

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>

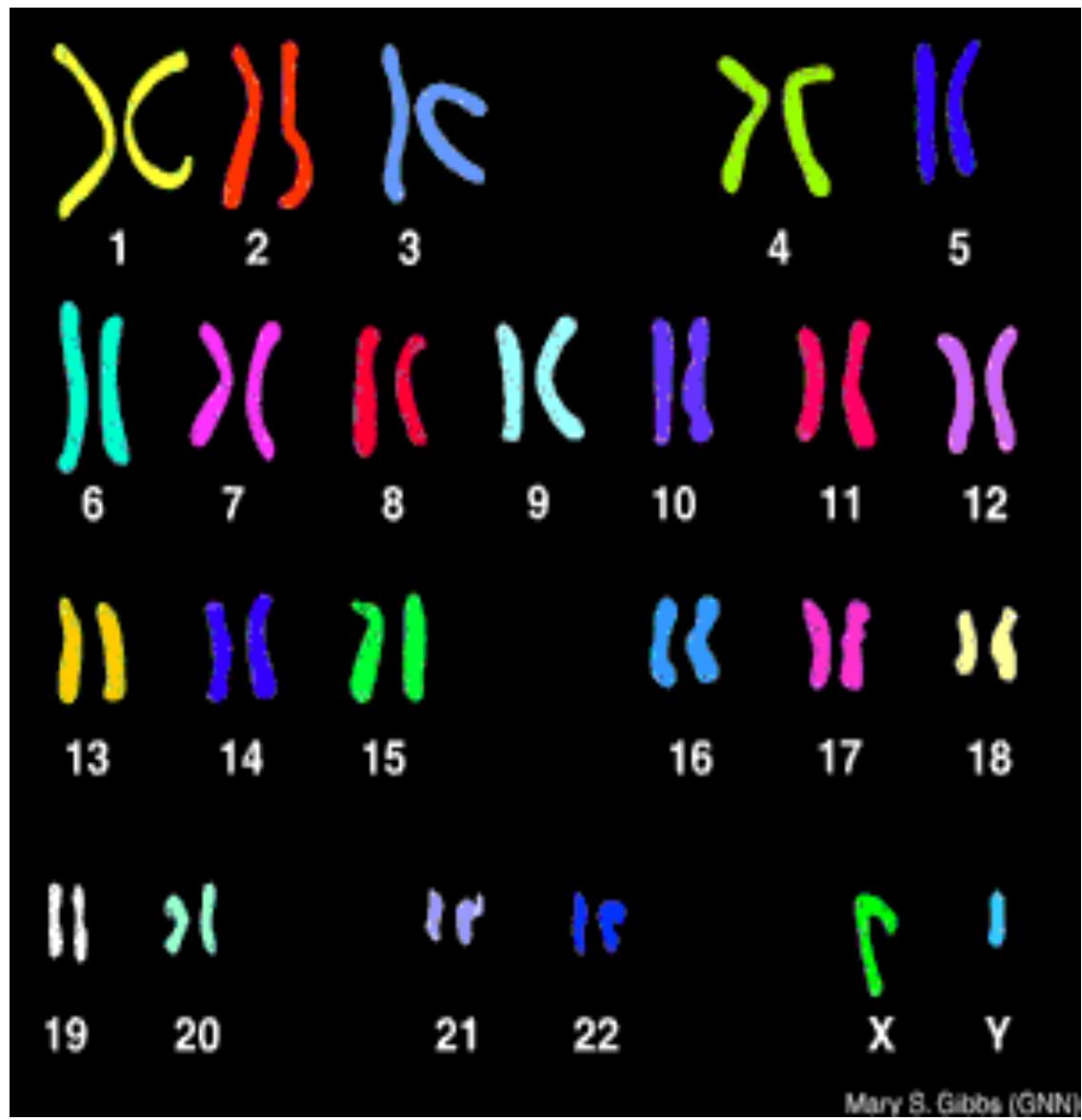
What's on tap today!

