# Computational Systems Biology
# Deep Learning in the Life Sciences

## 6.802 6.874 20.390 20.490 HST.506

David Gifford

Lecture 10

March 12, 2019

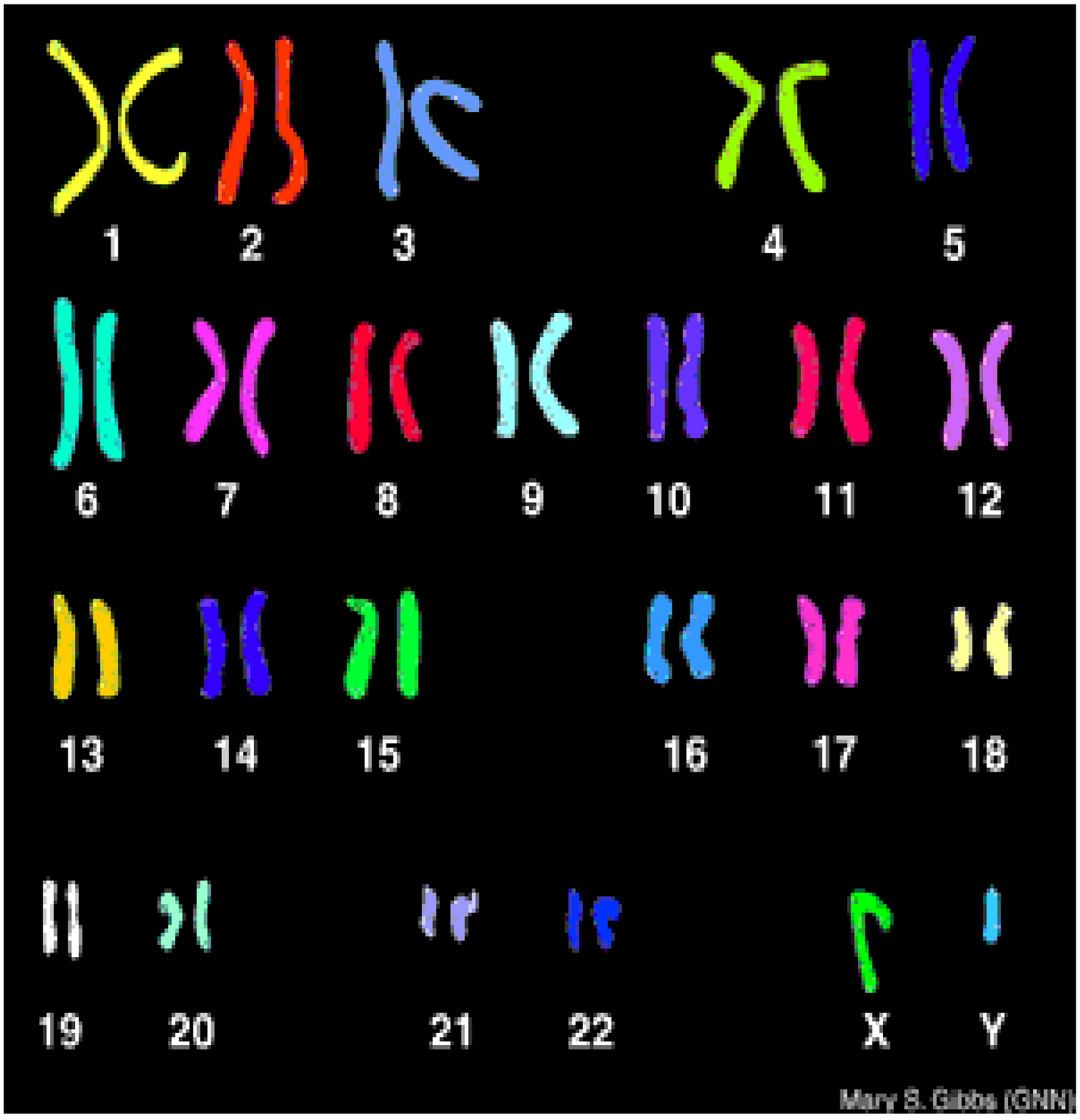
# Histone Marks
# Chromatin 3D Structure


http://mit6874.github.io

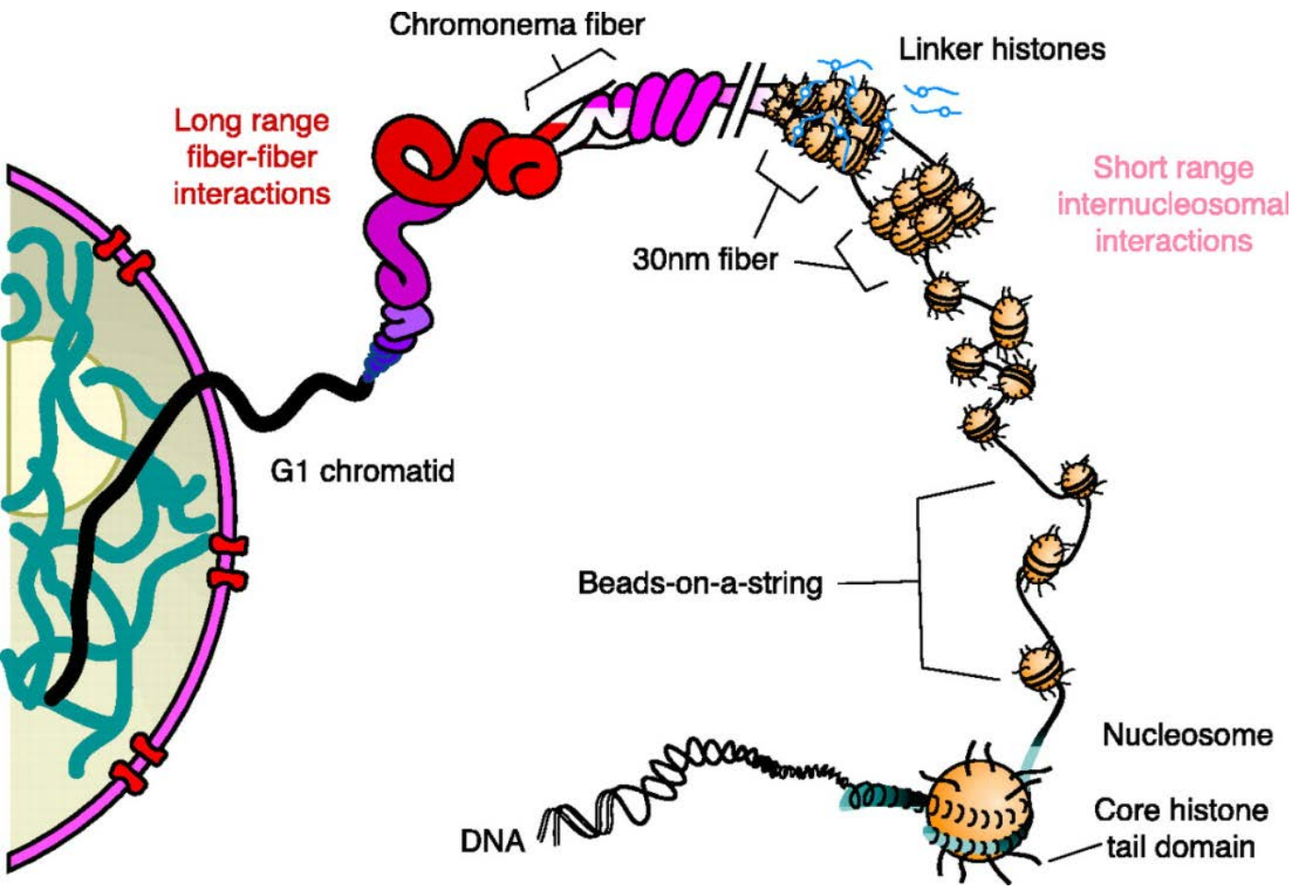
# Goals for today

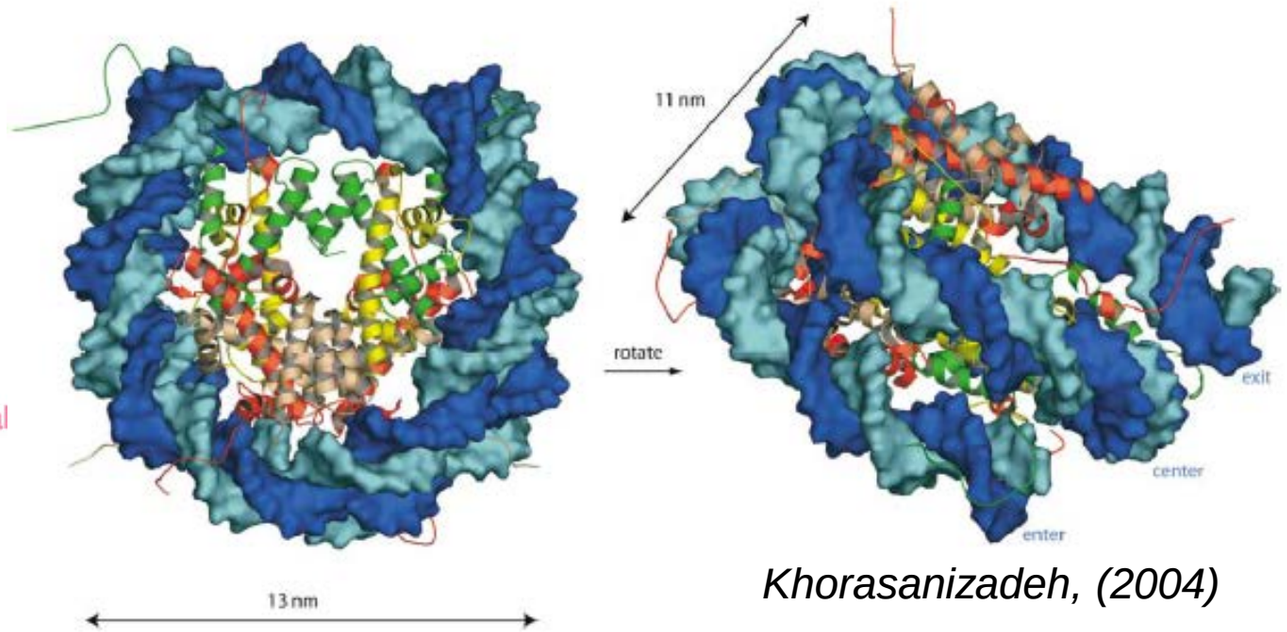
- Chromatin marks and their models
  - Hidden Markov Model (HMM)
  - Deep learning model (DeepSEA)
- Three-dimensional chromatin structure
  - Inferring it
  - Predicting it

# 1. Chromatin marks and biological state


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Mary S. Gibbs (GNN)

# Chromatin and Nucleosome Organization




*Khorasanizadeh, (2004)*

Green -H3, yellow - H4, red - H2A, pink - H2B.
Dark and light blue - DNA

**Nucleosome**

DNA - 146 base pairs, wrapped 1.7 times in a left-handed superhelix

Proteins - two copies of each Histones H2A, H2B, H3 and H4. Higher organisms have linker H1 histone

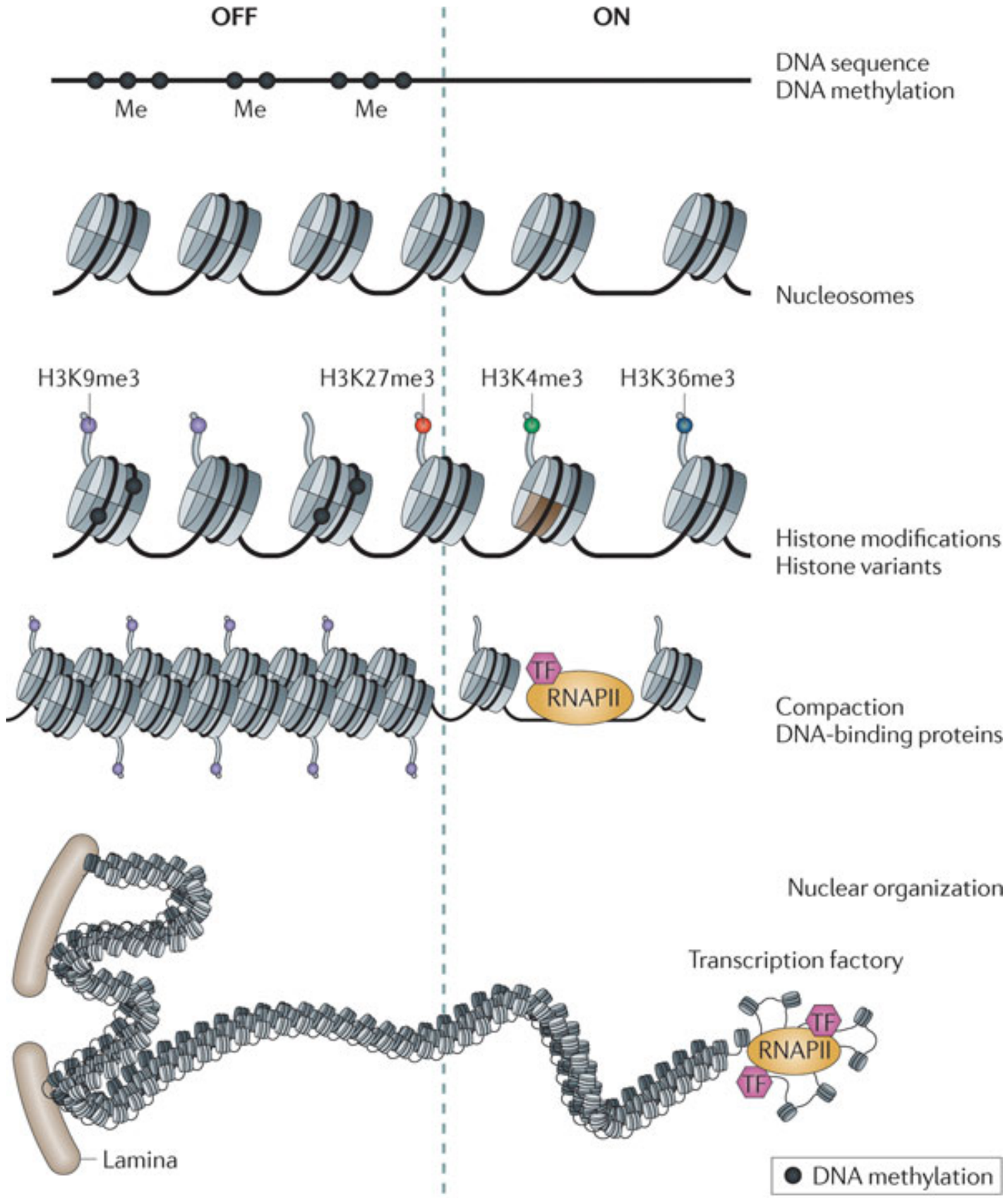
**Histone variants**

H3 variants: H3.3 - transcribed
CENP-A - centromeres

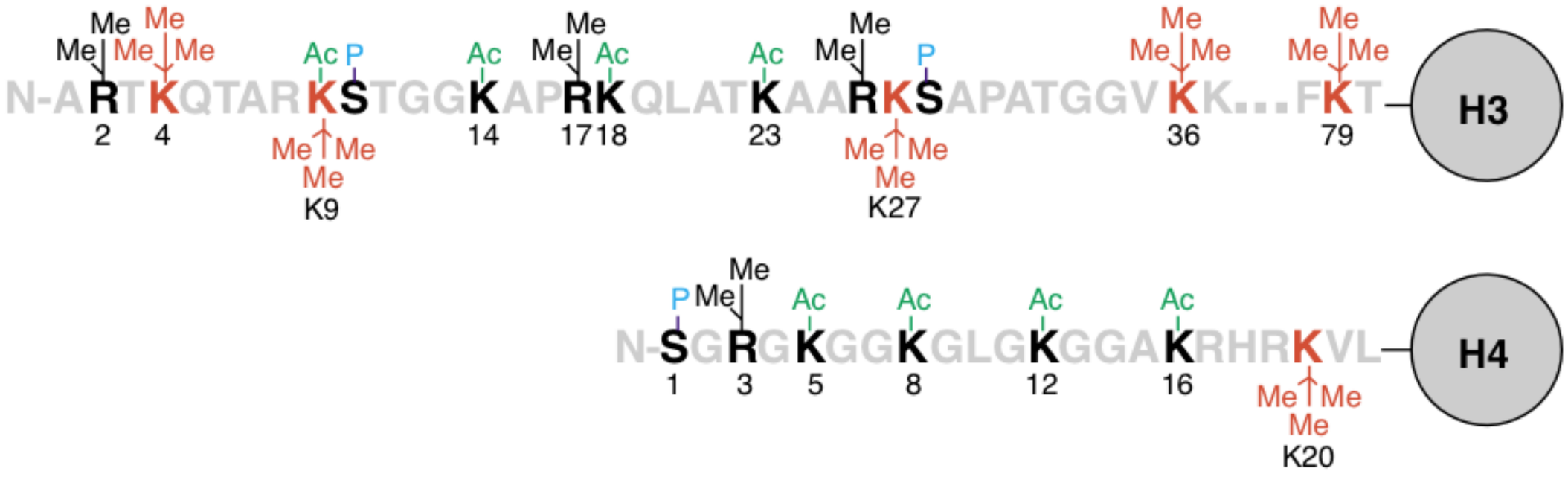
H2A variants: H2A.X - DNA damage
macroH2A - X chromosome
H2A.Z - transcribed regions

Chromatin organization has multiple structural layers and organizes chromatin into “domains”

Both DNA methylation and chromatin marks contain important functional information

