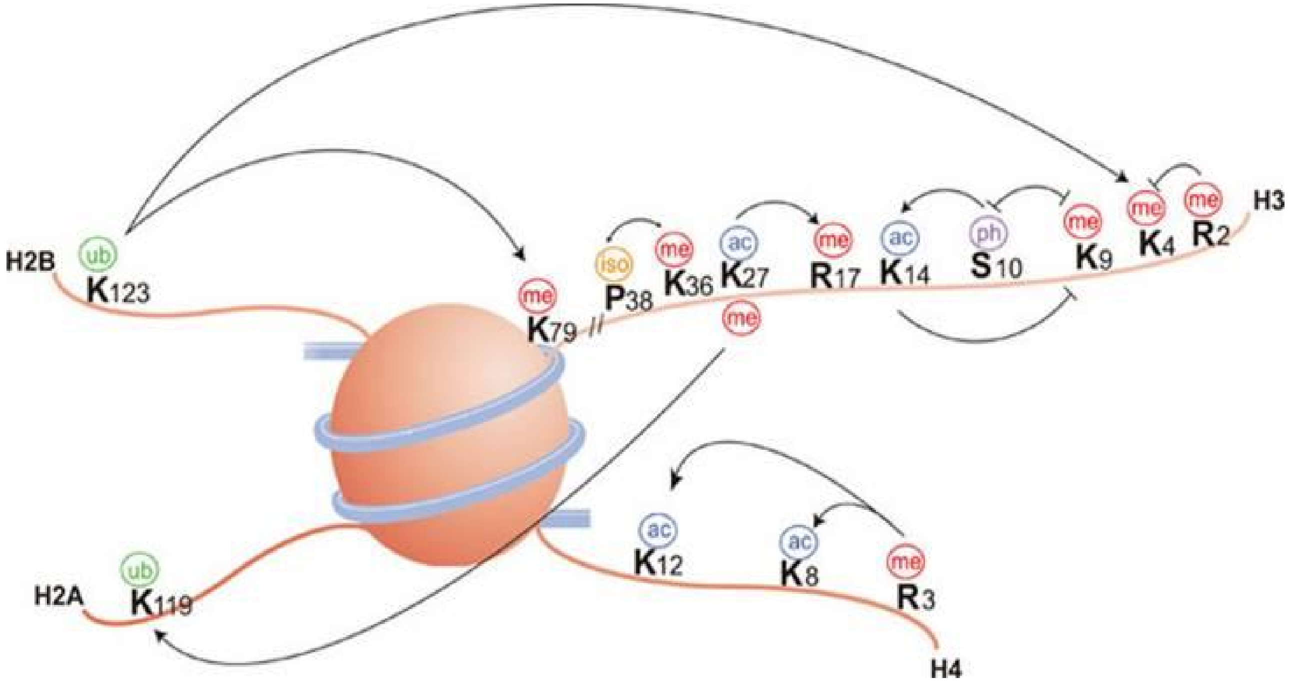
- Transcriptional activation or repression
- Changes in DNA replication
- Changes in DNA damage repair

Erasers
e.g., HDACs and KDMs

Readers
e.g., bromodomains, chromodomains and Tudor domains

Falkenberg and Jonstone, *Nature Reviews Drug Discovery, 2014*

Combinations of marks can have different effects



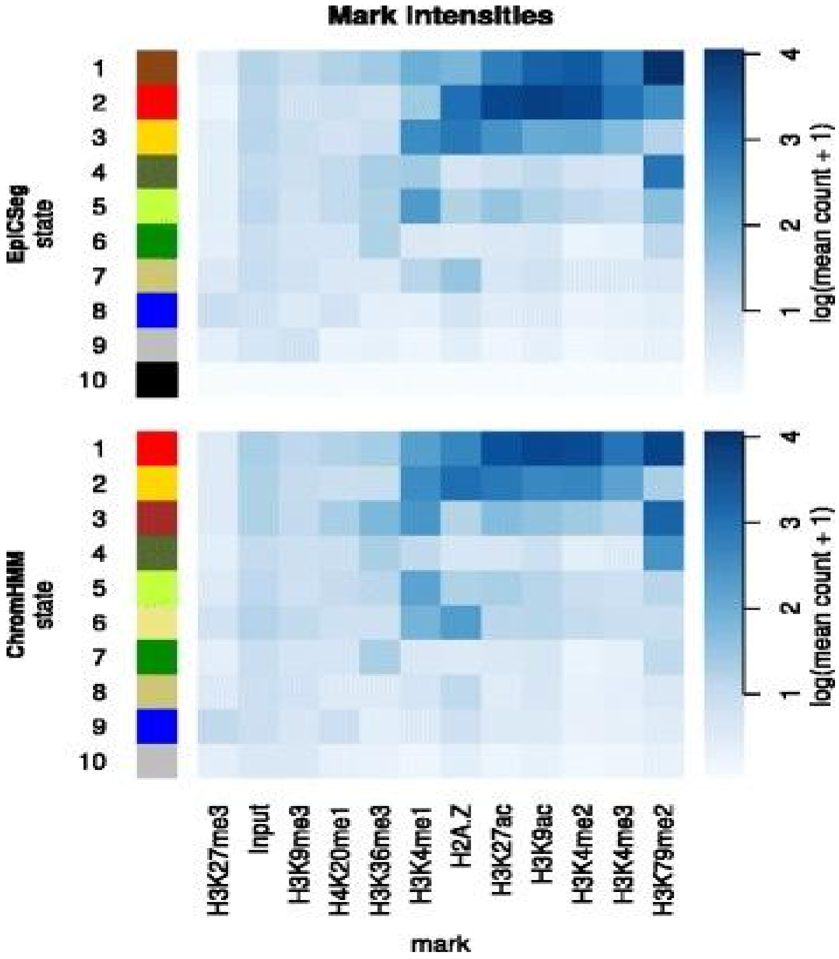
Bannister and Kouzarides (Cell Res. 2011)

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

Simplifying histone marks

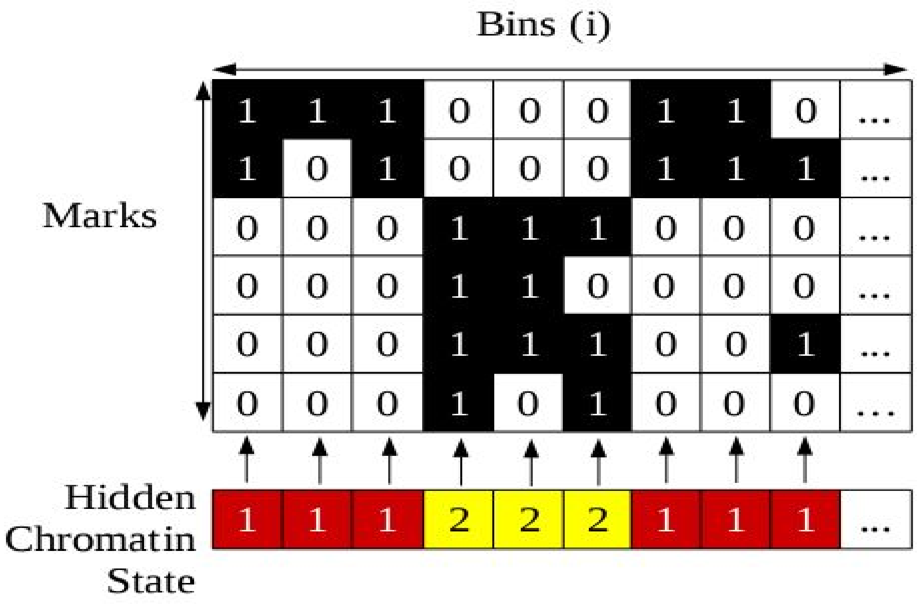
Unsupervised learning methods for *segmentation*;

- 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



Marks	1	1	1	0	0	0	1	1	0	...
	1	0	1	0	0	0	1	1	1	...
	0	0	0	1	1	1	0	0	0	...
	0	0	0	1	1	0	0	0	0	...
	0	0	0	1	1	1	0	0	1	...
	0	0	0	1	0	1	0	0	0	...
Hidden Chromatin State	1	1	1	2	2	2	1	1	1	...