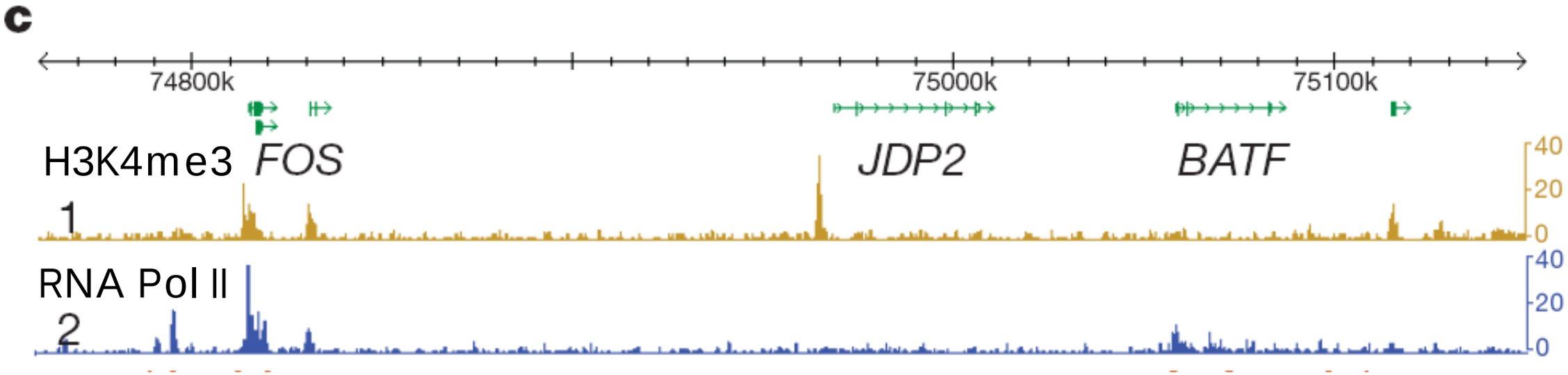


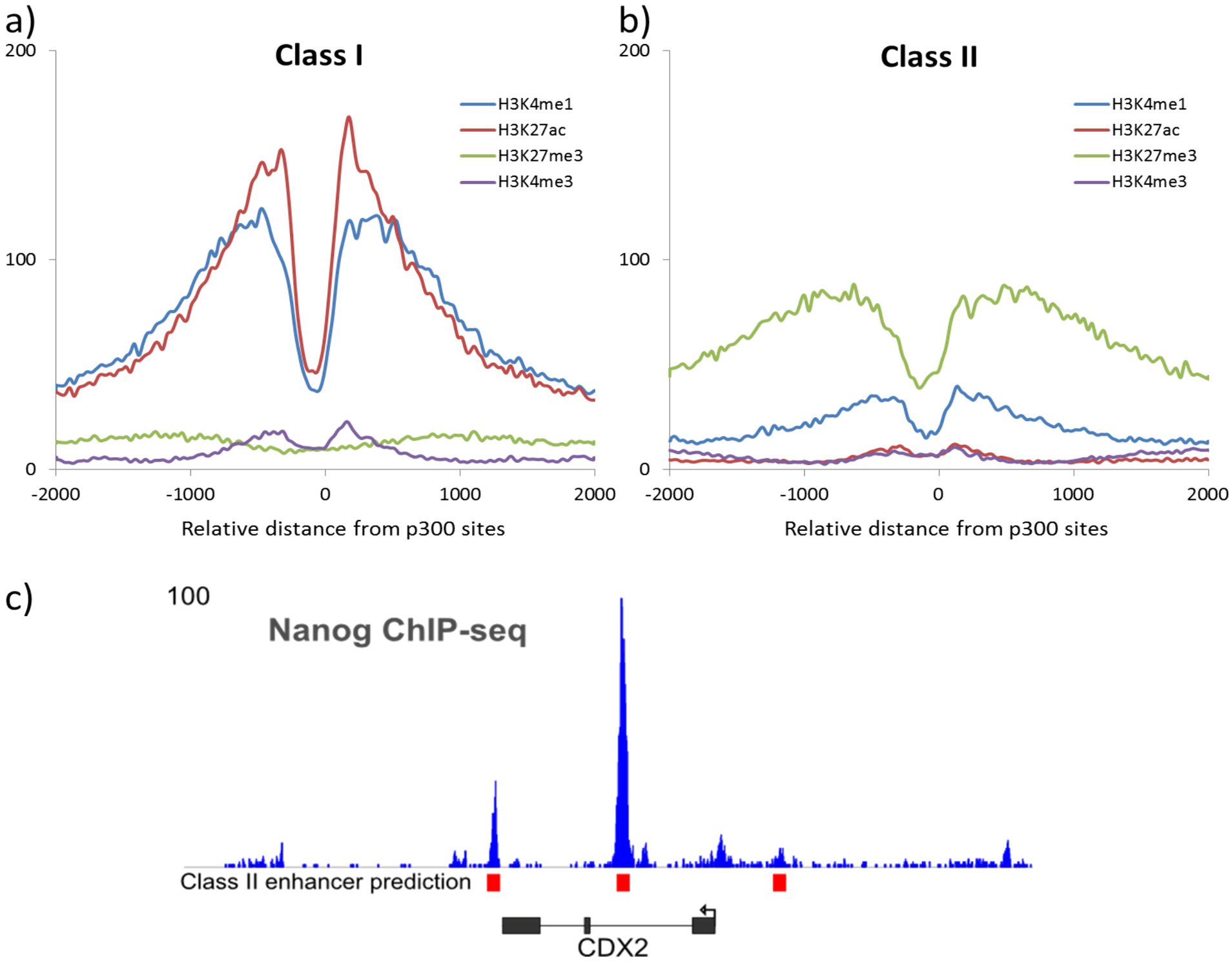
# Histone Tail Modifications



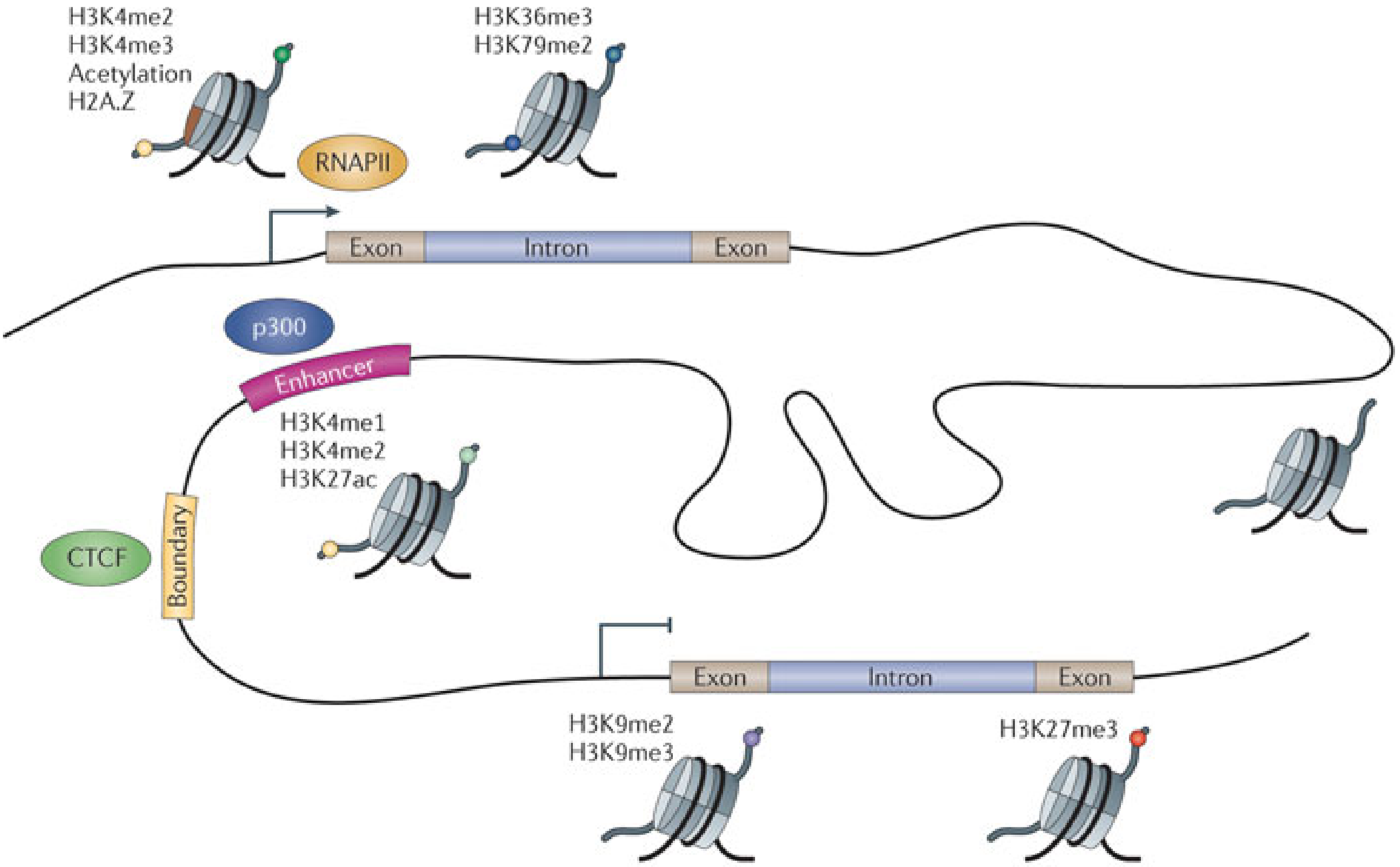

*Sims III et al., 2003*

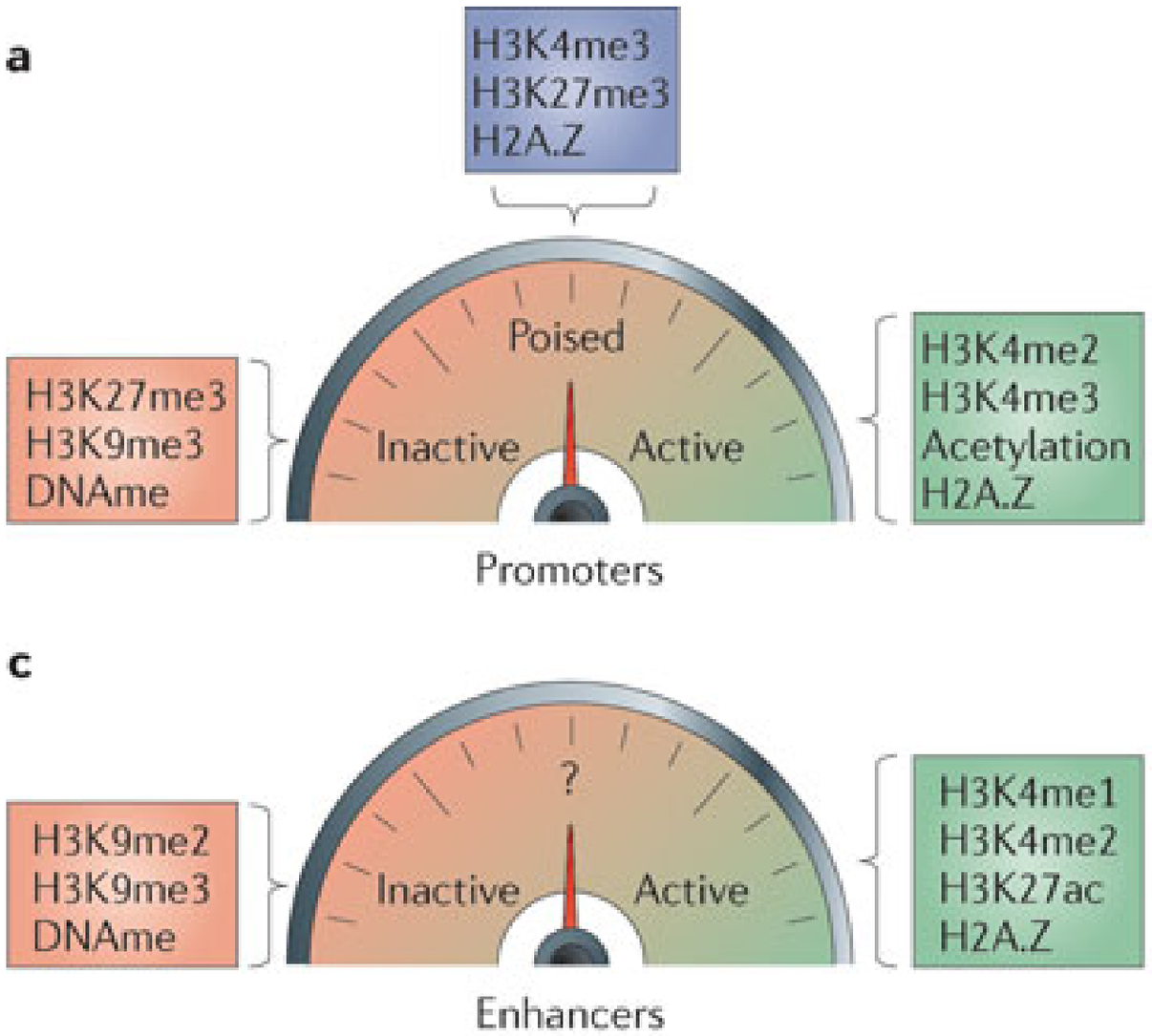
# We can observe chromatin marks and other genome associated proteins using ChIP-seq

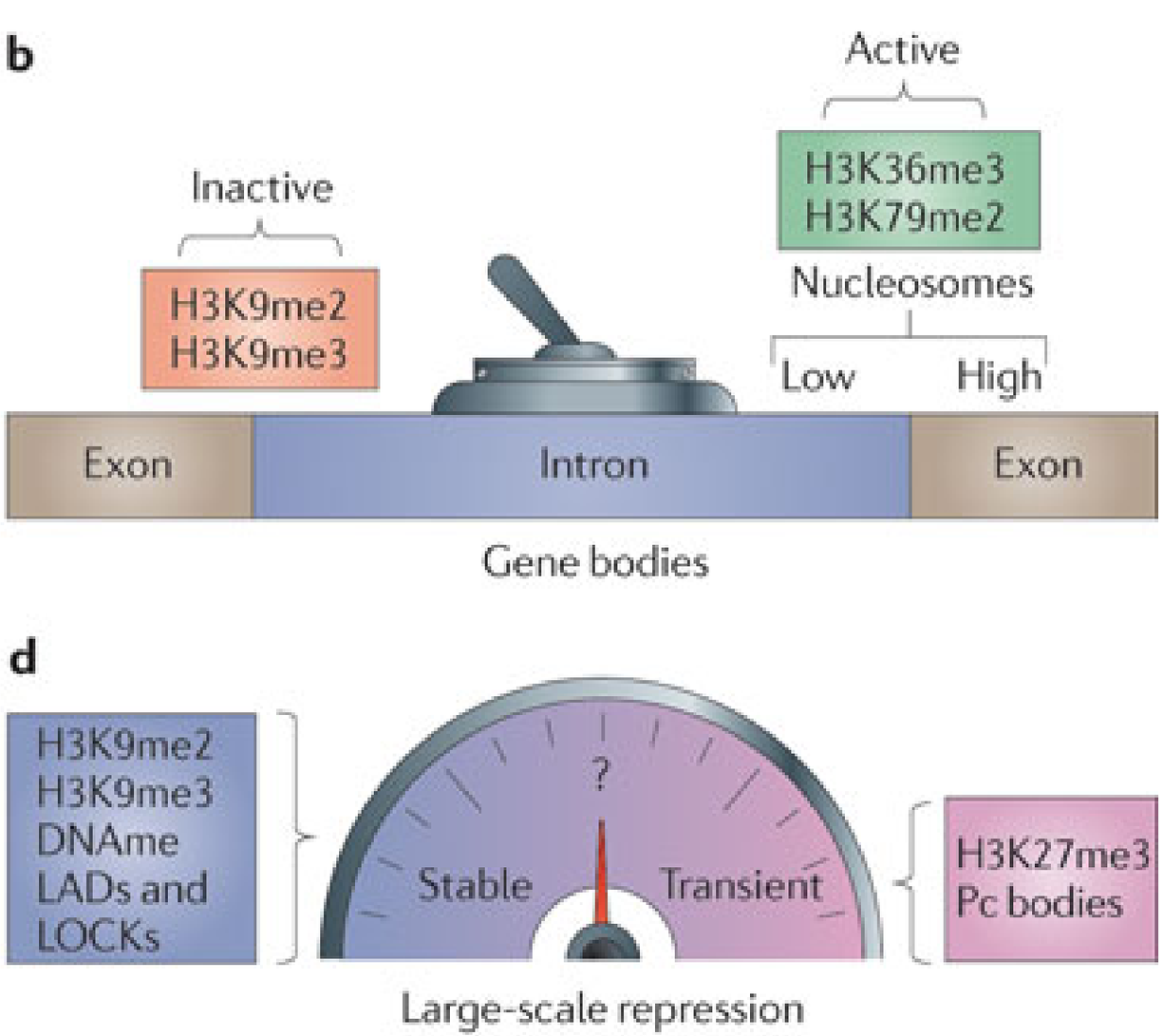





**Detection of Class I (active) and Class II (poised) enhancers.** a) b) hESC ChIP-seq read density profiles were generated for the indicated histone modifications centered on p300-bound regions in the top 1000 Class I and Class II enhancers, respectively. c) hESC Nanog ChIP-seq shows that Nanog binds at the three predicted Class II enhancer positions near the CDX2 gene.

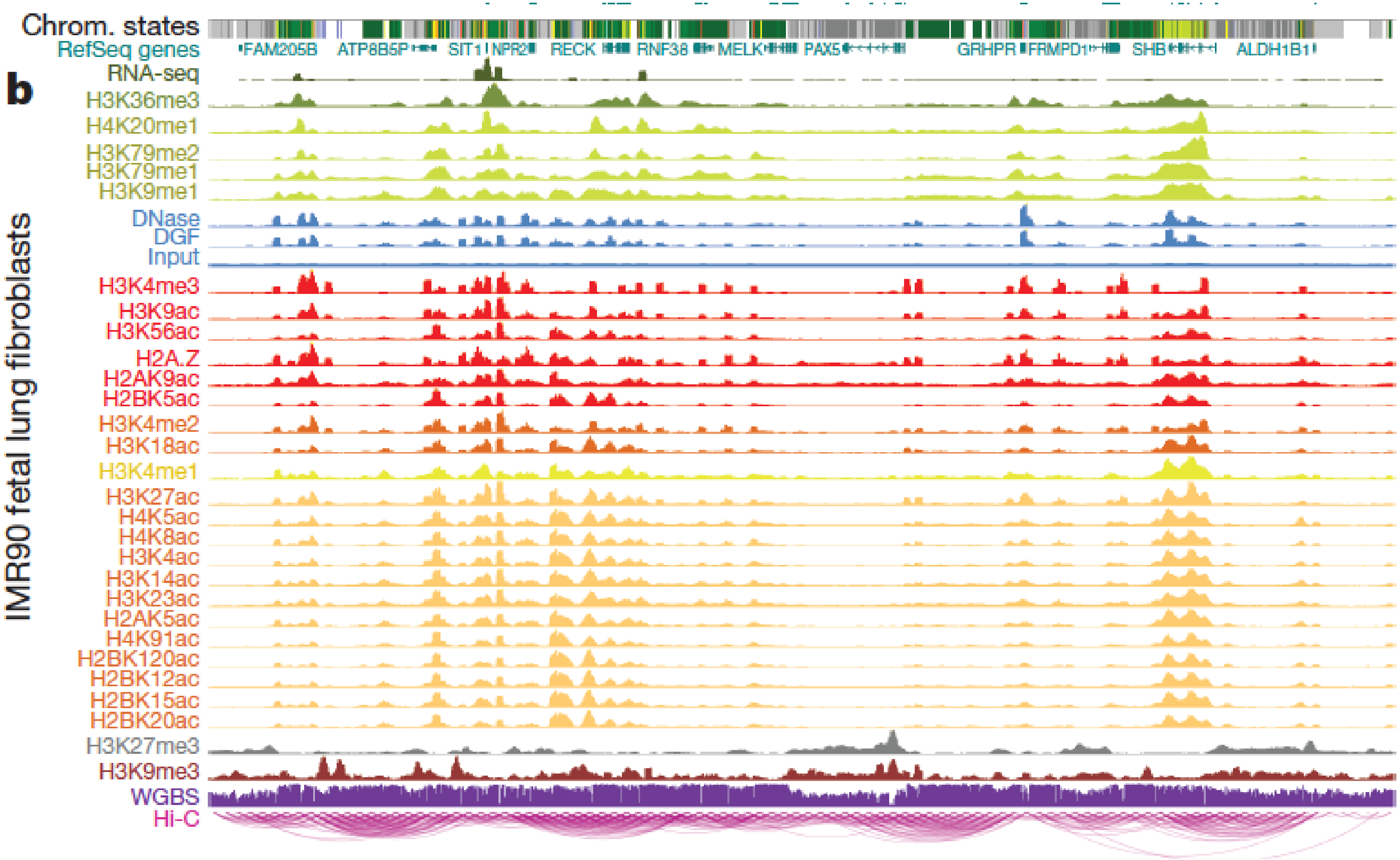

H3K4me2
H3K4me3
Acetylation
H2A.Z
RNAPII
H3K36me3
H3K79me2
Exon
Intron
Exon
p300
Enhancer
H3K4me1
H3K4me2
H3K27ac
CTCF
Boundary
Exon
Intron
Exon
H3K9me2
H3K9me3
H3K27me3


a
H3K4me3
H3K27me3
H2A.Z
Poised
Inactive
Active
H3K27me3
H3K9me3
DNAme
H3K4me2
H3K4me3
Acetylation
H2A.Z
Promoters
c
?
H3K9me2
H3K9me3
DNAme
Inactive
Active
H3K4me1
H3K4me2
H3K27ac
H2A.Z
Enhancers


b
Active
H3K36me3
H3K79me2
Inactive
H3K9me2
H3K9me3
Nucleosomes
Low
High
Exon
Intron
Exon
Gene bodies
d
?
H3K9me2
H3K9me3
DNAme
LADs and
LOCKs
Stable
Transient
H3K27me3
Pc bodies
Large-scale repression

# 2. Learning chromatin states

# Can we find latent state to explain observed marks?




Roadmap Epigenomics Consortium *et al. Nature* **518**, 317-330 (2015) doi:10.1038/nature14248

# Hidden Markov Models

Hidden state x in [1 .. m]