Chromatin Segmentation

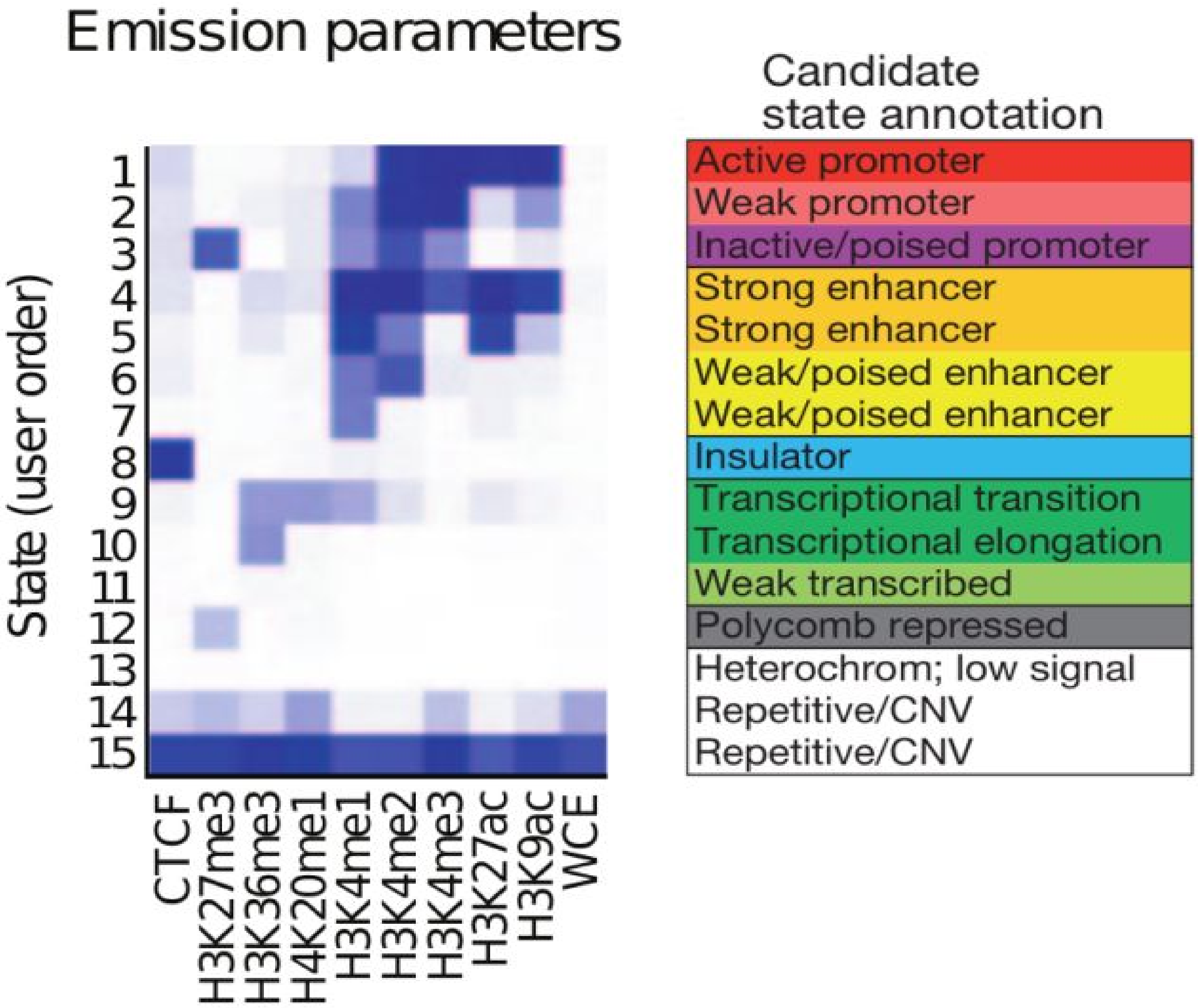
Advantages:

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

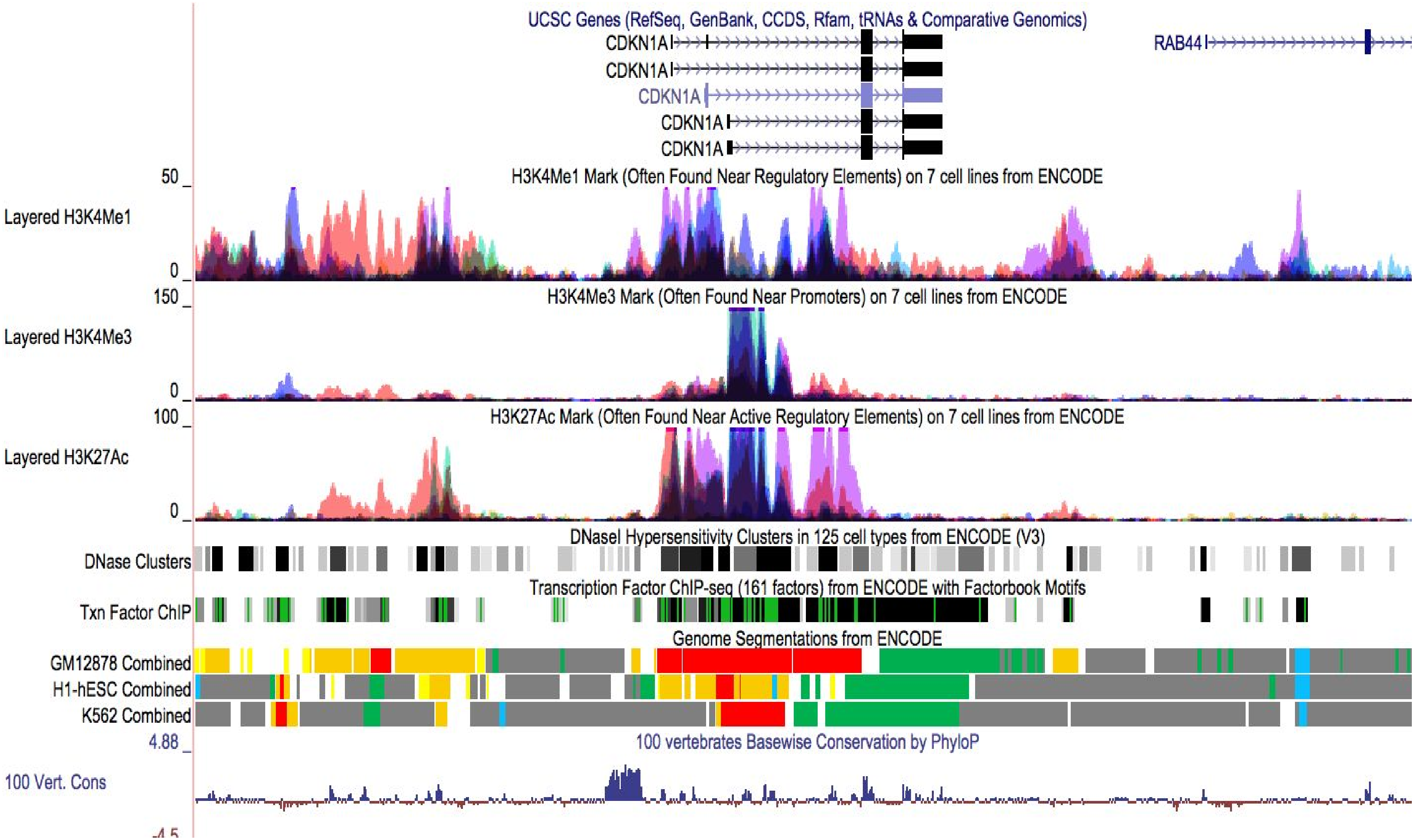
Disadvantages:

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

Chromatin Colours



Visualizing Chromatin Marks



UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
CDKN1A
CDKN1A
CDKN1A
CDKN1A
CDKN1A
RAB44
50
Layered H3K4Me1
0
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
150
Layered H3K4Me3
0
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Txn Factor ChIP
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Genome Segmentations from ENCODE
GM12878 Combined
H1-hESC Combined
K562 Combined
4.88
100 Vert. Cons
100 vertebrates Basewise Conservation by PhyloP
-4.5

Chromosomal Conformation Capture

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

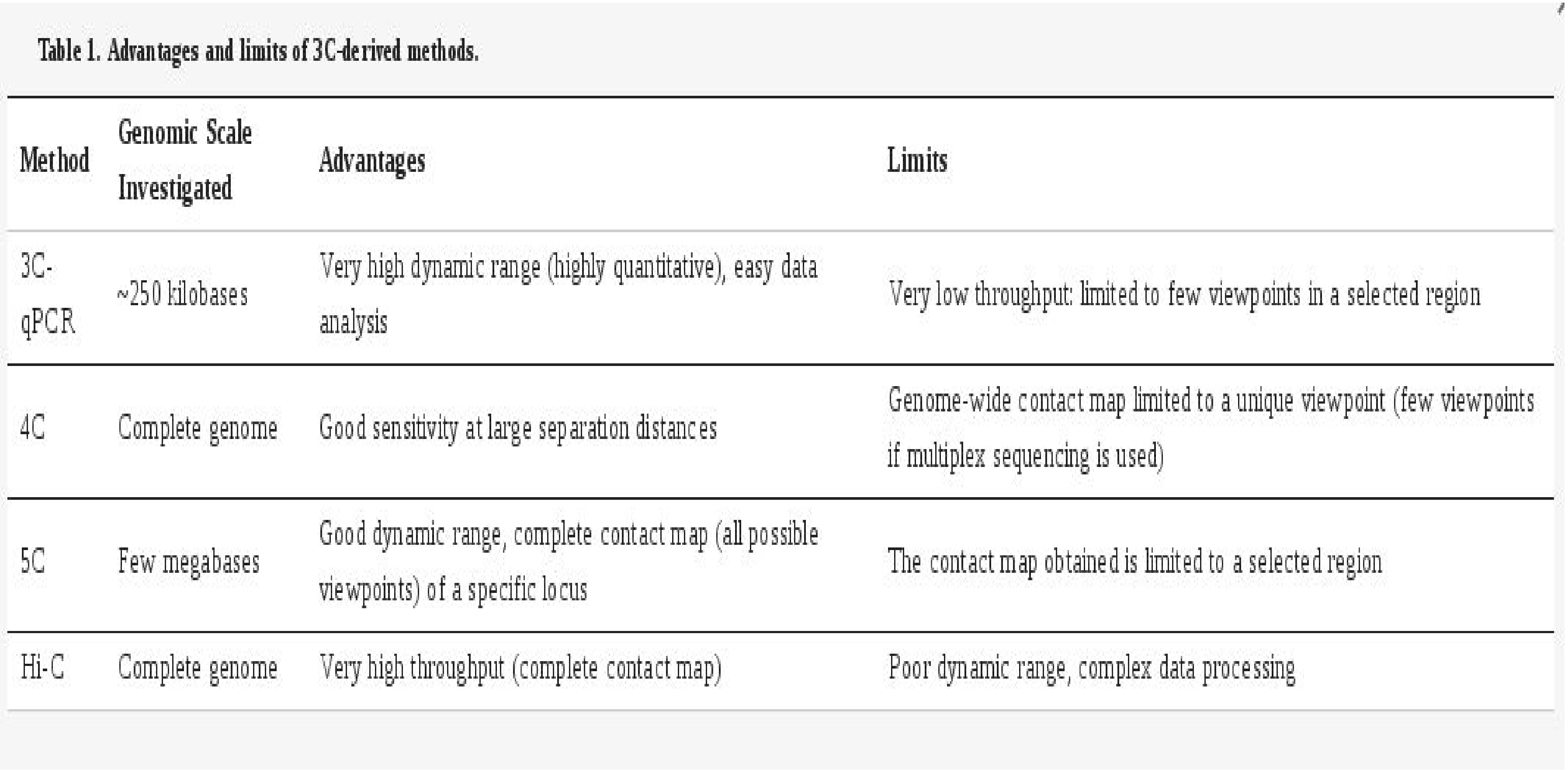
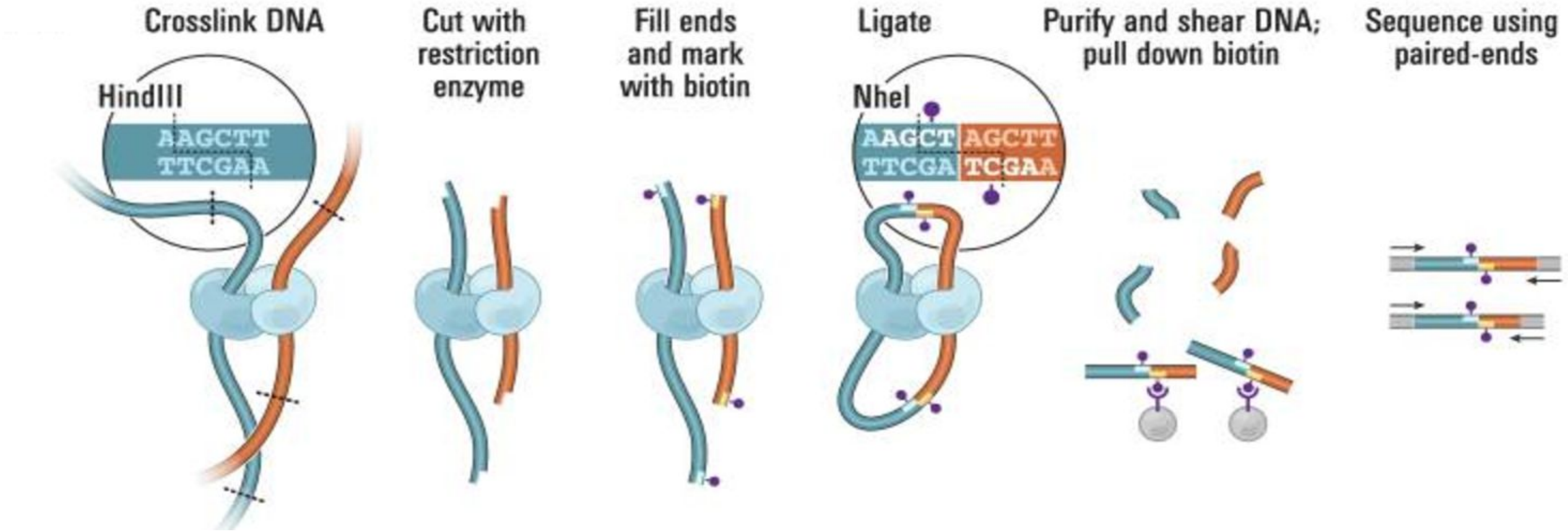


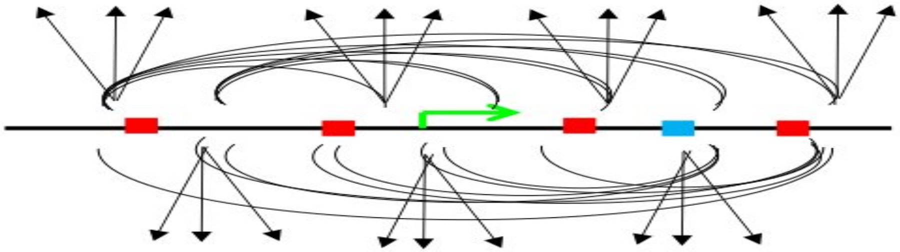
Table 1. Advantages and limits of 3C-derived methods.

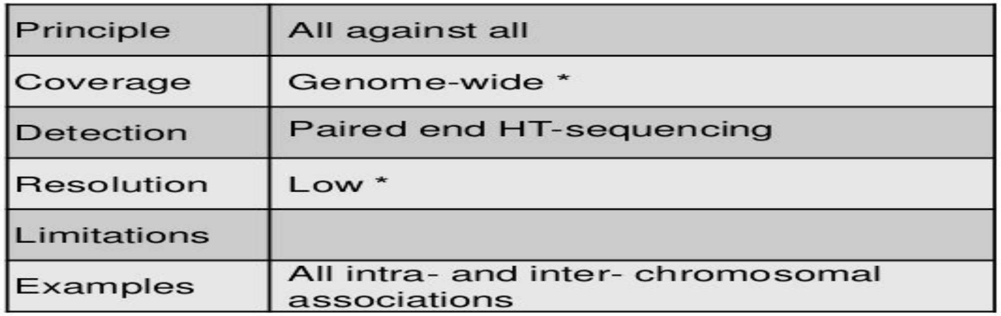
Method	Genomic Scale Investigated	Advantages	Limits
3C-qPCR	~250 kilobases	Very high dynamic range (highly quantitative), easy data analysis	Very low throughput: limited to few viewpoints in a selected region
4C	Complete genome	Good sensitivity at large separation distances	Genome-wide contact map limited 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

Hi-C: long distance interactions


Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

Lieberman-Aiden, Science 2009



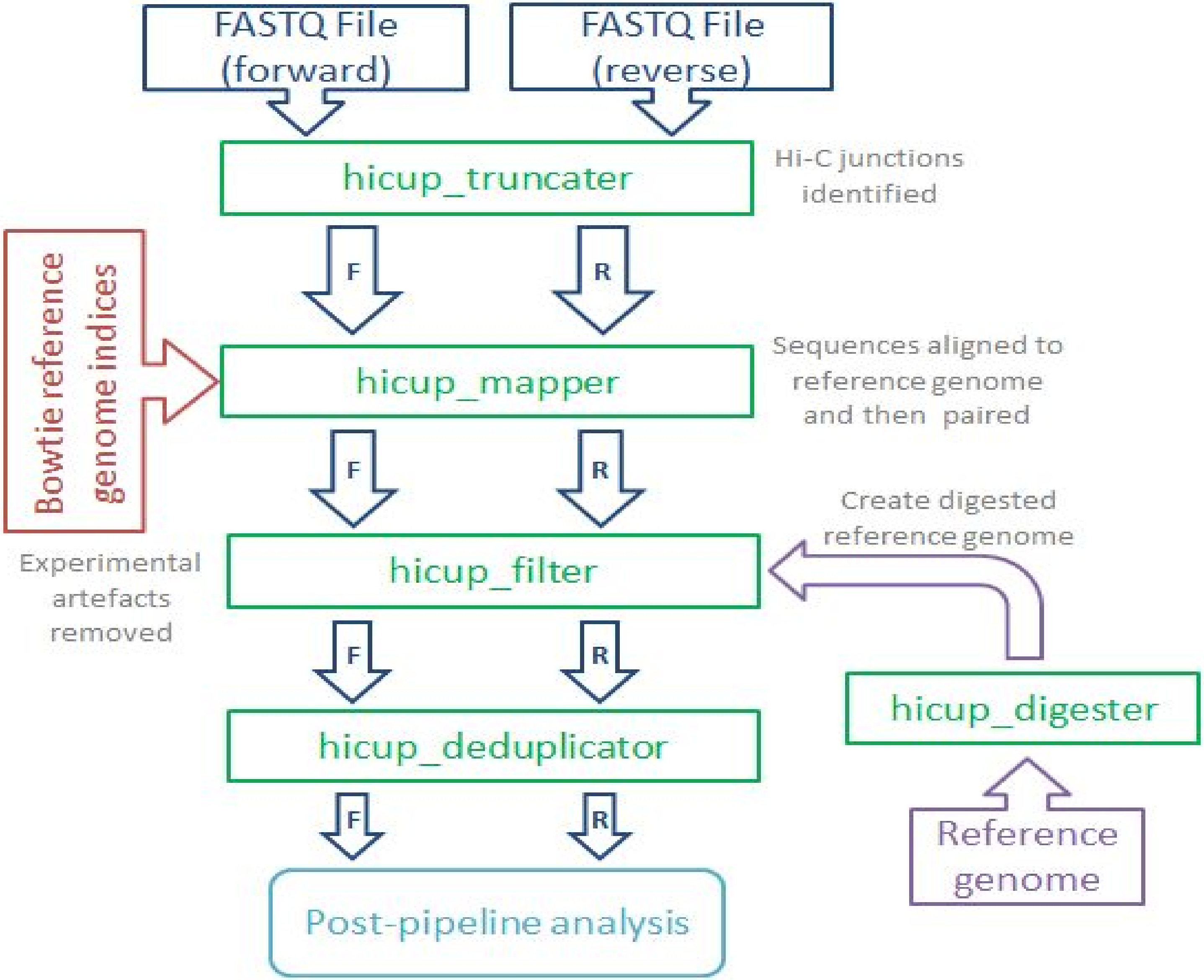


Principle	All against all
Coverage	Genome-wide *
Detection	Paired end HT-sequencing
Resolution	Low *
Limitations	
Examples	All intra- and inter- chromosomal associations

HiCUP

- HiCUP is a bioinformatics pipeline for processing Hi-C data.
- The pipeline maps FASTQ data using bowtie2 against a reference genome and filters out frequently encountered experimental artefacts.
- The pipeline also produces paired-read files in SAM/BAM format, each read pair corresponding to a putative Hi-C di-tag.
- HiCUP also produces summary statistics at each stage of the pipeline and provides quality control reports, helping pinpoint potential problems and refine the experimental protocol.