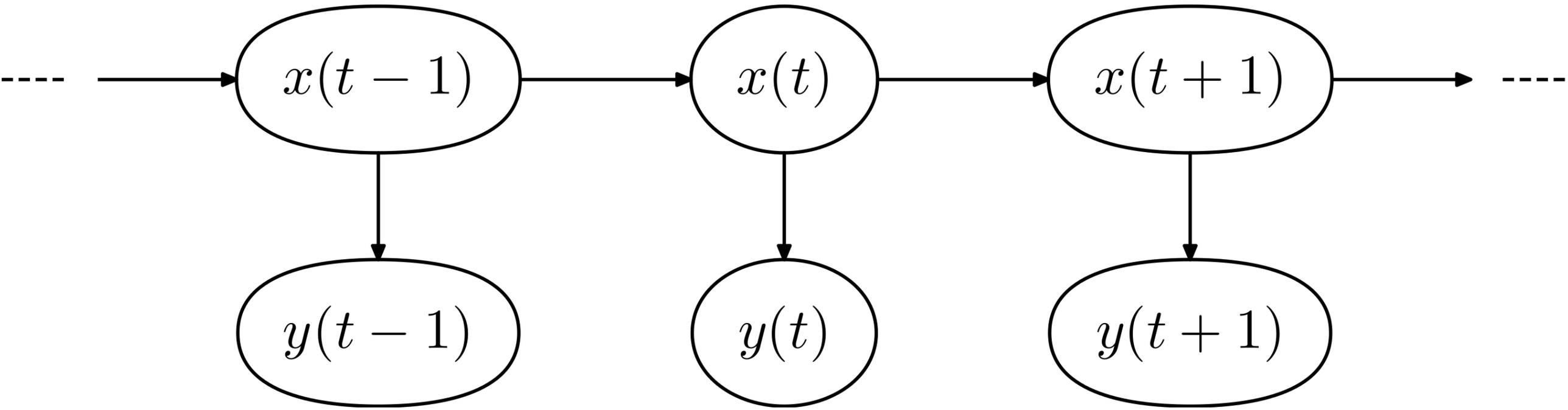
For example, m can 15

Emitted symbol y can be multi dimensional

For example, histone and accessibility data at genomic locus t

One node every 200bp down genome

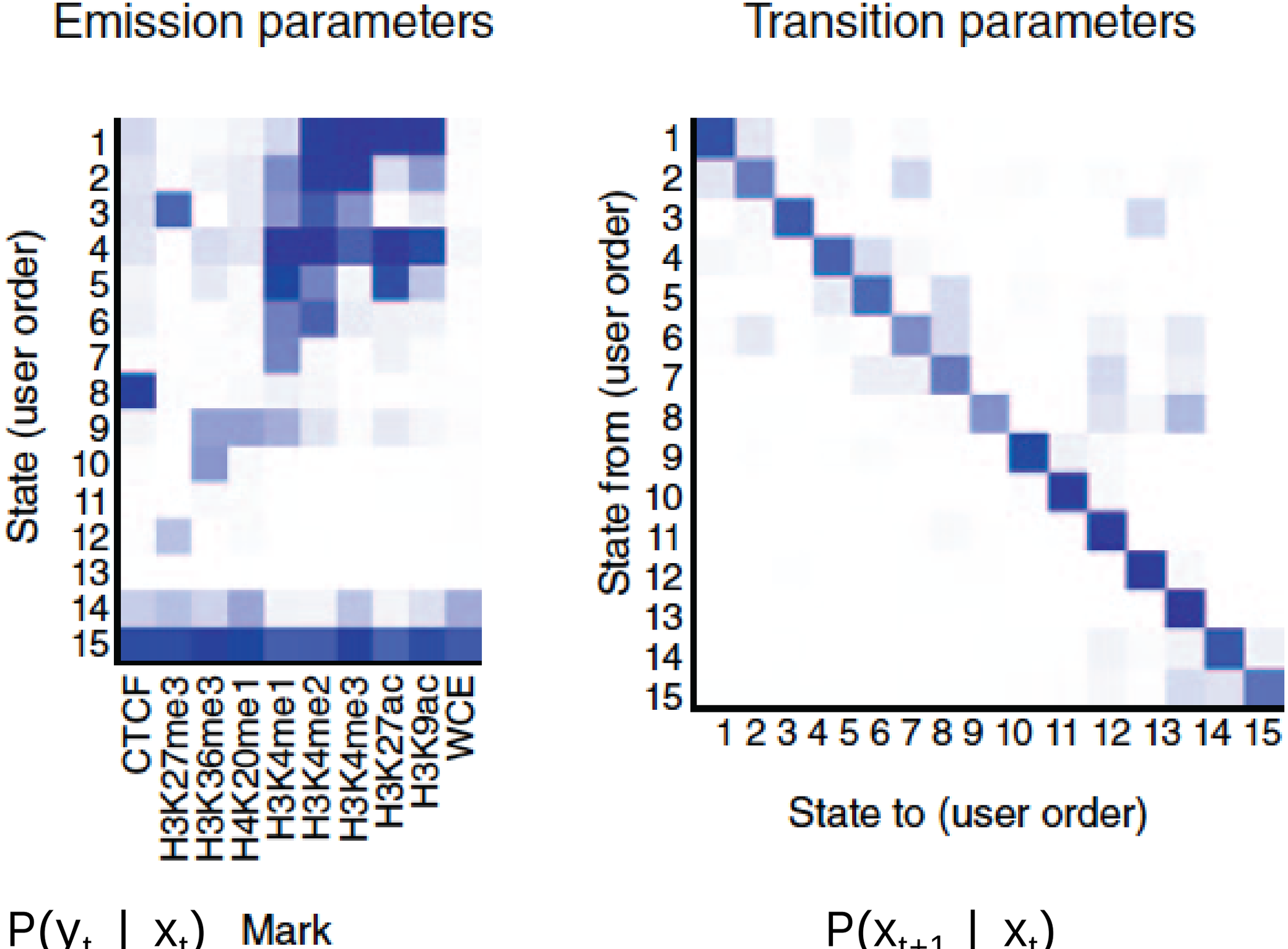
Parameters are $P(x_{t+1} \mid x_t)$, $P(y_t \mid x_t)$



# Hidden Markov Models can be used to create latent states that generate chromatin marks

Hidden Markov Model (ChromHMM)
Divide genome into 200bp windows
Hidden state for a 200bp window models what histone marks are present in the window
Unsupervised – resulting states must be interpreted with independent data
The number of states is fixed and is a modeling decision

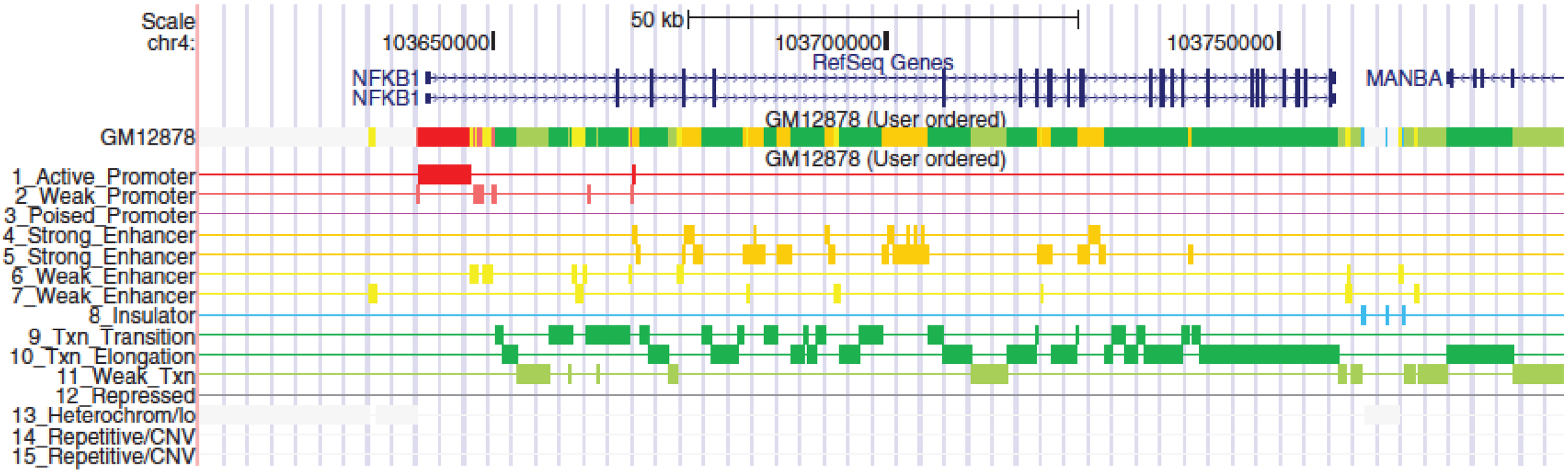
# ChromHMM Model Parameter Visualization.



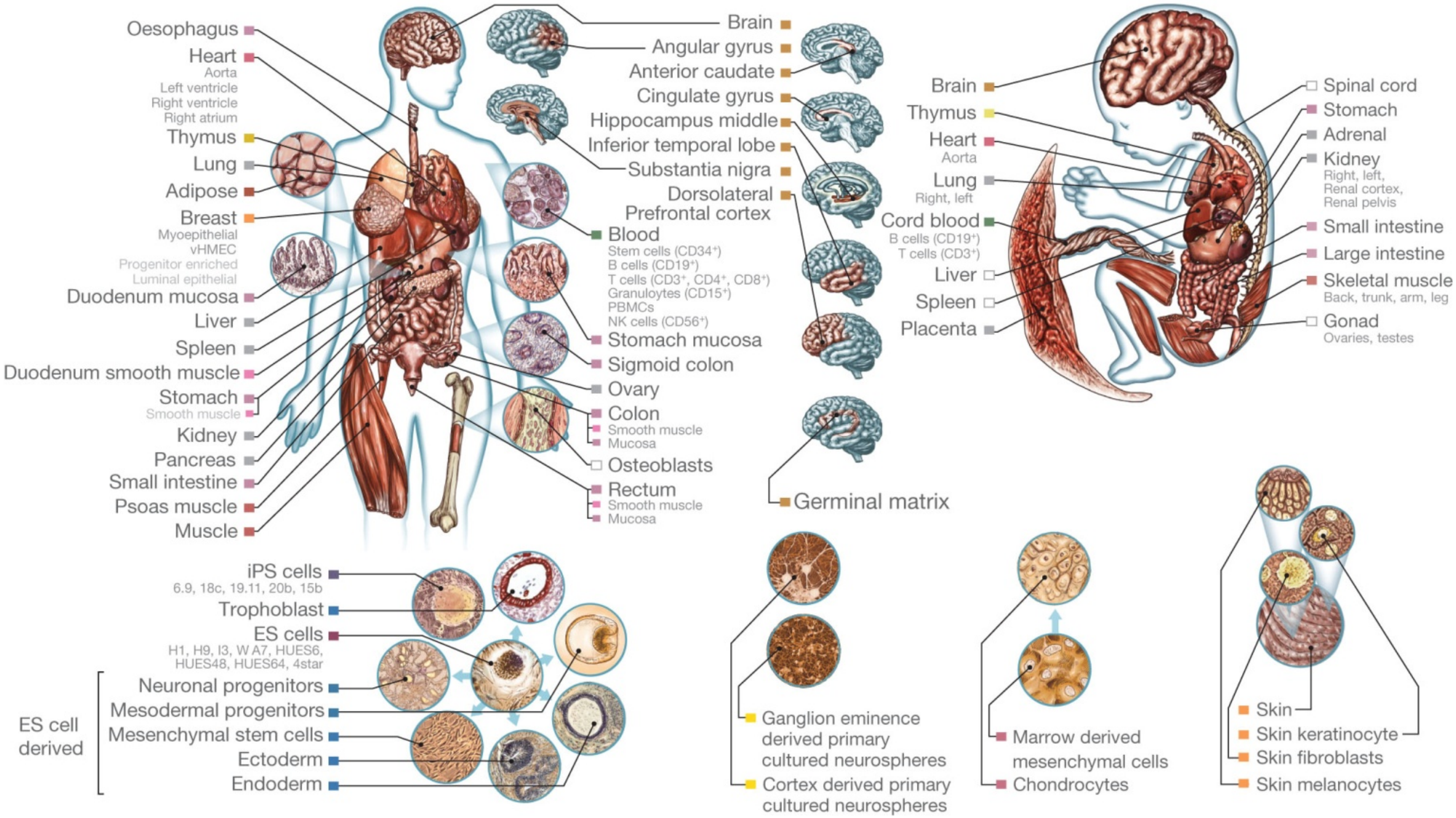

$P(y_t \mid x_t)$

$P(x_{t+1} \mid x_t)$

# ChromHMM segment based chromatin states



# Tissues and cell types profiled in the Roadmap Epigenomics Consortium.

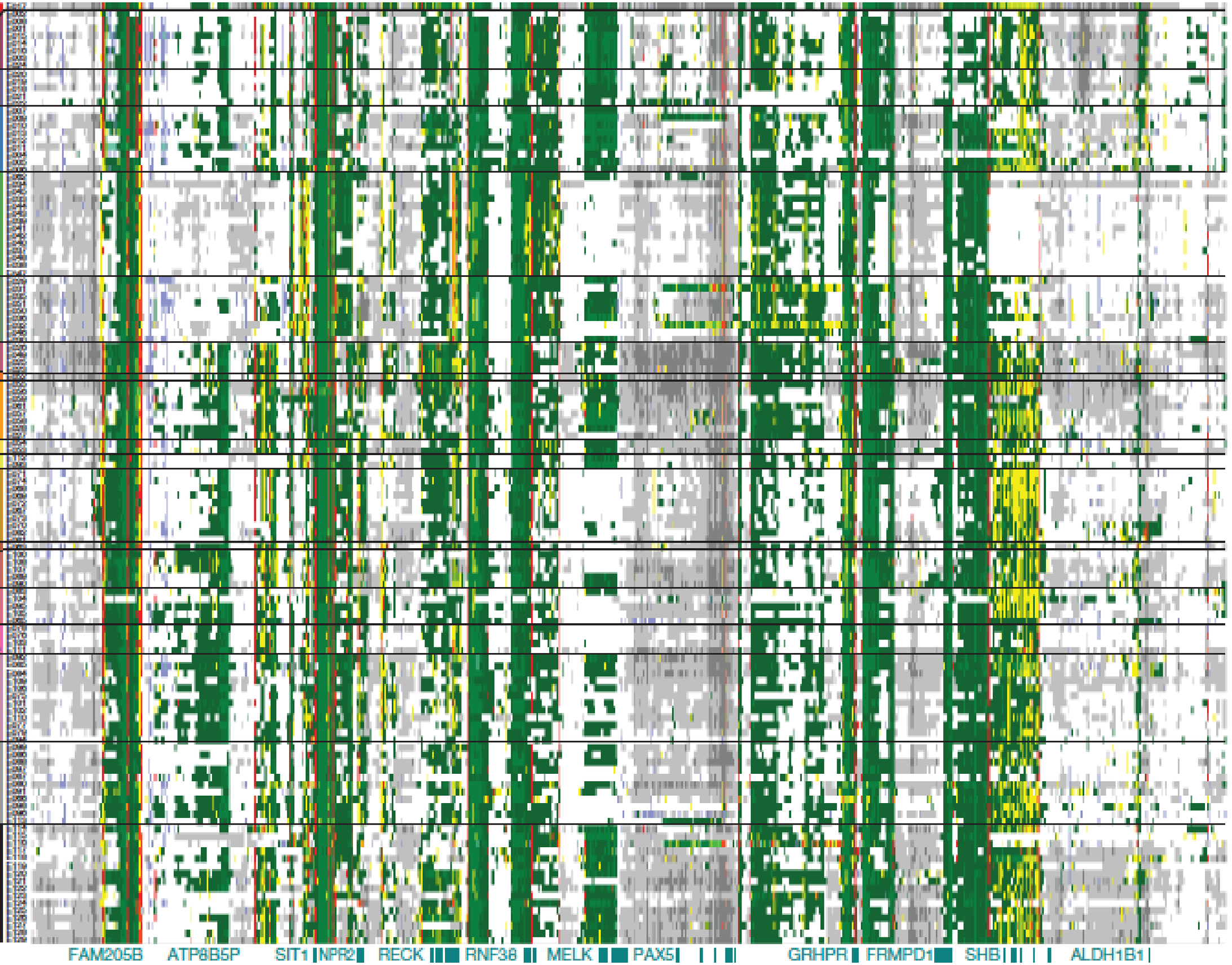



Roadmap Epigenomics Consortium *et al. Nature* **518**, 317-330 (2015) doi:10.1038/nature14248

**a**

Chromatin state annotations in 127 epigenomes

IMR90
ES cell
iPSC
ES-deriv.
Blood & T cell
HSC & B cell
Mesench.
Epithelial
Brain
Muscle
Heart
Sm. musc.
Digestive
Other
ENCODE 2012

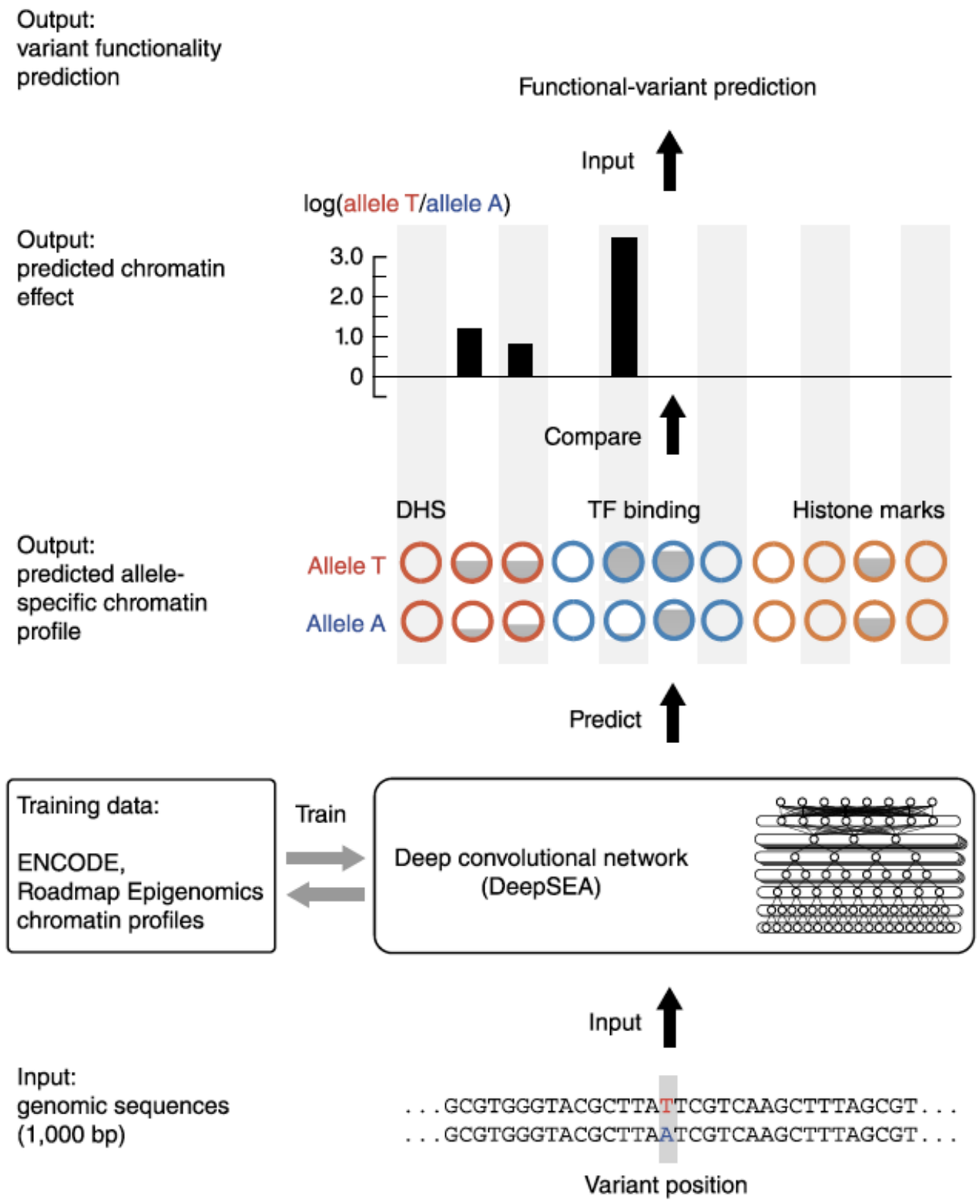


FAM205B ATP8B5P SIT1 NPR2 RECK RNF38 MELK PAX5 GRHPR FRMPD1 SHB ALDH1B1

Roadmap Epigenomics Consortium *et al. Nature* **518**, 317-330 (2015) doi:10.1038/nature14248

# 3. Predicting chromatin state from sequence

# DeepSea learns TF binding, accessibility, and chromatin marks

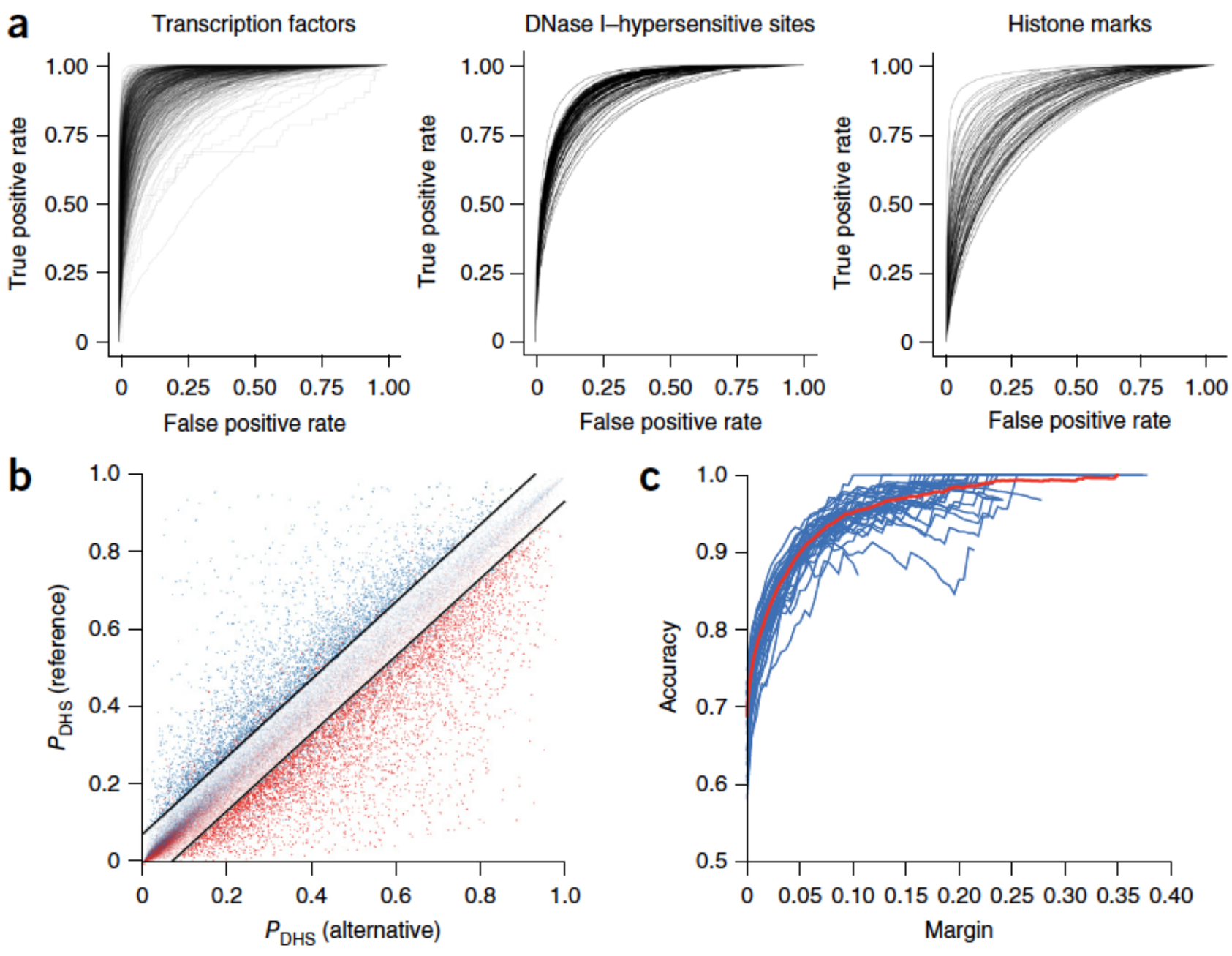



17% of genome
690 TF binding profiles for 160 different TFs, 125 DHS profiles and 104 histone-mark profiles
Chr 8 and 9 excluded

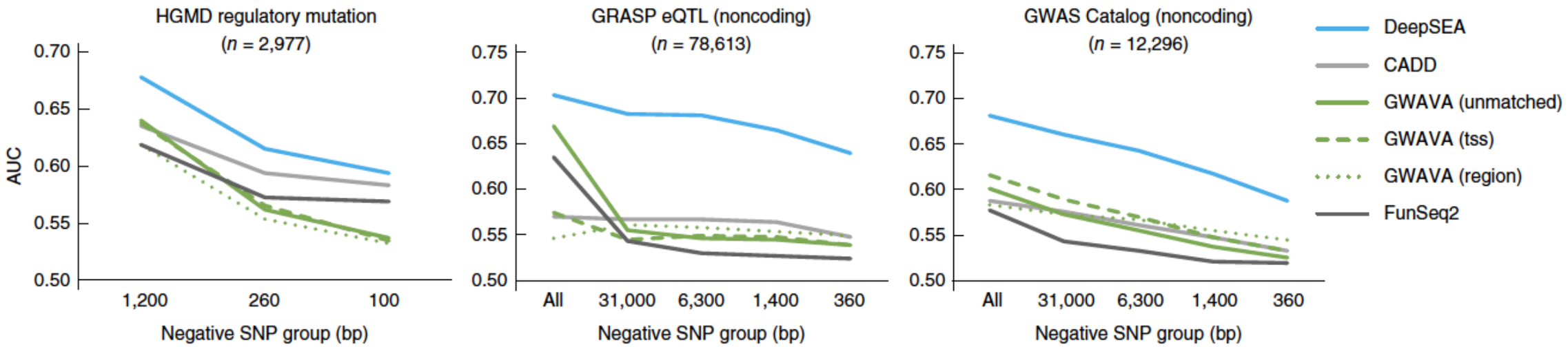
125 DNase features, 690 TF features, 104 histone features

three convolution layers with 320, 480 and 960 kernels
1000 bp window

# DeepSea can predict differentially accessible regions based upon SNP value



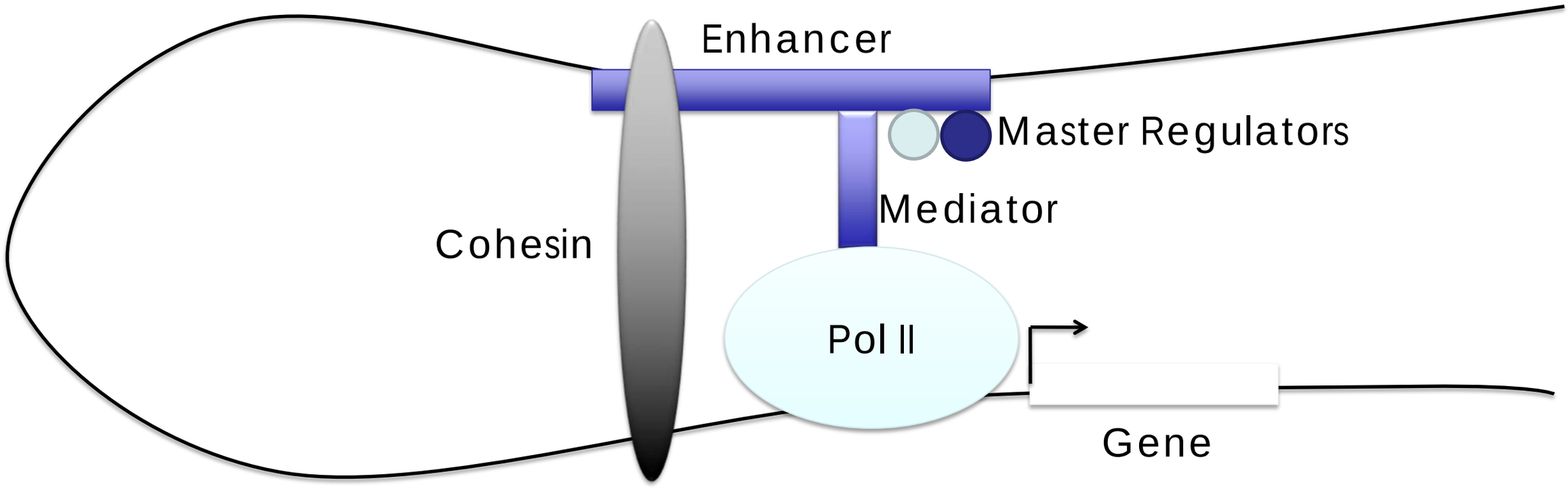
# An ensemble logistic regression classifier based on DeepSea output can identify regulatory variants



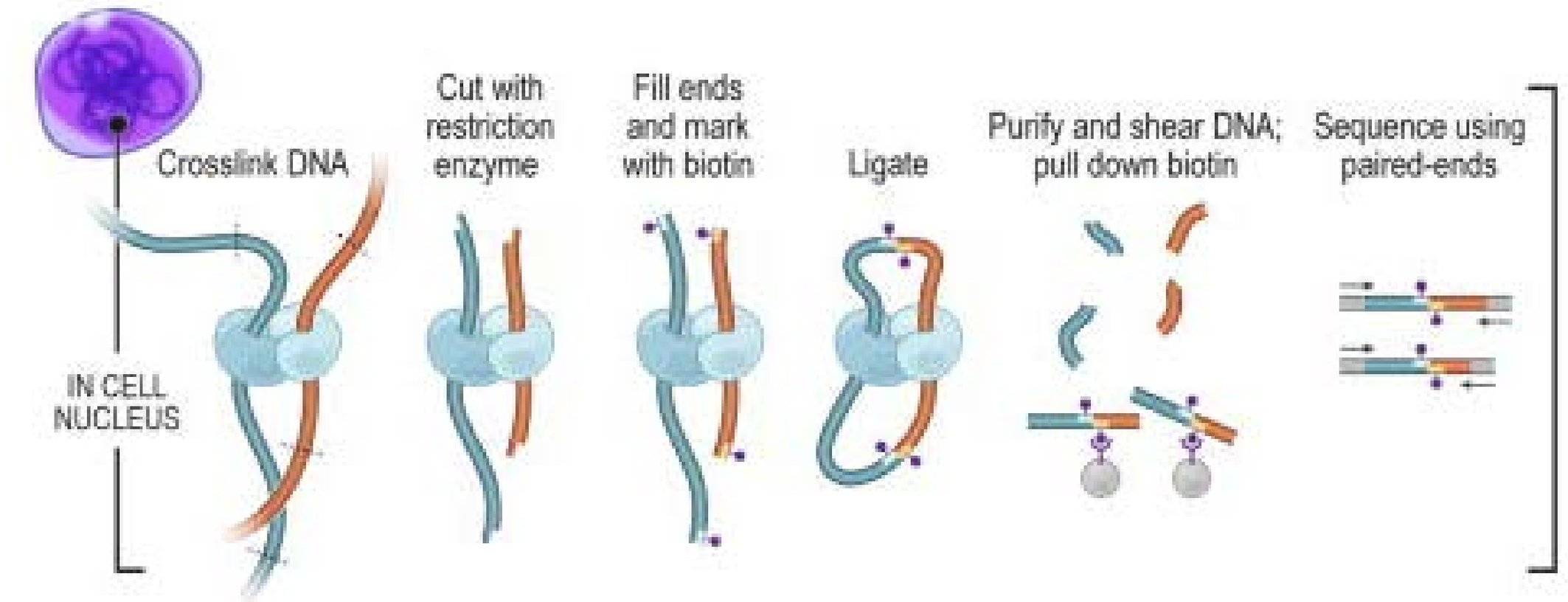
# 4. Three-dimensional interactions

# HiC, HiChip, and ChIA-PET data reveal distal genome interactions

# Enhancers regulate distal target genes by genome looping

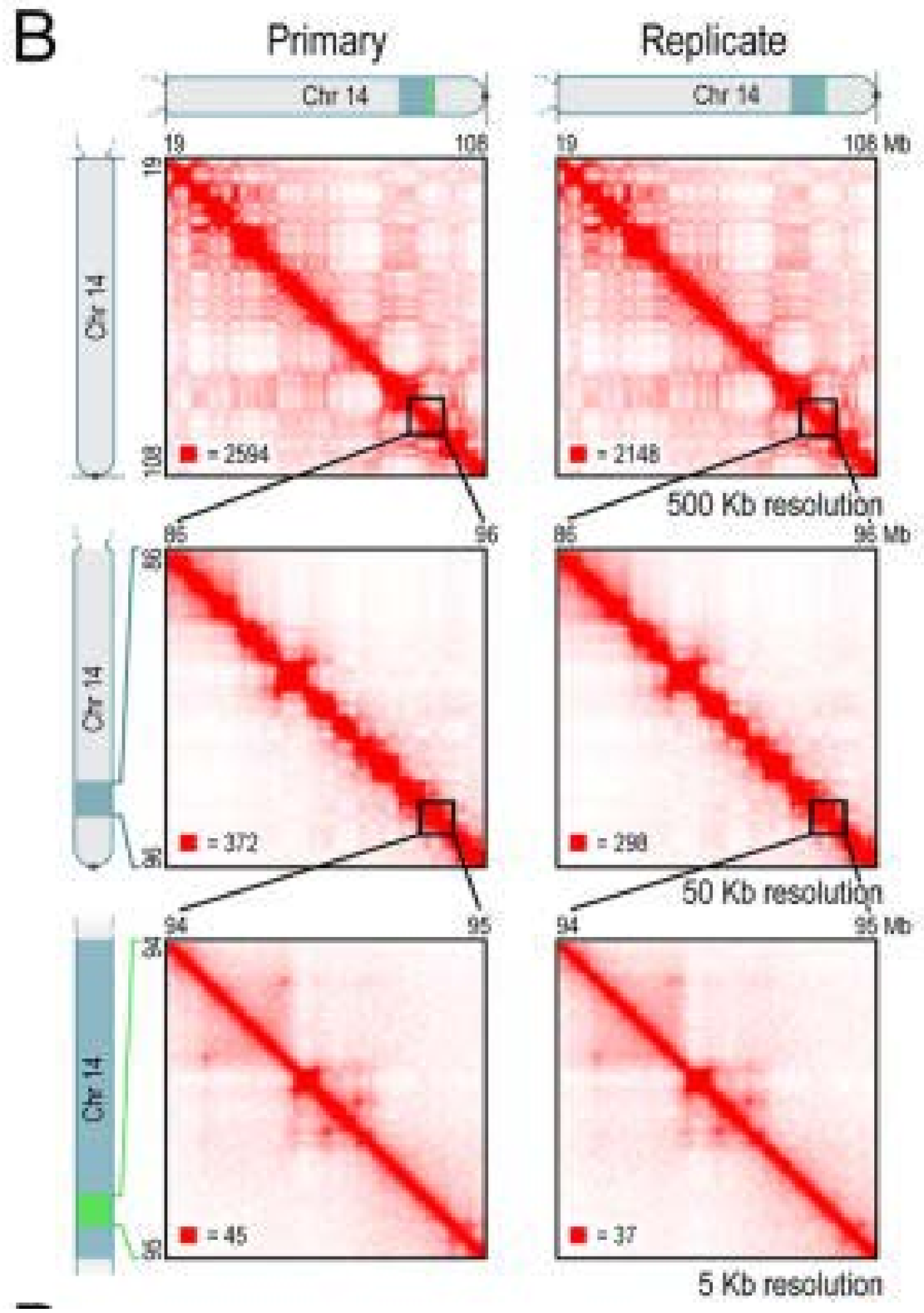


# *in situ* HiC identifies proximal genomic contacts

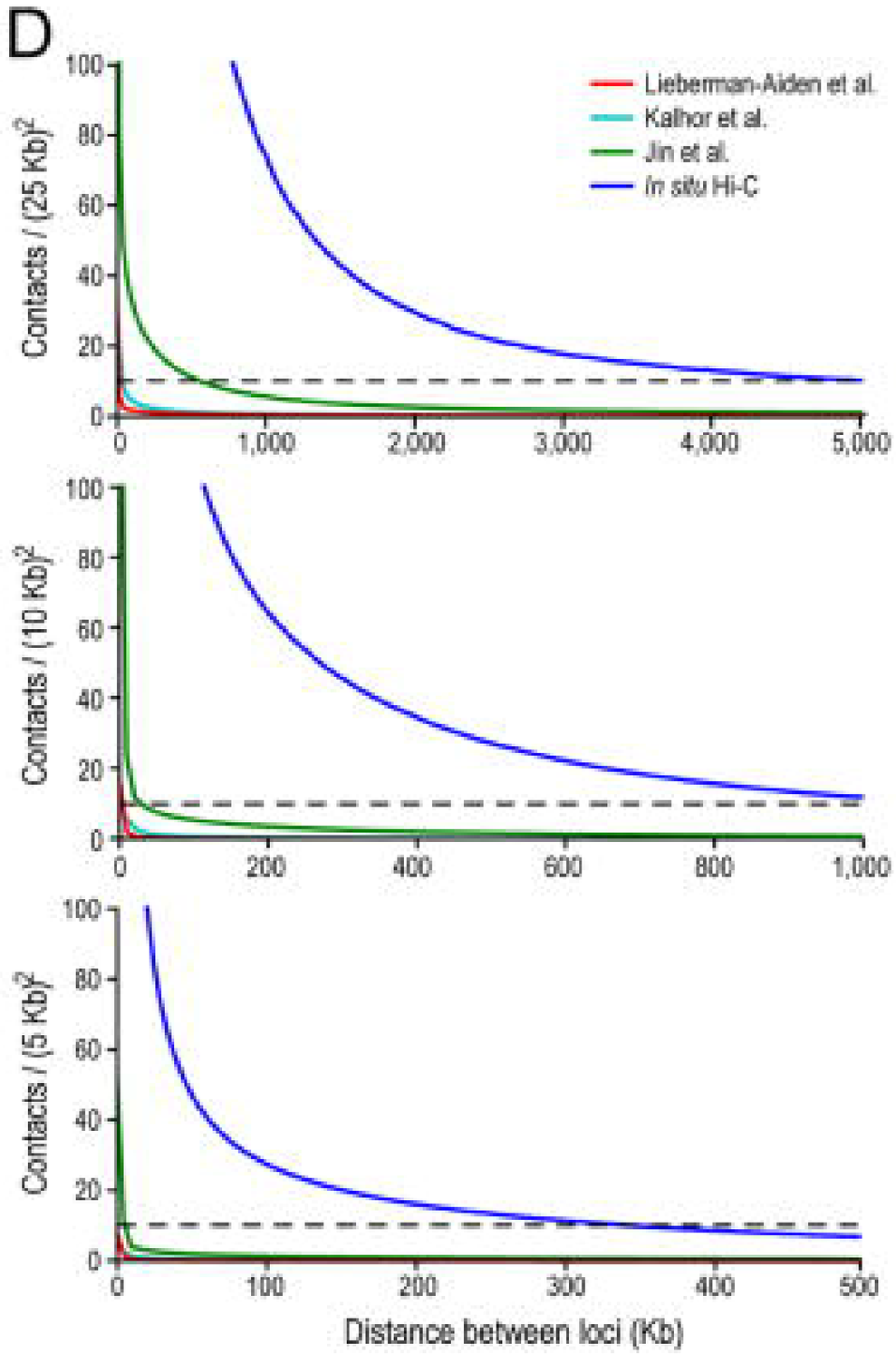



Cell. 2014 Dec 18; 159(7): 1665–1680.