HiCUP workflow



FASTQ File (forward)

FASTQ File (reverse)

hicup_truncater

Hi-C junctions identified

F

R

Bowtie reference genome indices

hicup_mapper

Sequences aligned to reference genome and then paired

F

R

Create digested reference genome

hicup_filter

Experimental artefacts removed

F

R

hicup_digester

hicup_deduplicator

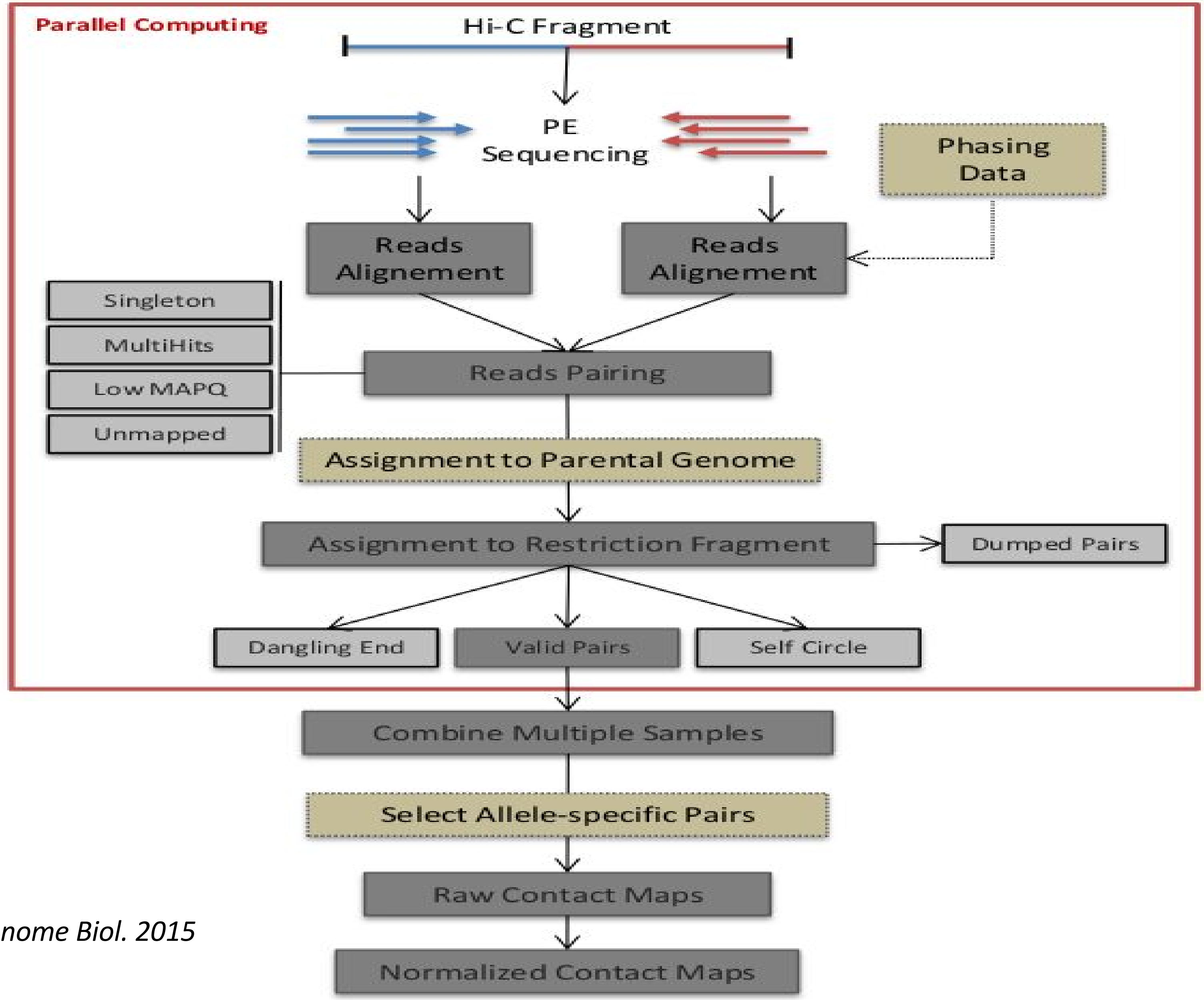
F

R

Reference genome

Post-pipeline analysis

HiC: Computational workflow



Servant, Genome Biol. 2015

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.

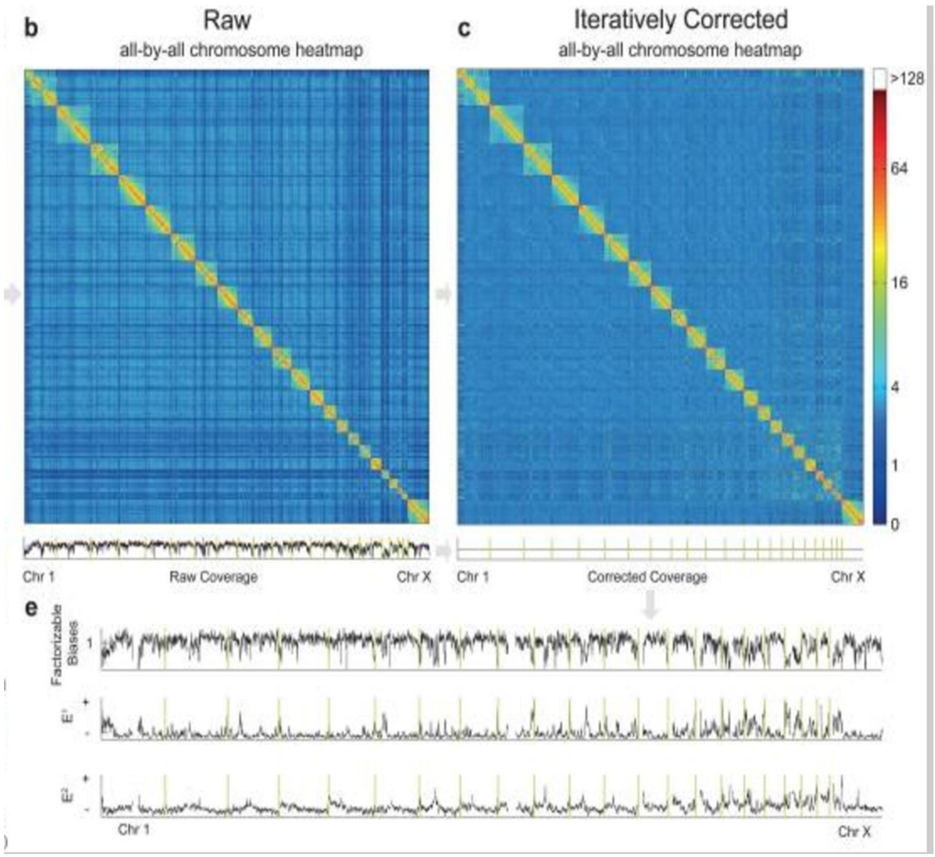
Biases:

- 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

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

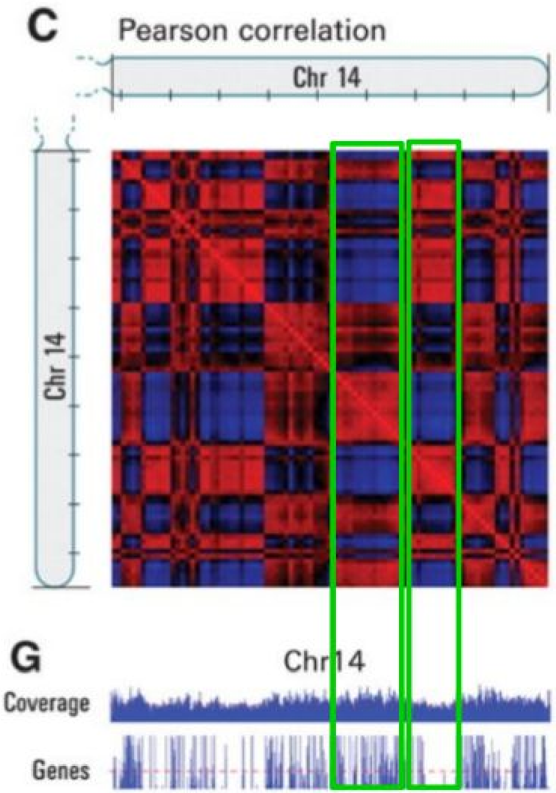


Imakaev et al. 2012

Chromatin: A-B compartments

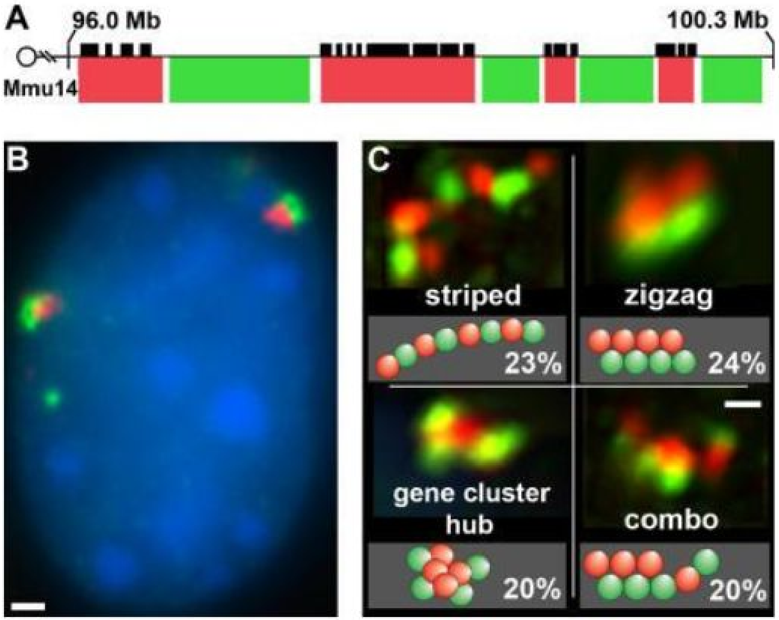
- The genome (at the Mb scale) can be divided into cell type or condition specific A/B compartments that are associated with open and closed chromatin.
- The A compartment is associated with gene rich, transcriptionally active, open chromatin state regions.
- Interactions are more likely between A-A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