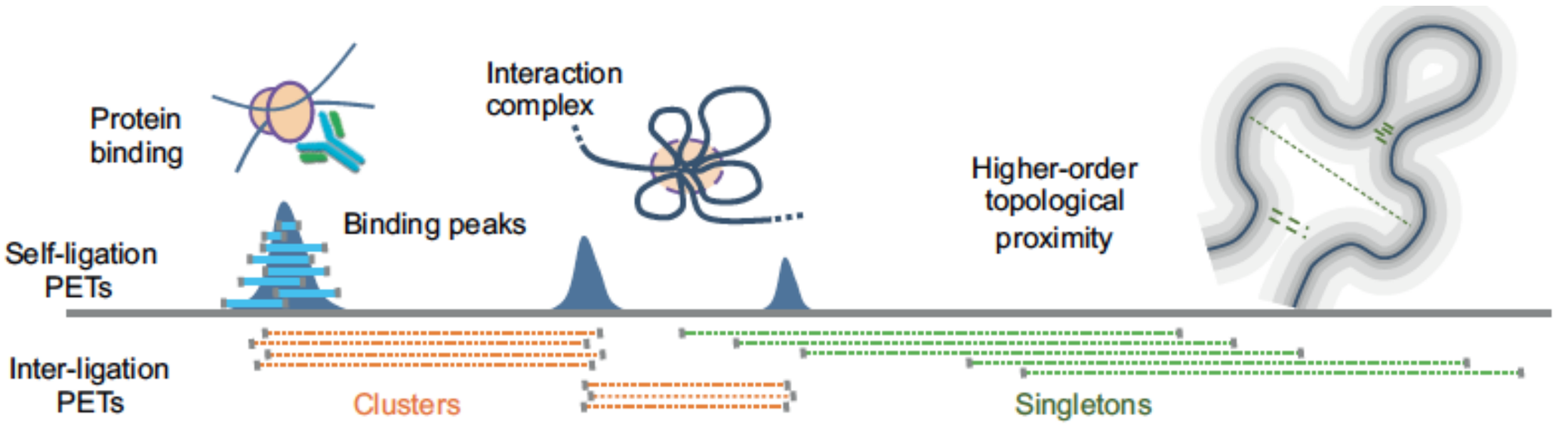
# *in situ* HiC reveals interactions at 1 – 5 KB resolution



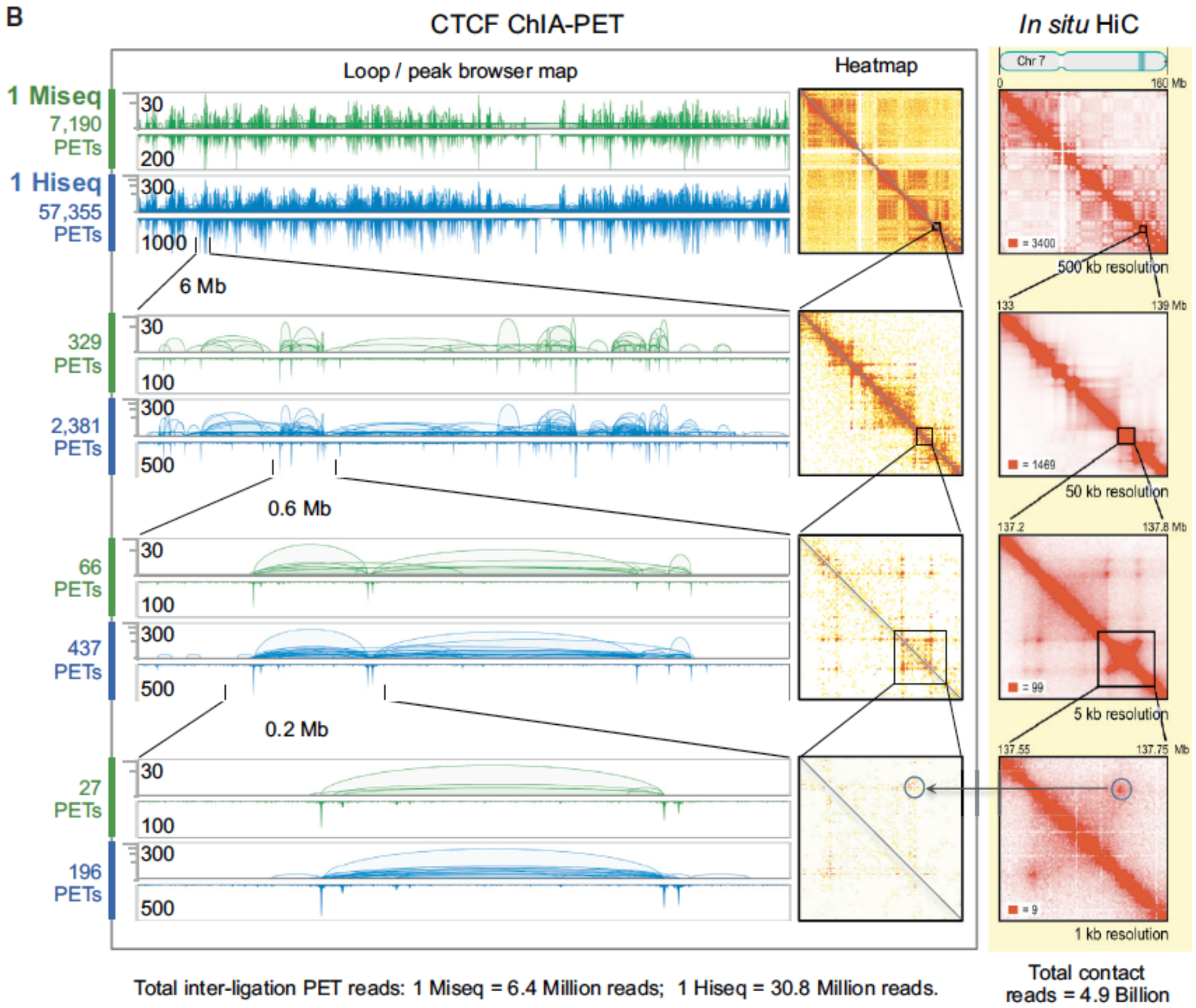
# Observed interchromosomal interaction distances fall off exponentially



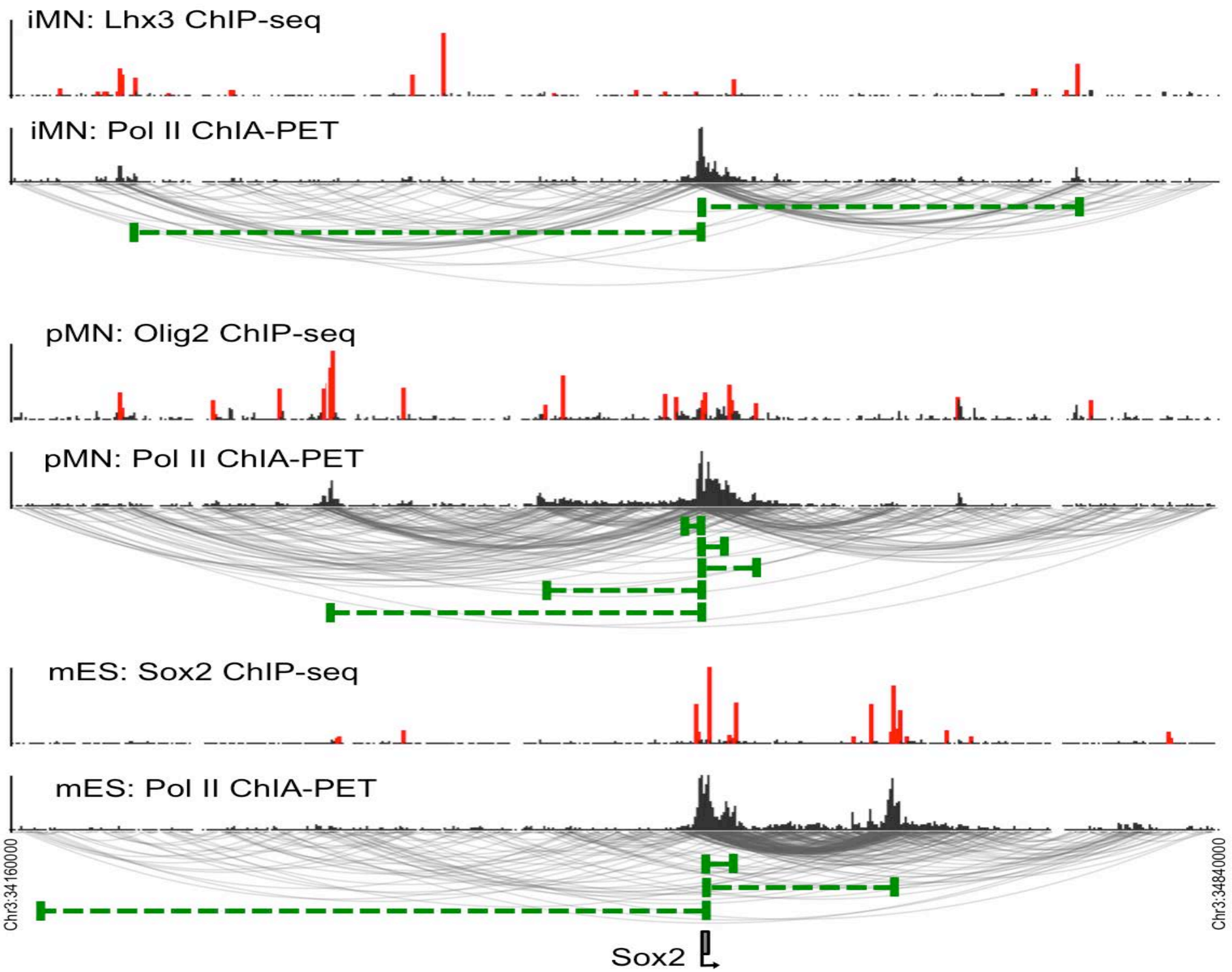
# ChIA-PET identifies protein mediated interactions and improves resolution for those events



# ChIA-PET data are consistent with HiC data



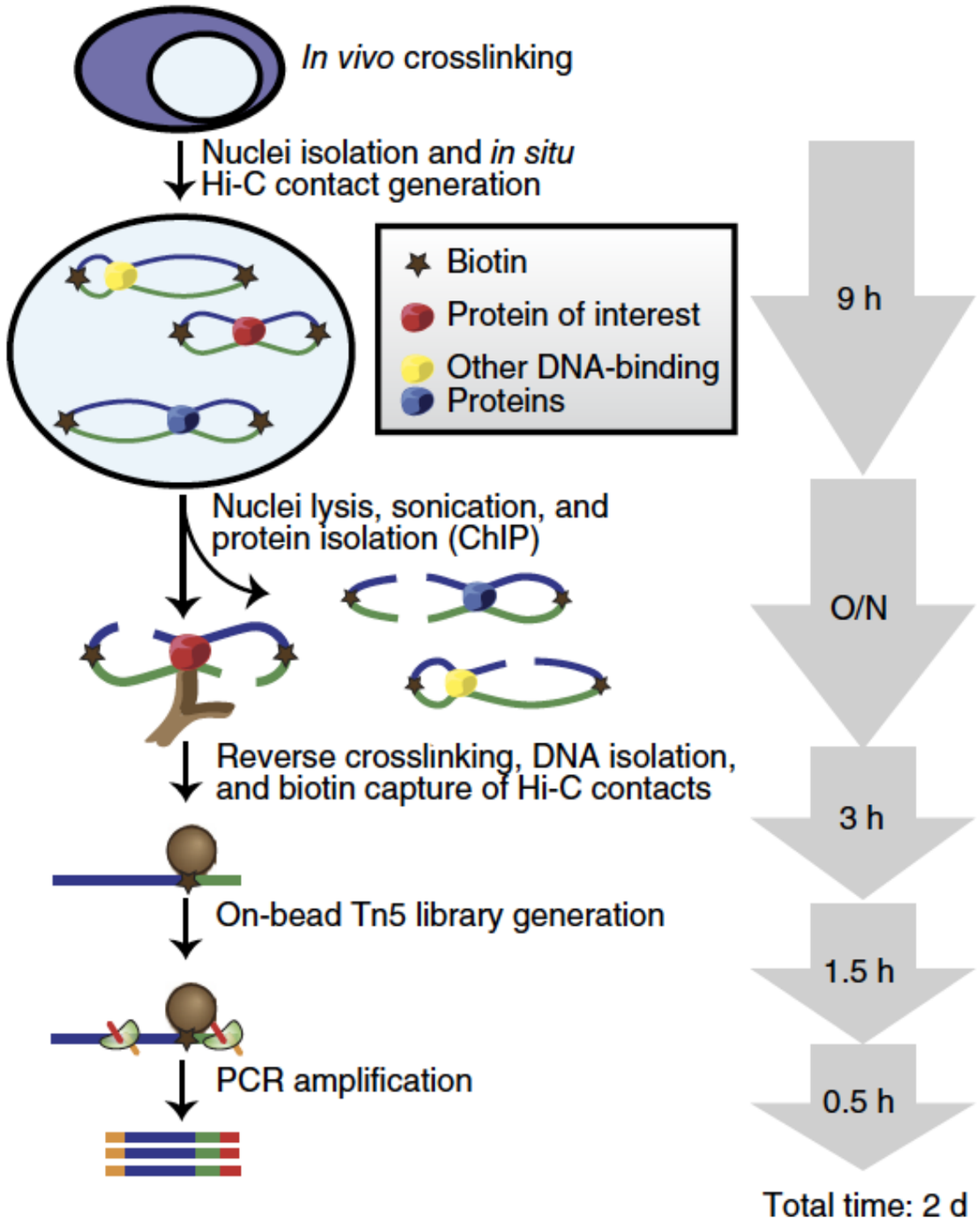
# ChIA-PET discovered enhancer linkages



# Issues with ChIA-PET

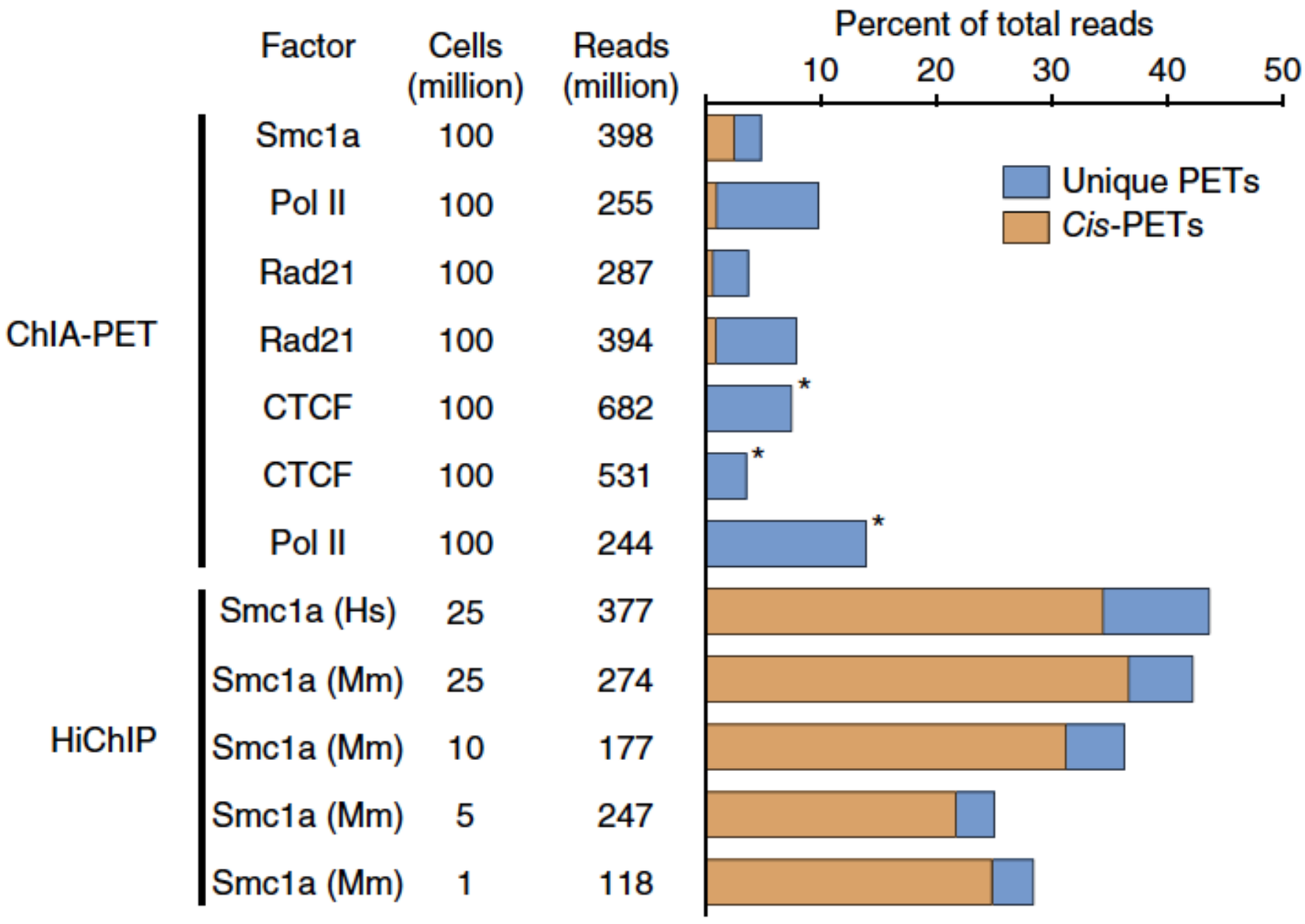
1. High false negative rate. Libraries produced are not complex enough to permit further discovery by additional sequencing.
2. Specific to a protein (RNA Polymerase II in our example)
3. Hi-C and derivatives may solve these problems eventually

# HiChIP identifies protein mediated interactions

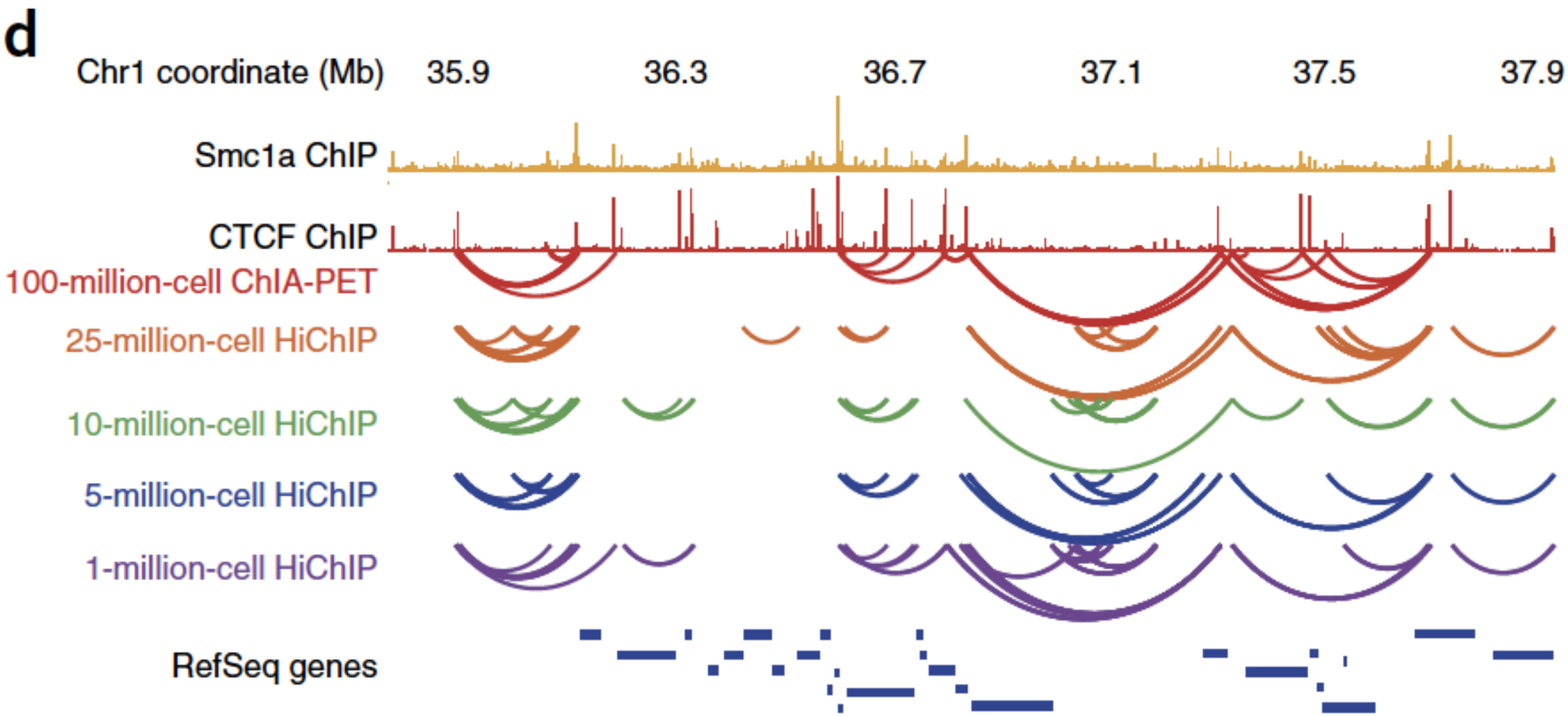


# HiChIP is more sensitive than ChIA-PET

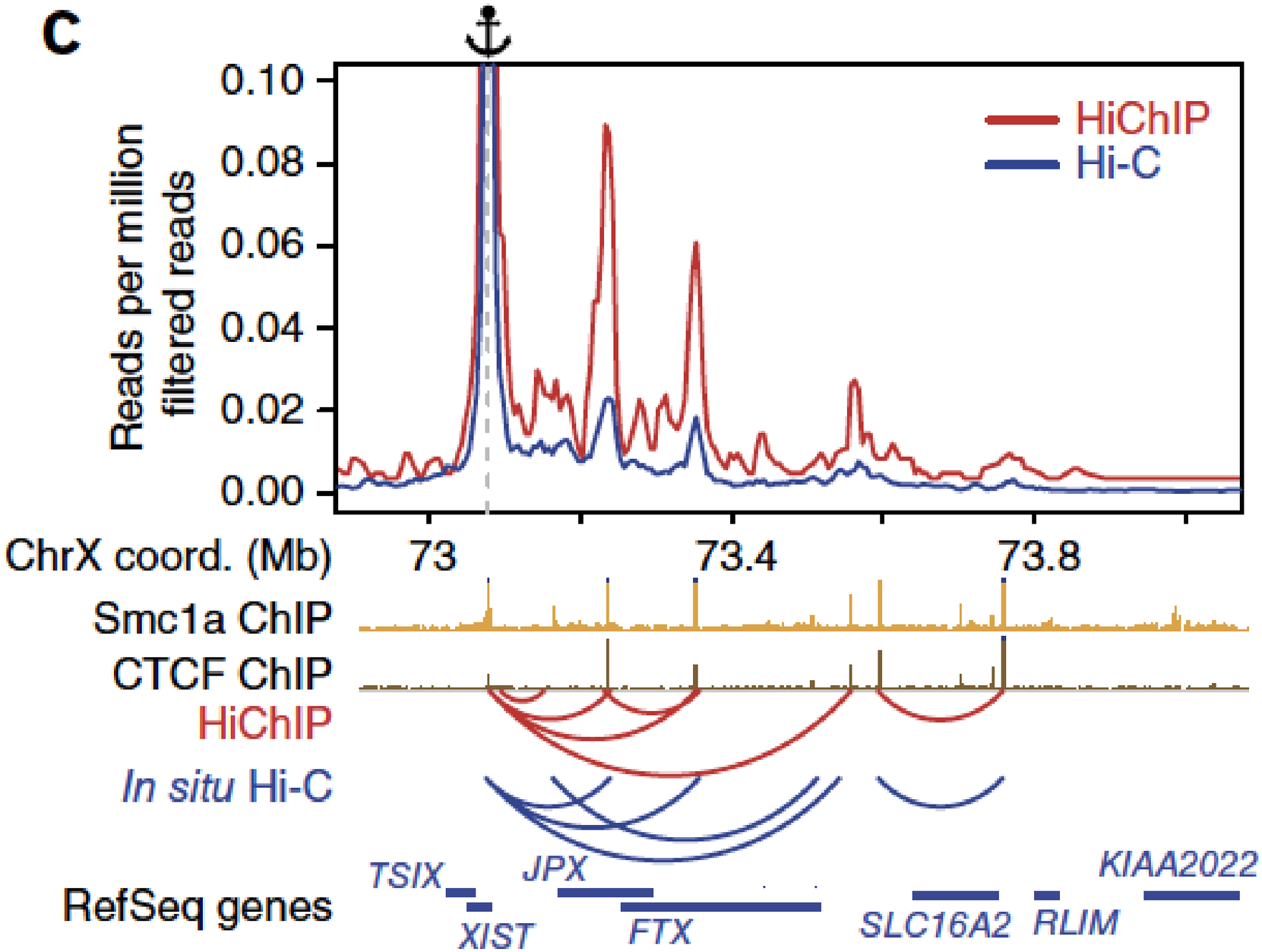
**b**



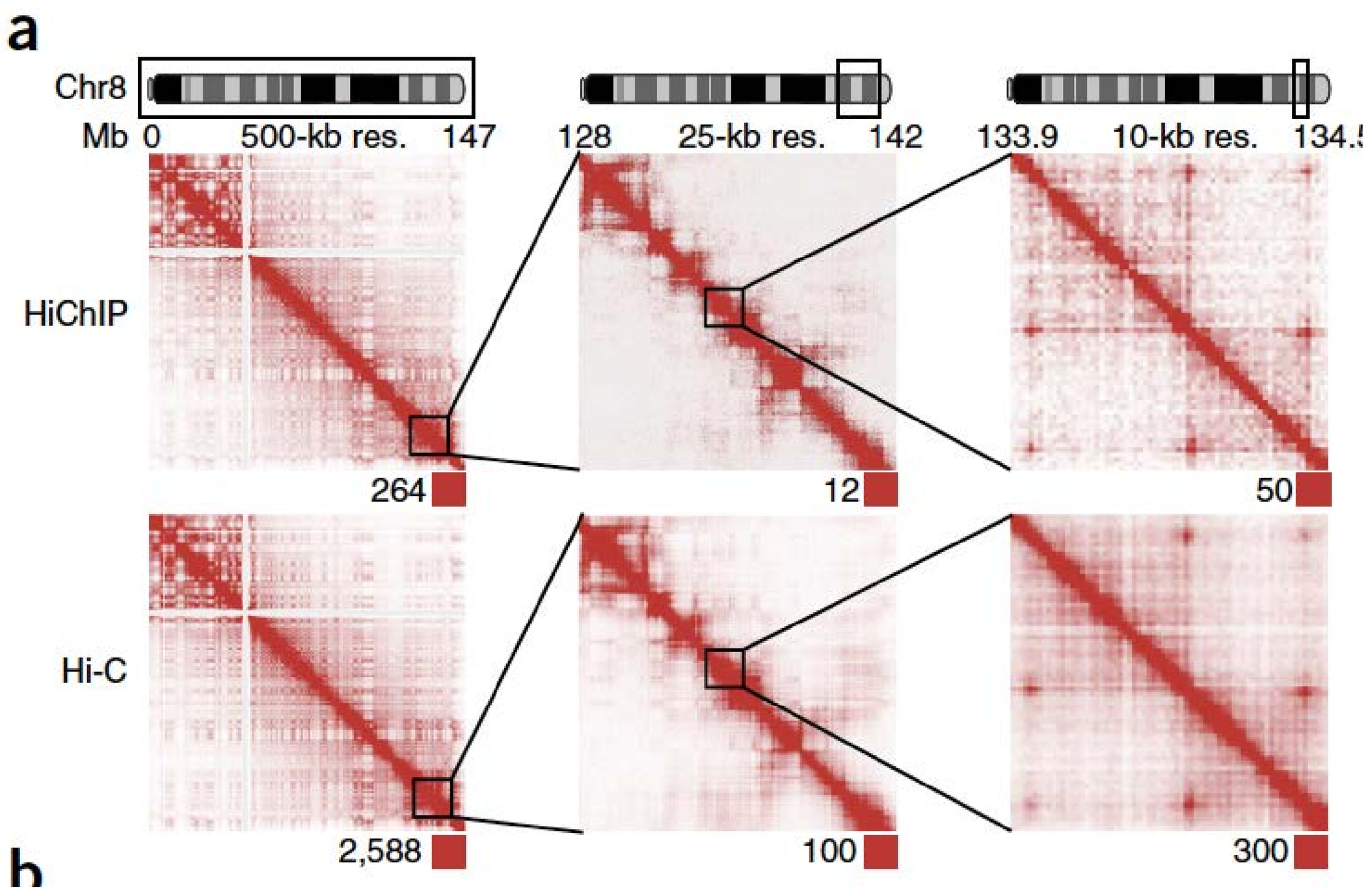
# HiChIP and ChIA-PET interactions compared Smc1a antibody (part of cohesion complex)



# XIST promoter interactions show more support from HiChIP than Hi-C



# HiChIP (Smc1a) is more sensitive than HiC



# 5a. Discovering interactions: Anchor-based

# Method 1: Discover anchors using ChIP-seq methods

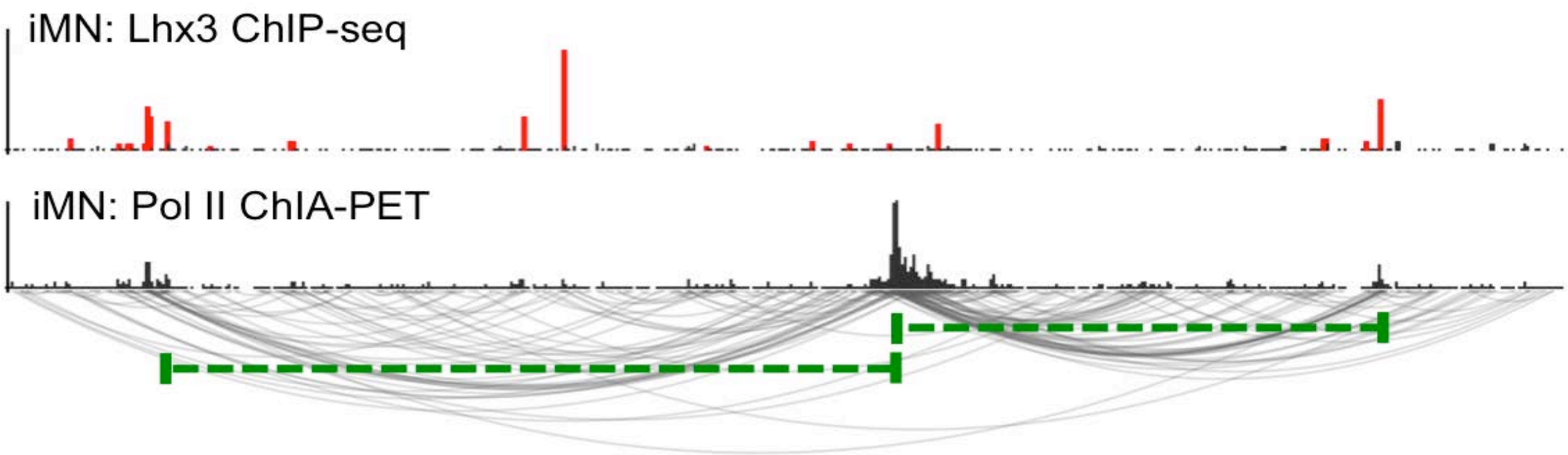
Given anchors, what is the chance of observing an interaction by chance?

N total ends

$l_{a,b}$ interactions observed

$c_a$ ends

$c_b$ ends



# What is the chance of observing an interaction by chance?

$$P(I_{A,B}|N,c_A,c_B) = \frac{\binom{c_A}{I_{A,B}}\binom{N-c_A}{c_B-I_{A,B}}}{\binom{N}{c_B}}$$

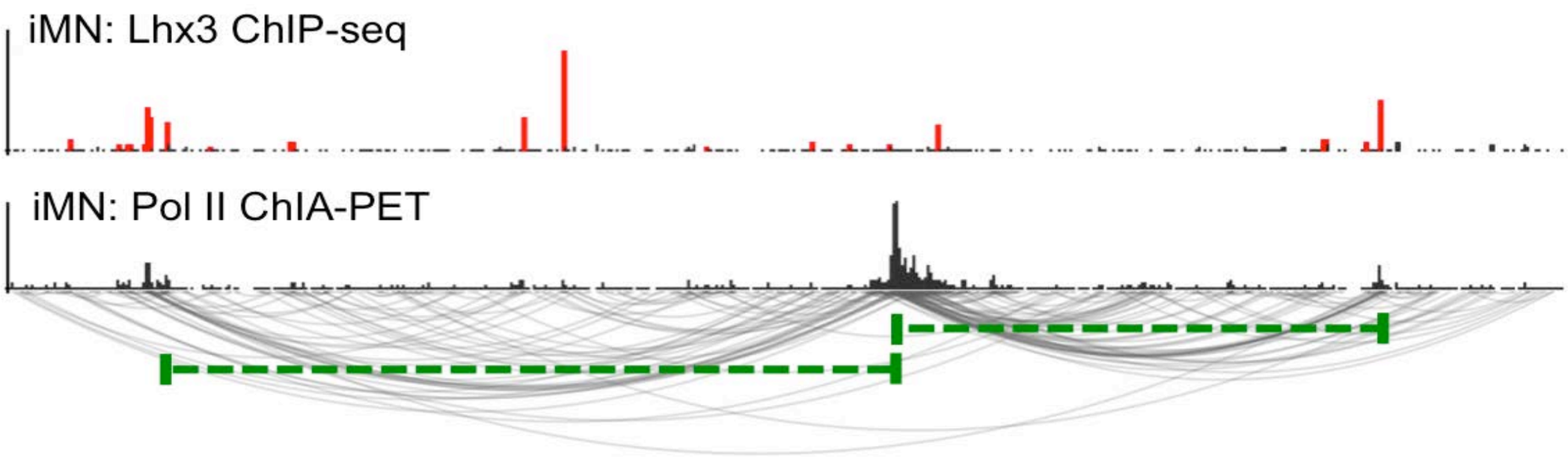
$$\phi = \sum_{i=I_{A,B}}^{\min\{c_A,c_B\}} P(i|N,c_A,c_B)$$

N total ends

$I_{a,b}$ interactions observed

$c_a$ ends

$c_b$ ends



# Estimating total events from overlap

Imagine we perform two biological replicates of an experiment and obtain 1000 events in each, of which 900 are identical.

We can use a hypergeometric model to infer how many possible events exist (*N*) given two sample sizes (*m* and *n*) and an overlap (*k*):

$$\hat{N} = \underset{N}{\operatorname{argmax}} \left[ P(X = k; N, m, n) \right]$$

Using this model, we predict ~1100 total events

# Approximate closed form solution for total number of events

The ML estimate of *N* is approximately:

$$\hat{N}(m,n,k) = \frac{mn}{k}$$

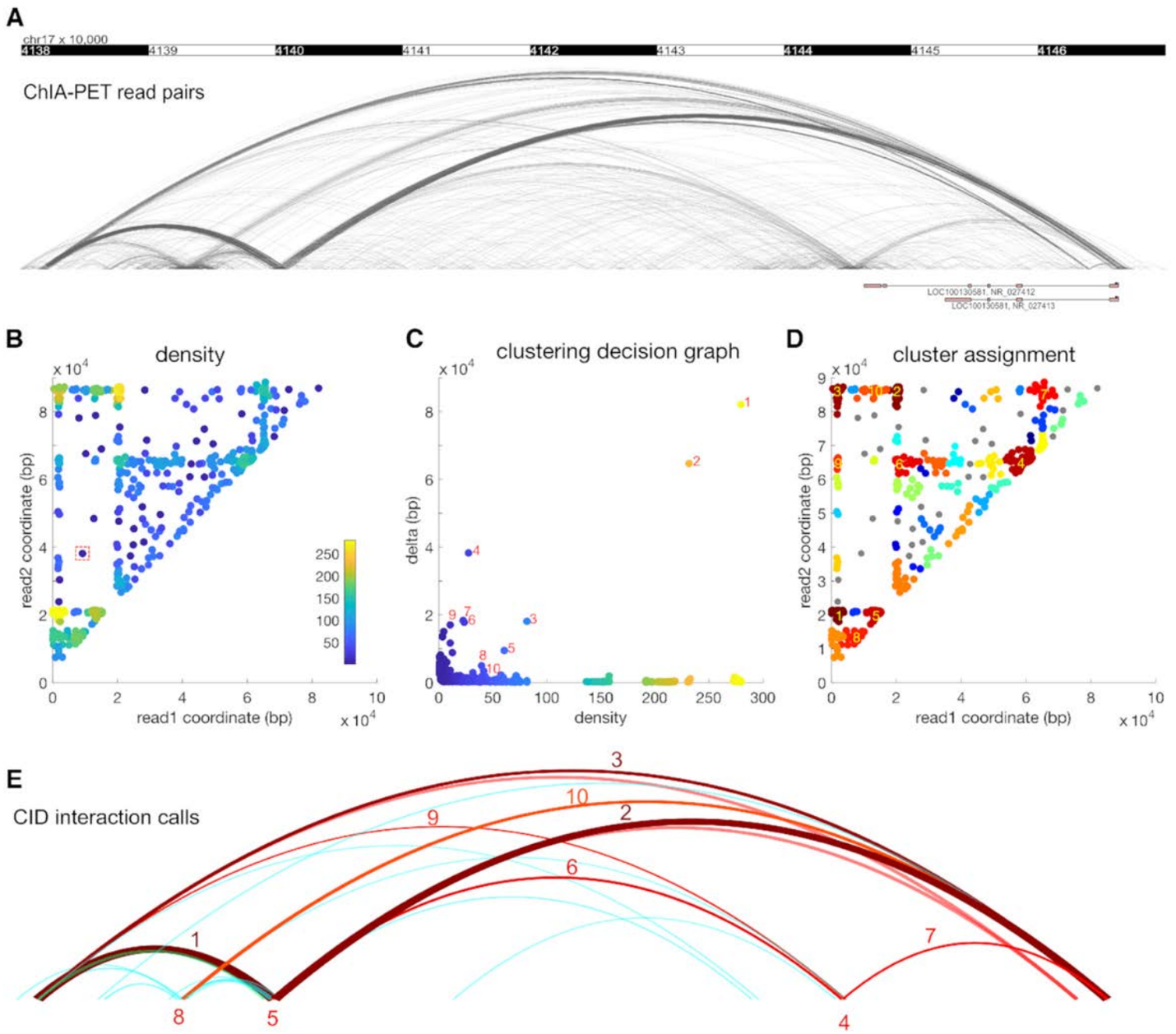
One way to see this is by using the normal approximation of the binomial approximation to the hypergeometric distribution:

$$P(X=k;N,m,n) \approx \text{Binomial}\left(X=k; n=n, p=\frac{m}{N}\right)$$

$$\approx \text{Normal}\left(X=k; \mu=\frac{mn}{N}, \sigma^2=\frac{mn}{N}(1-\frac{m}{N})\right)$$

# 5b. Discovering interactions: Density-based

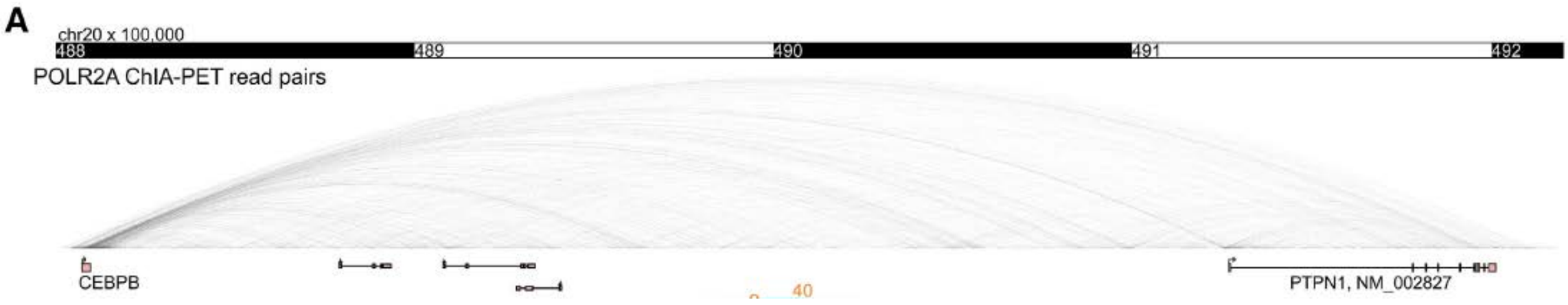
# Method 2: CID uses density-based clustering to discover chromatin interactions




Nucleic Acids Research, 14 February 2019, gkz051, https://doi.org/10.1093/nar/gkz051

•**Figure 1.** CID uses density-based clustering to discover chromatin interactions. (A) ChIA-PET interactions can be discovered as groups of dense arcs connecting two genomic regions. Each arc is a PET. (B) The PETs plotted on a two-dimensional map using the genomic coordinates of the two reads. Each point is a PET. The colors represent the density values, defined as the number of PETs in the neighborhood. The red dashed square represents the size of the neighborhood. (C) The clustering decision graph. Each point is a PET. The points with high density and high delta values are selected as cluster centers. For simplicity, only large clusters are labelled. (D) The read pairs are assigned to the nearest cluster centers. The clusters are labeled as in (C). (E) The clusters are visualized as arcs. The clusters are labeled as in (C) and (D).