A-B compartments



Compartment A B

Compartment A is gene rich

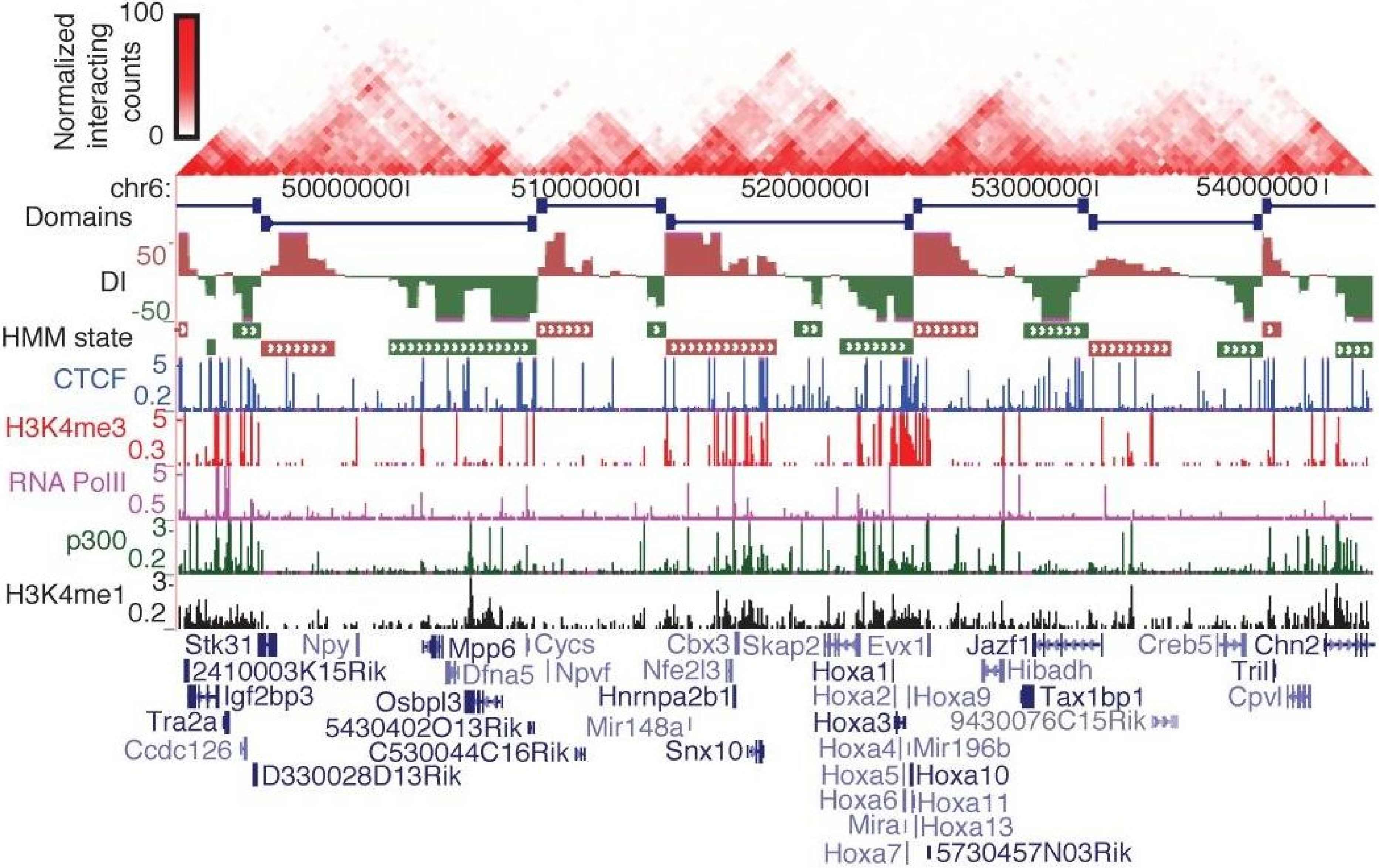


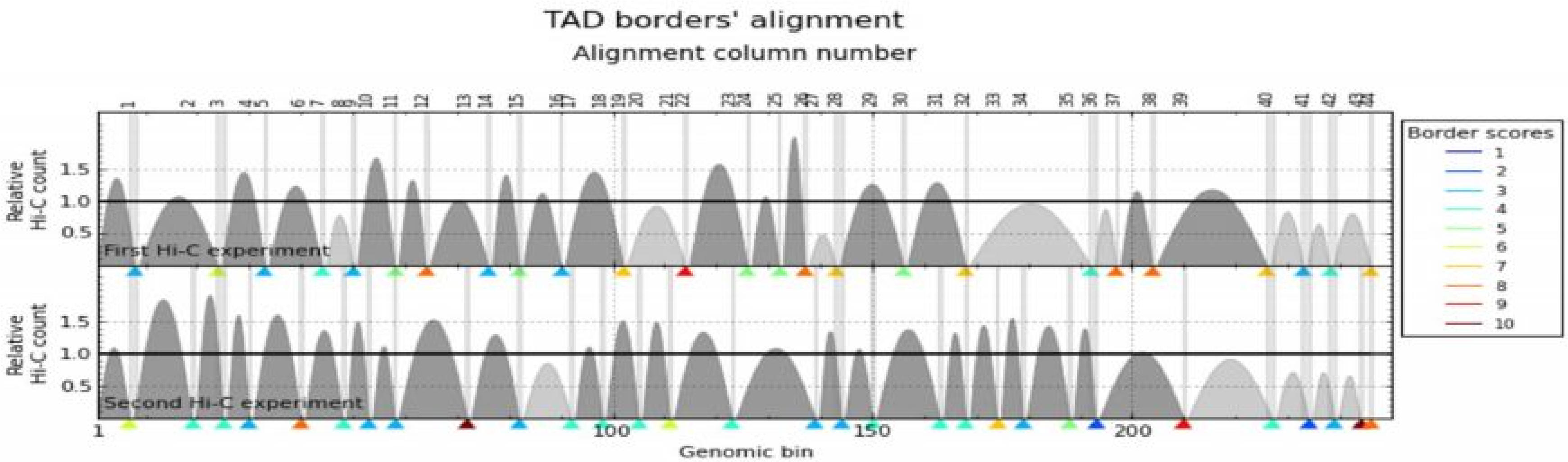
Compartment A
Genes Spearman's ρ = 0.431
Expression Spearman's ρ = 0.476
Accessible chromatin, Spearman's ρ = 0.651
H3K36 trimethylation, Spearman's ρ = 0.601 (active)
H3K27 trimethylation, Spearman's ρ = 0.282 (repressive)
A is more closely associated with open, accessible, actively transcribed chromatin.

Liebermann-Aiden et al., 2009

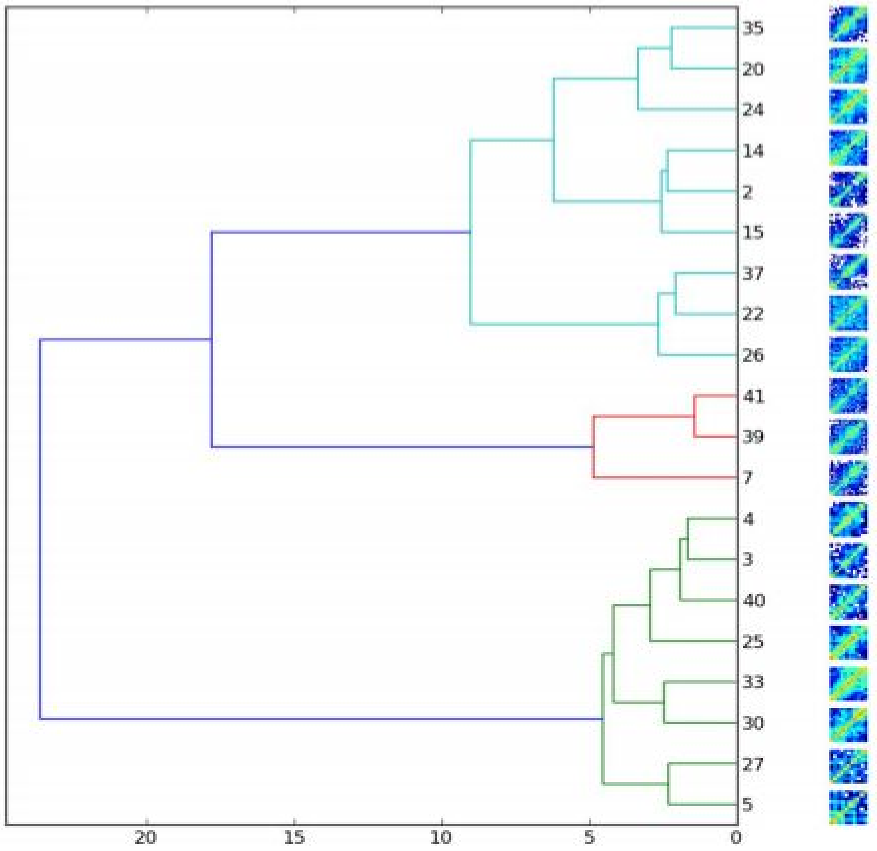
Shopland, et al 2006

Topologically Associated Domains

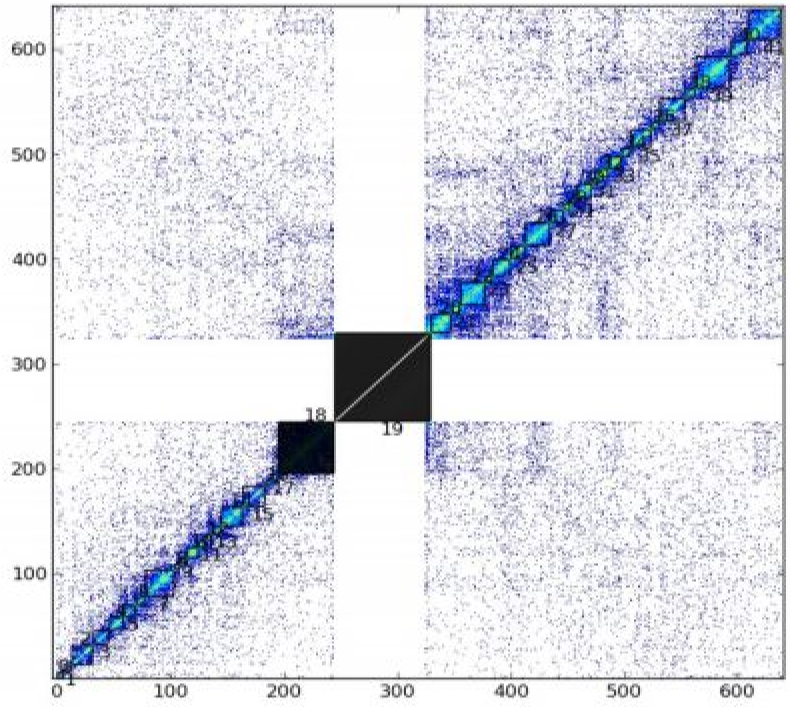




TAD borders' alignment
Alignment column number
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44
Relative Hi-C count
1.5
1.0
0.5
First Hi-C experiment
Second Hi-C experiment
1
100
150
200
Genomic bin
Border scores
1
2
3
4
5
6
7
8
9
10