# Method 2: Density cluster interaction origins




$$Distance\left(PET_i, PET_j\right) = max\left(\left|C_{i,L} - C_{j,L}\right|, \left|C_{i,R} - C_{j,R}\right|\right)$$

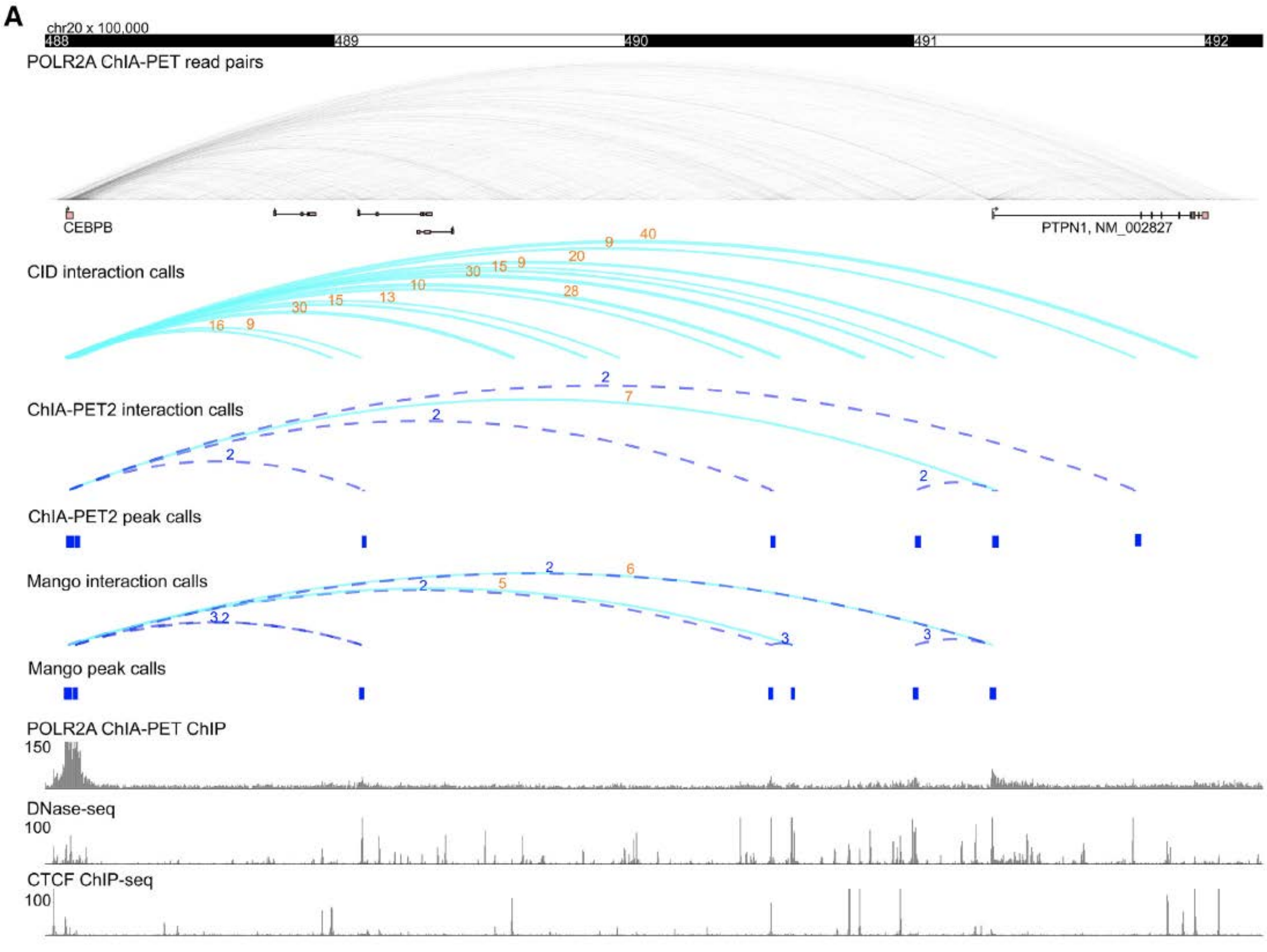
https://academic.oup.com/nar/advance-article/doi/10.1093/nar/g

We use a three-component mixture model to describe conditional distribution of PET-count from all the PET clusters. One component represents true interaction PET cluster (TiPC), and the other two for random collision PET cluster (RcPC) and random ligation PET cluster (RlPC), respectively.

TiPC and RcPC models include $d_{a,b}$ distance between clusters

https://academic.oup.com/bioinformatics/article/31/23/3

# Cluster interaction origins

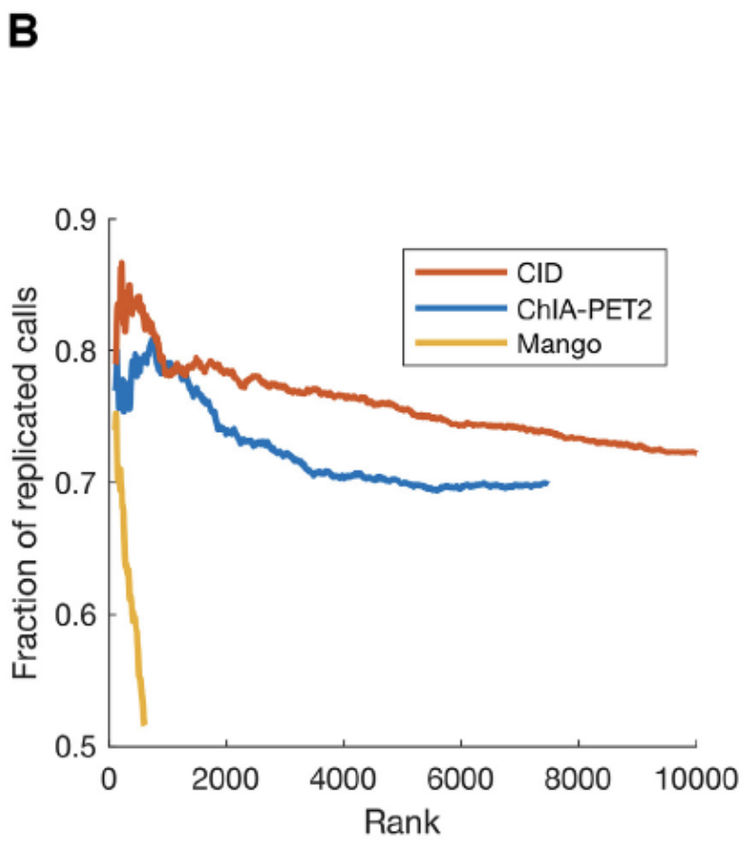


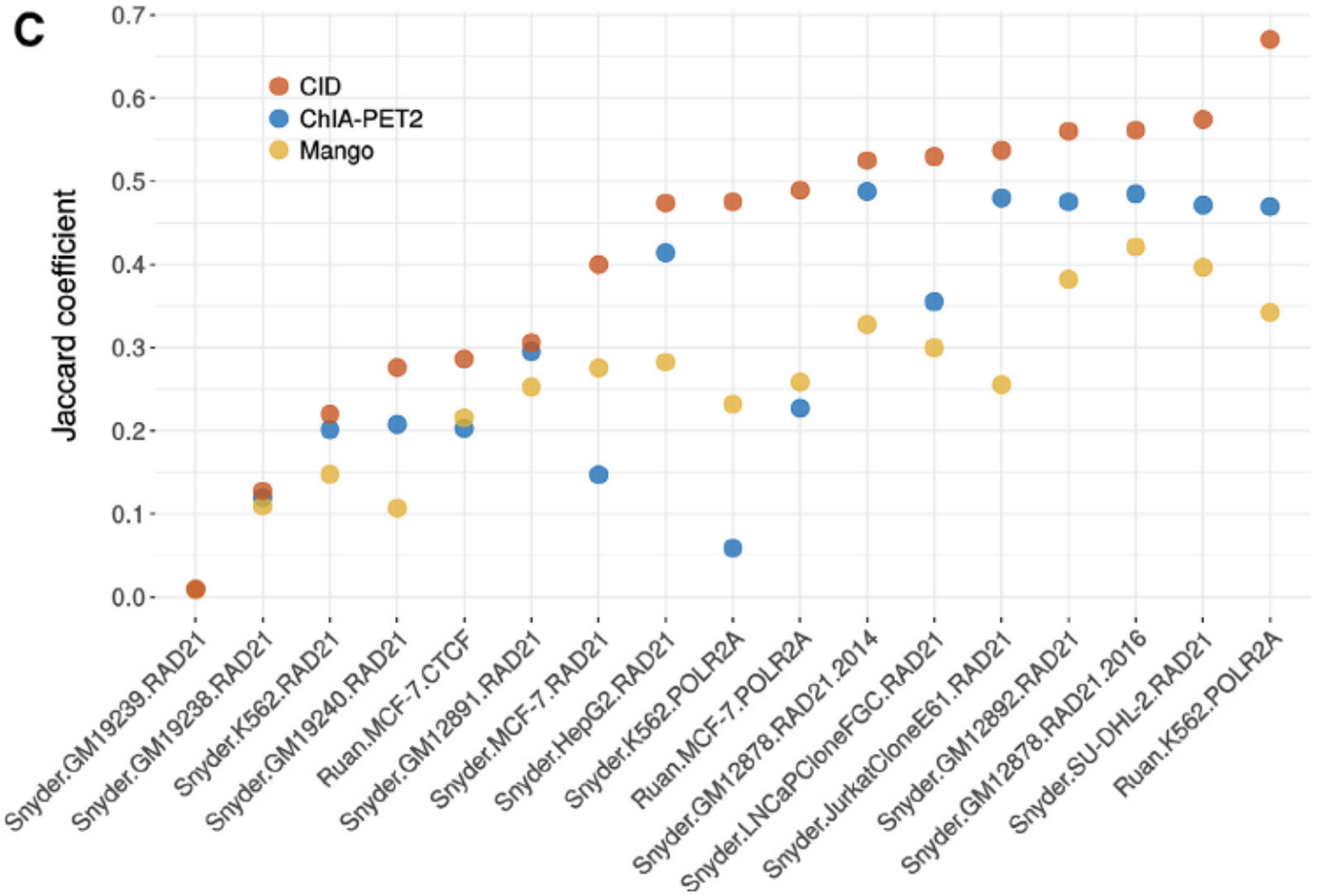
# Jaccard coefficient – measure of set similarity

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

$$0 \leq J(A,B) \leq 1.$$

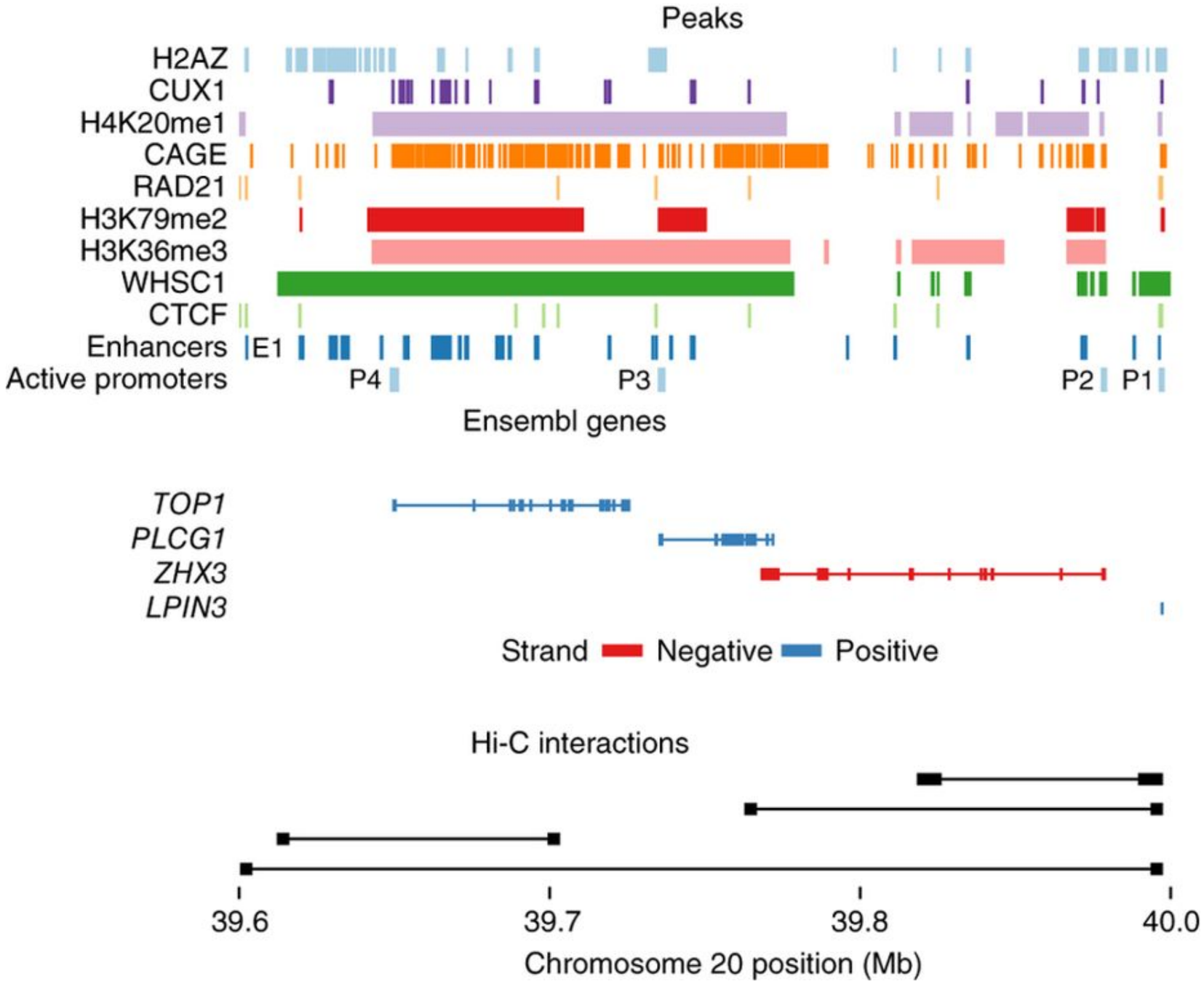
# CID is more reproducible and sensitive



# 6. Predicting enhancer-promoter interactions

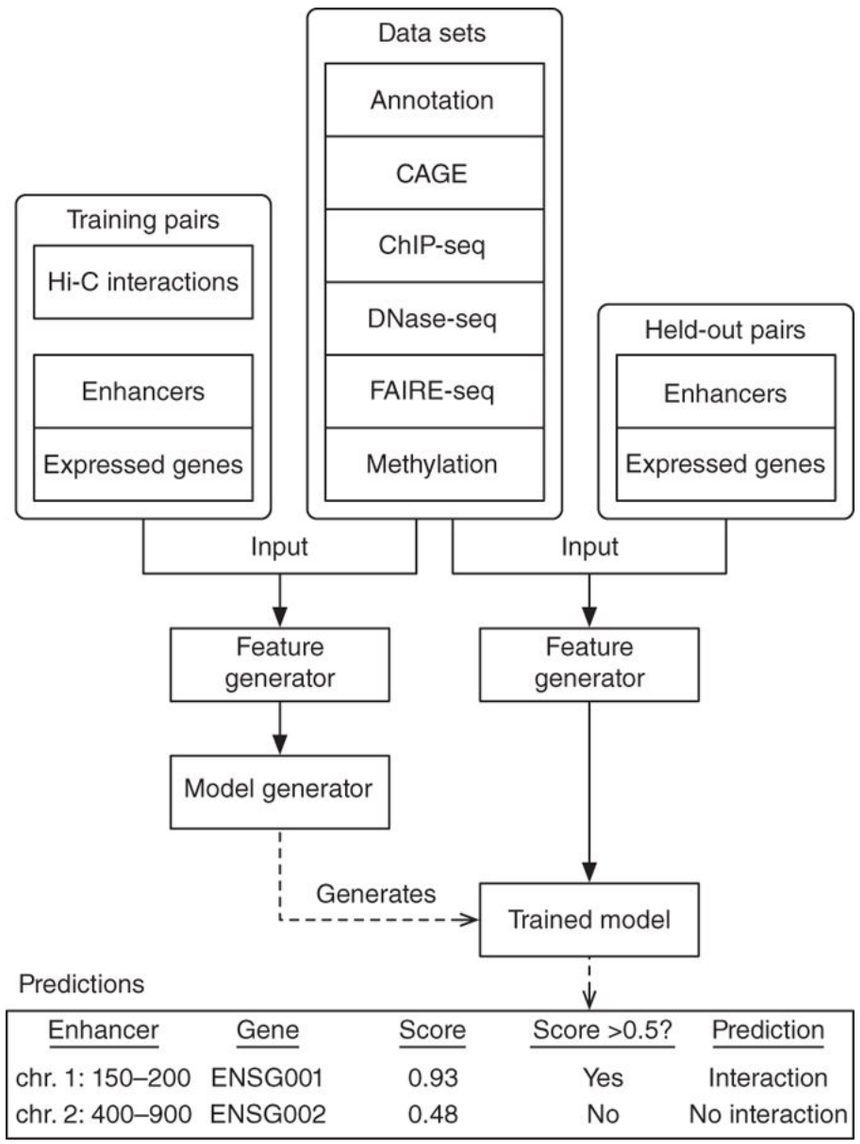
# TargetFinder uses multiple data types to predict HiC interactions



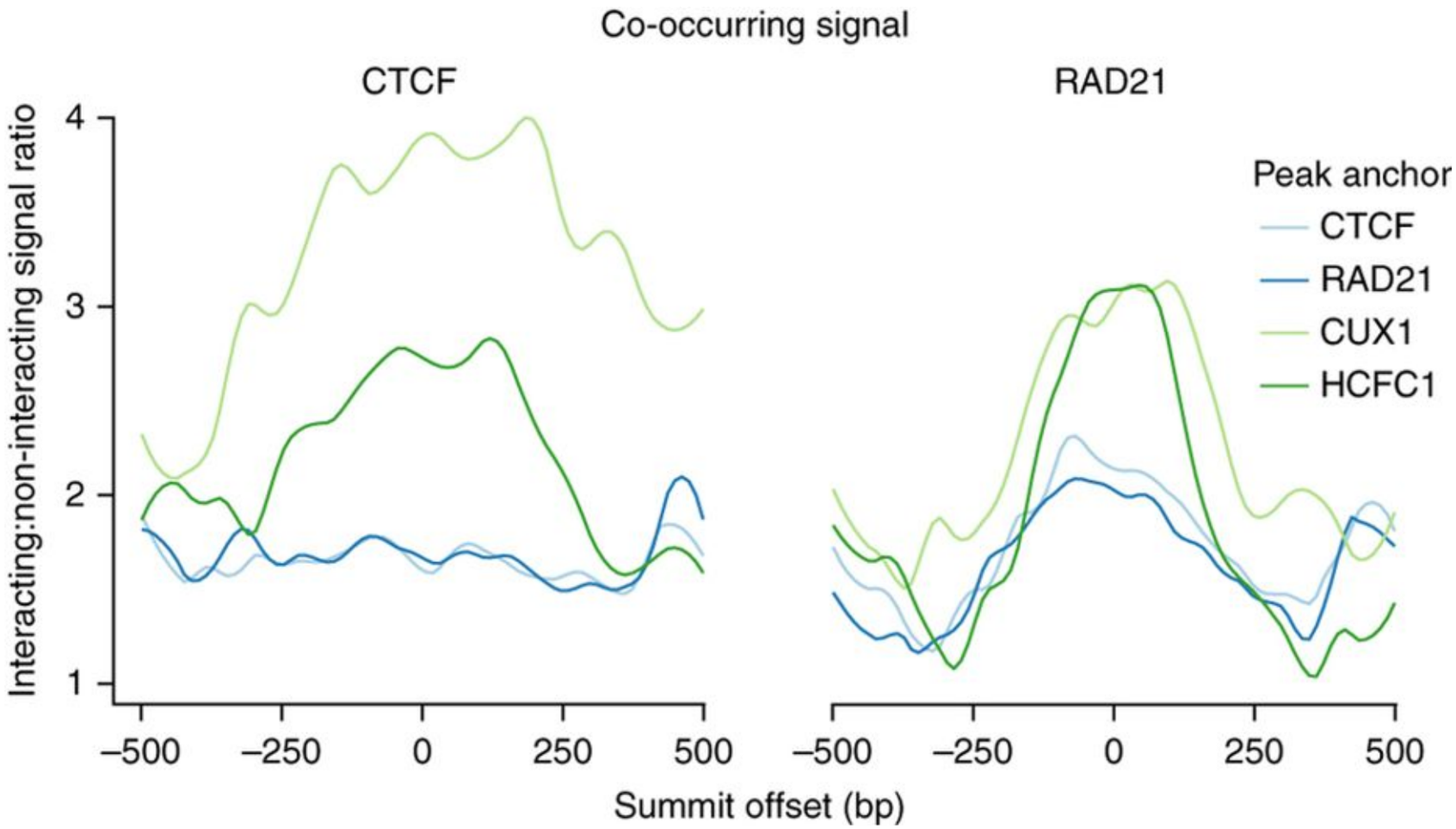

https://www.nature.com/articles/ng.3539

# TargetFinder Training Data

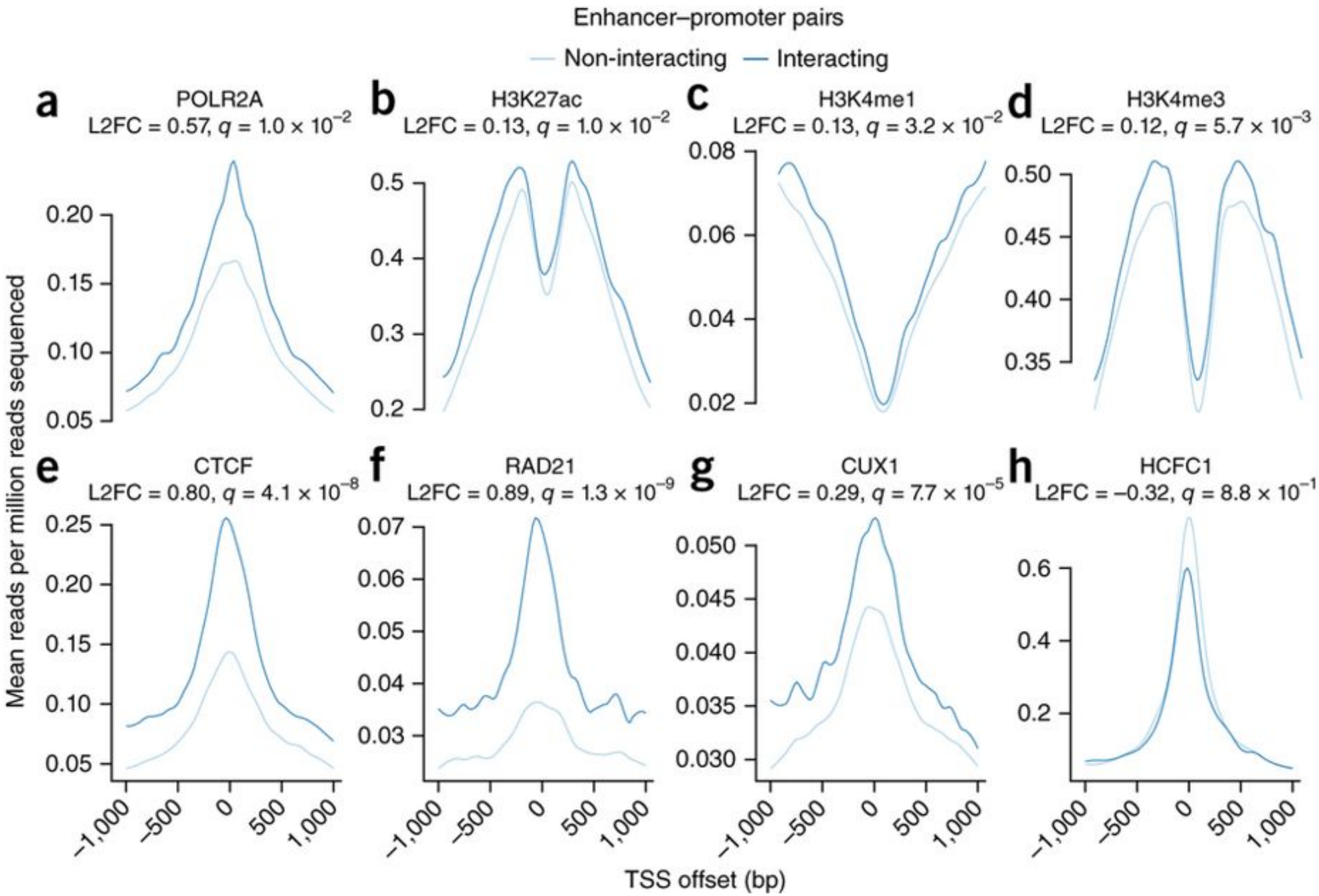
| Cell Line | Promoters | Enhancers | Interacting Pairs | Non-Interacting Pairs |
|---|---|---|---|---|
| K562 | 8196 | 82806 | 1977 | 39500 |
| GM12878 | 8453 | 100036 | 2113 | 42200 |
| HeLa-S3 | 7794 | 103460 | 1740 | 34800 |
| HUVEC | 8180 | 65358 | 1524 | 30400 |
| IMR90 | 5253 | 108996 | 1254 | 25000 |
| NHEK | 5254 | 144302 | 1291 | 25600 |



# TargetFinder – Ratio of the CTCF and RAD21 ChIP-seq signals occurring within interacting enhancers and non-interacting enhancers

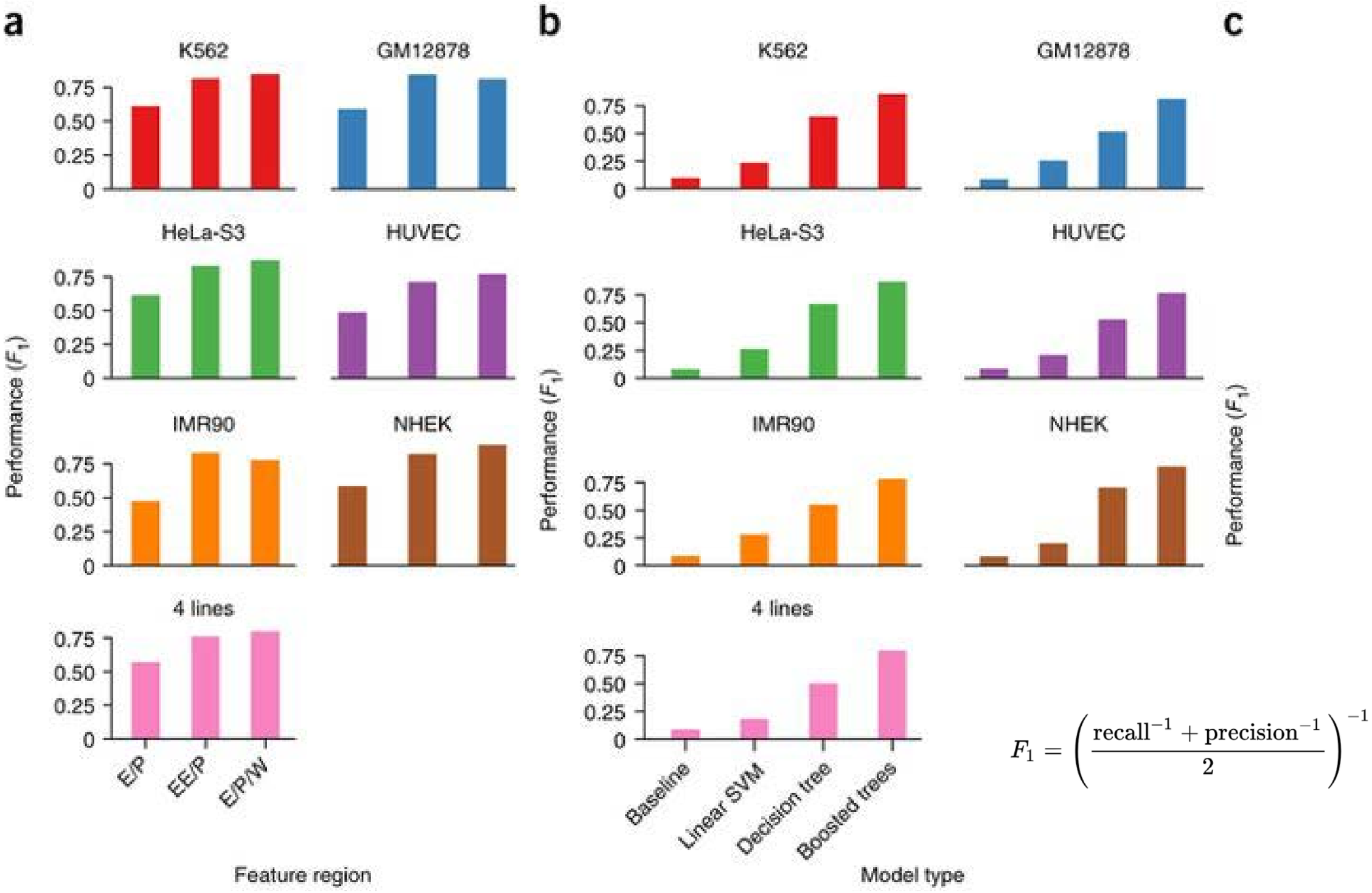


# TargetFinder – Enrichment of signals at transcription start sites (TSS)




Dark – interacting; Light – non-interacting

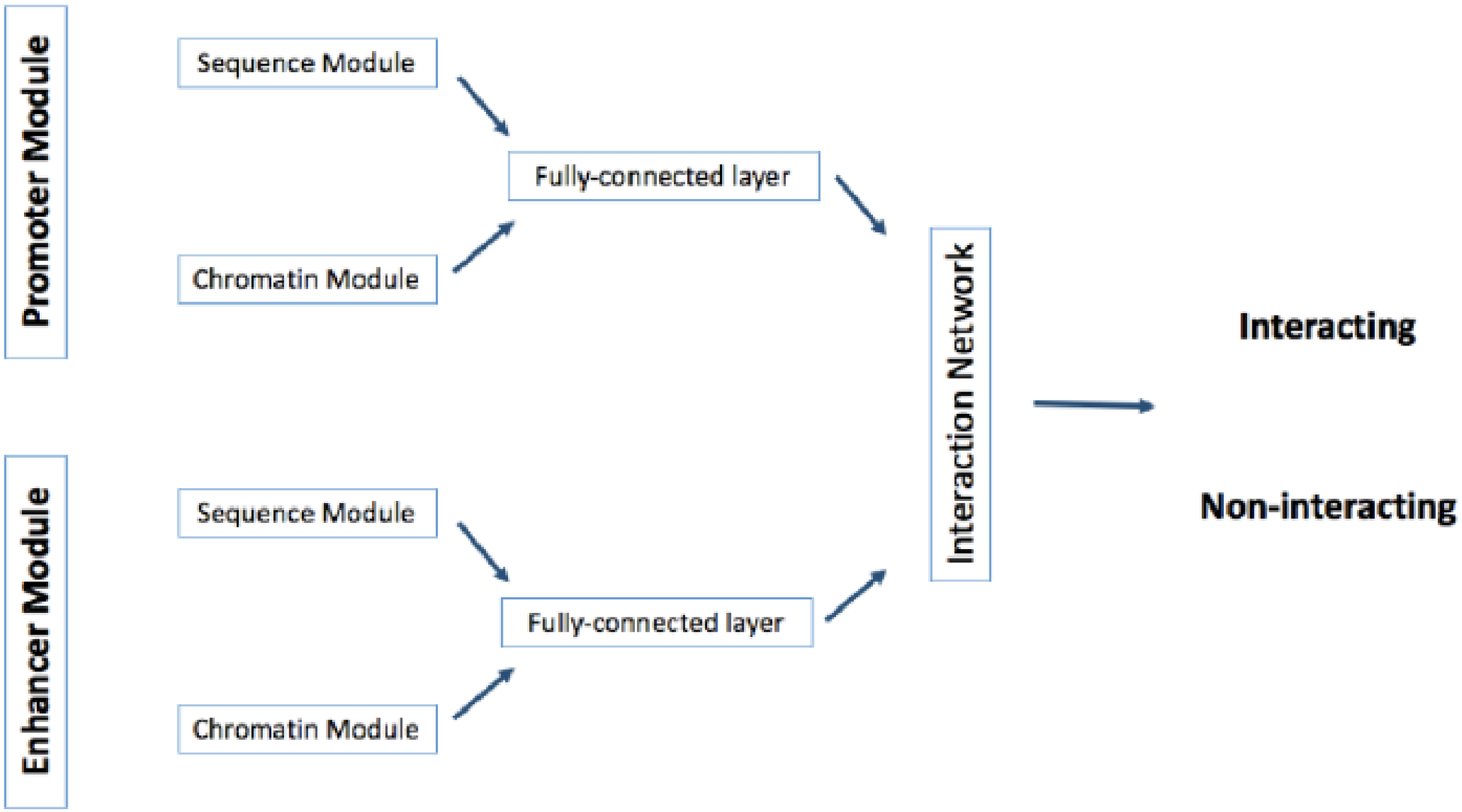
# TargetFinder – Performance




$$F_1 = \left( \frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1}$$

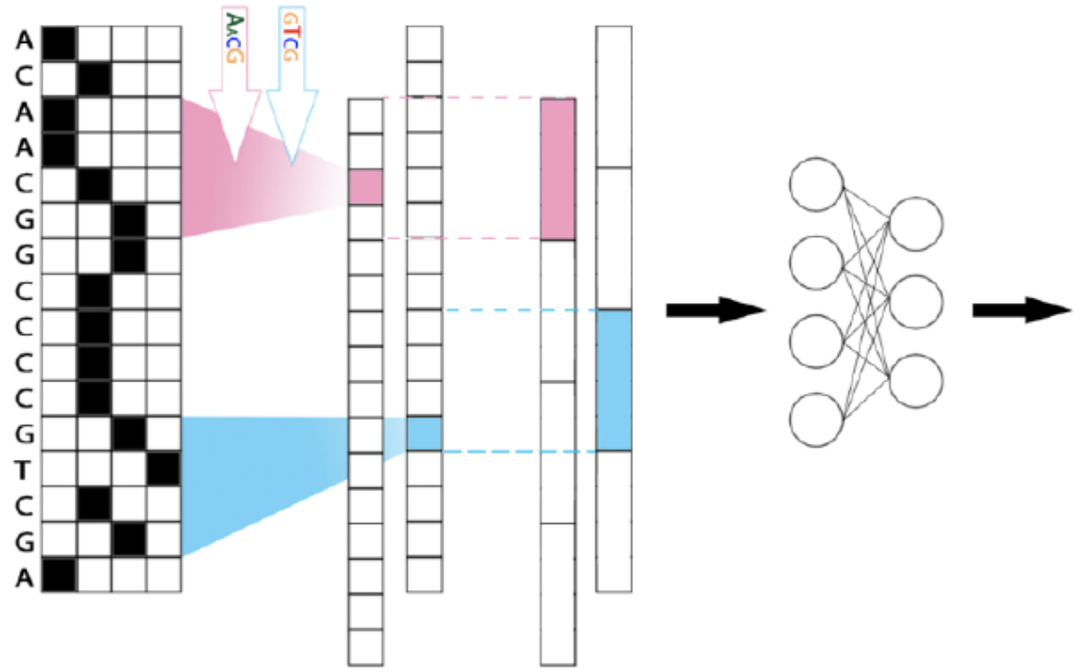
Features for enhancers and promoters only (E/P), extended enhancers and promoters (EE/P), and enhancers and promoters plus the windows bet

# Deep learning network for predicting enhancer-promoter interactions

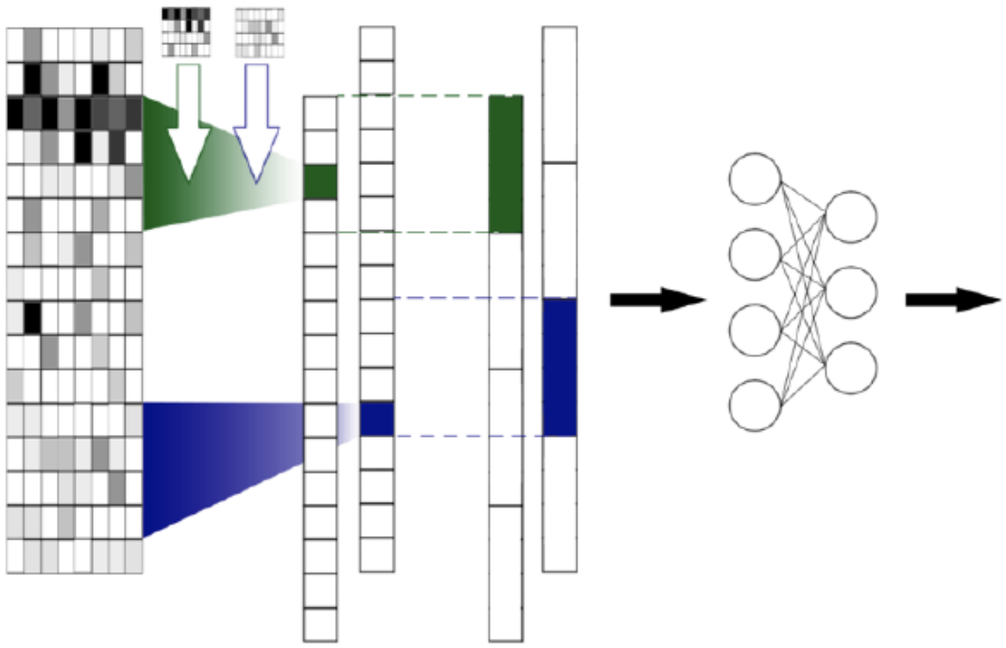


# Sequence and chromatin anchor networks outputs are concatenated

Sequence
-2kb
sequence
windows




Chromatin –
10 kb / 200 bp bins
DNase-seq,
H3K4me1, H3K4me2,
H3K27ac,
H3K27me3,
H3K36me3, and
H3K9me3



# Enhancer promoter prediction performance with varying feature sets