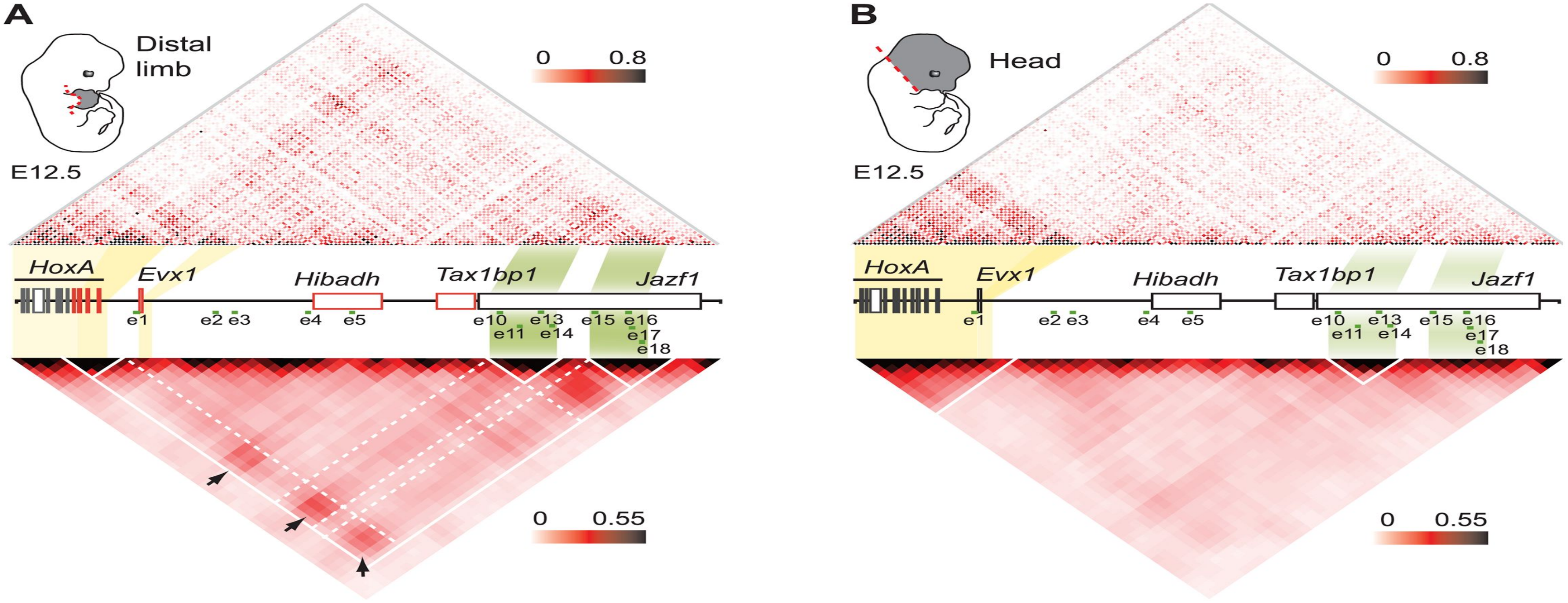


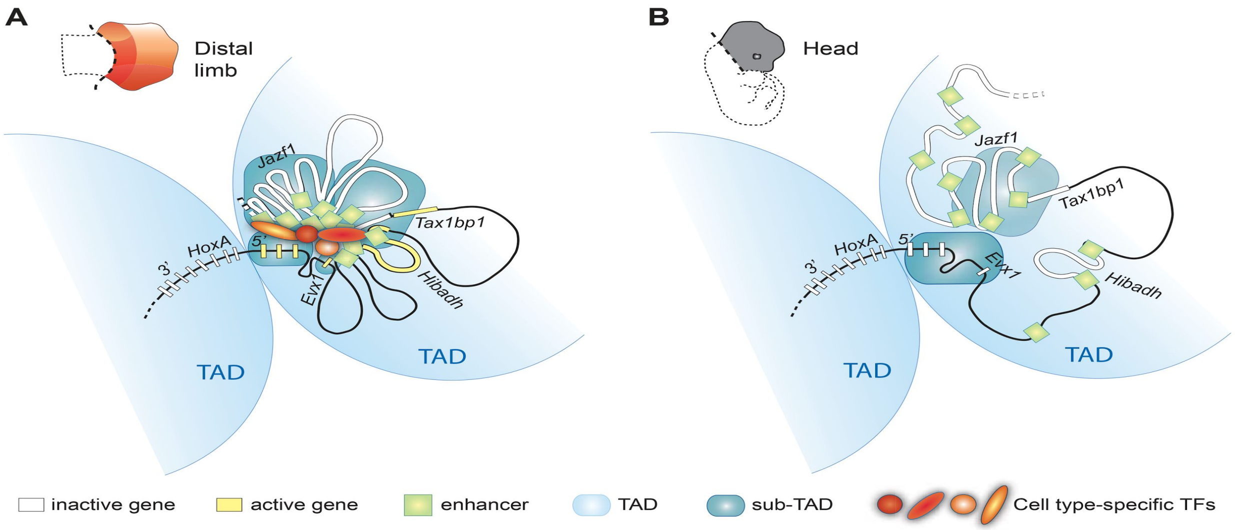
35
20
24
14
2
15
37
22
26
41
39
7
4
3
40
25
33
30
27
5
20
15
10
5
0



600
500
400
300
200
100
0
100
200
300
400
500
600
18
19

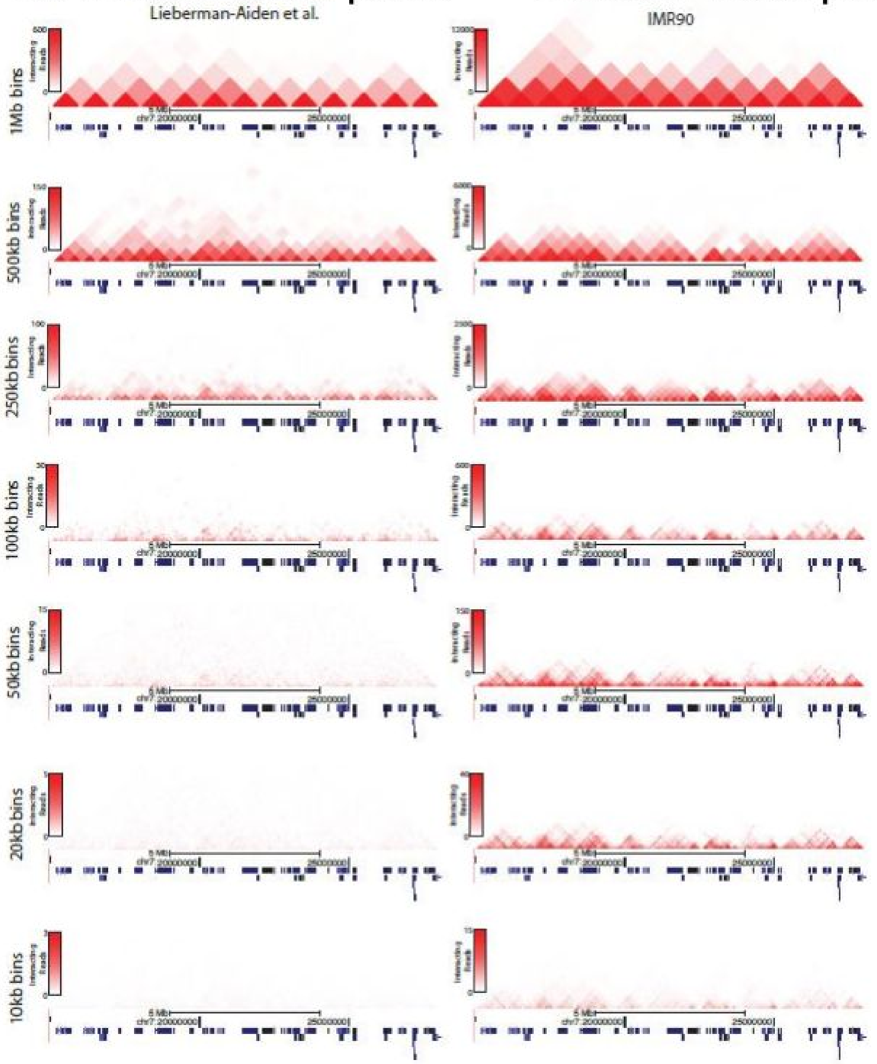
Long distance interactions





Read depth and bin size

~ 20 million read pairs ~ 1 billion read pairs



Liebermann-Aiden et al., 2009

Dixon et al., 2012