- Predicting hidden chromatin state
- Using chromatin state to predict causal variants
- Discovering enhancer-promoter interactions
- Predicting interactions
 - Anchor based methods
 - Clustering based methods

What you should know

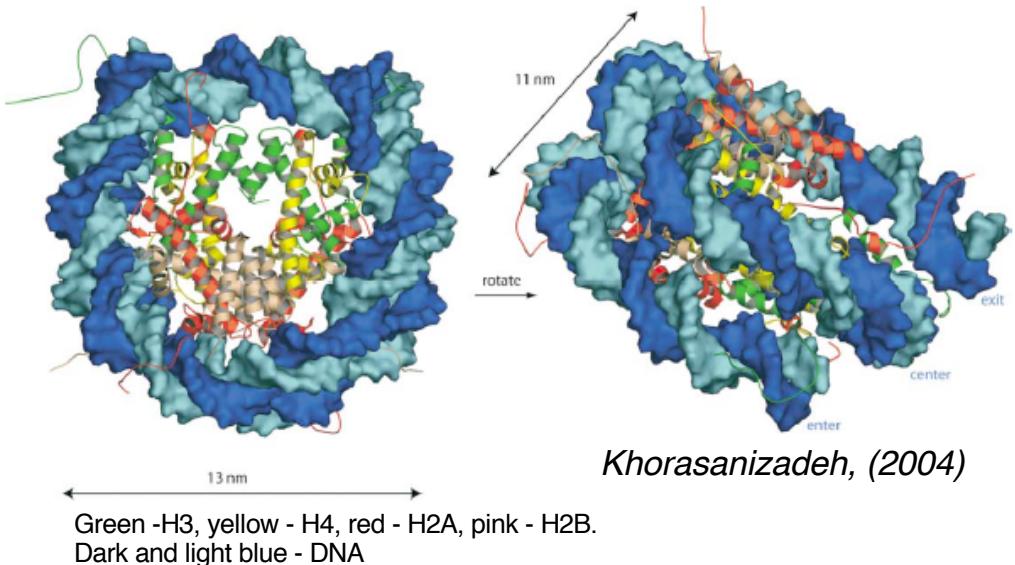
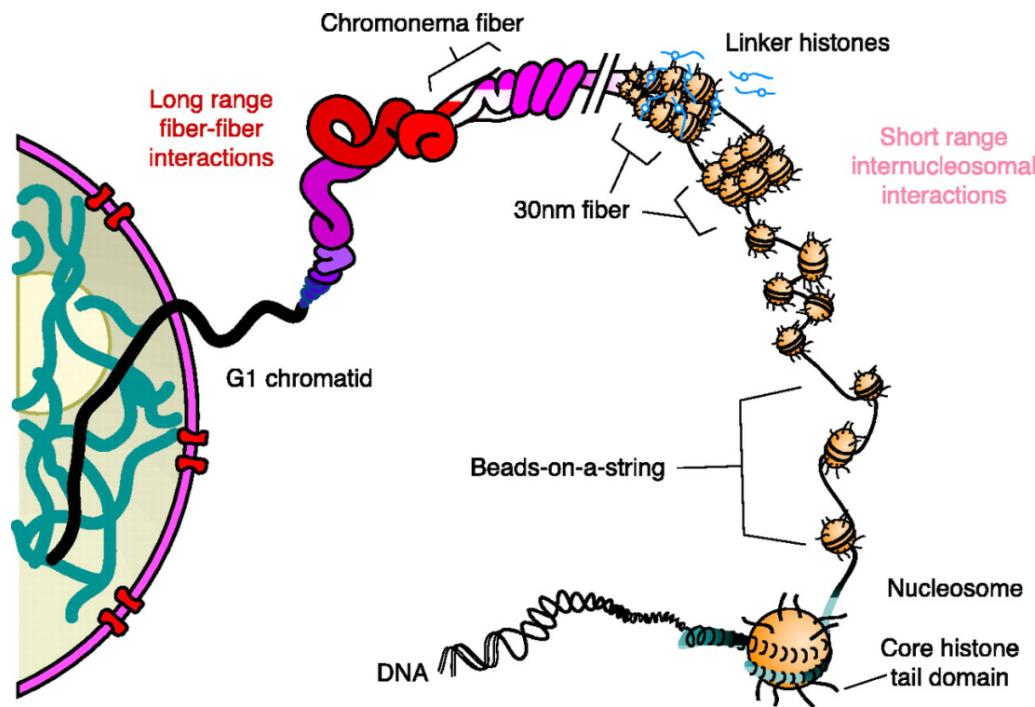
- Chromatin marks and their models
 - Hidden Markov Model (HMM)
 - Deep learning model (DeepSEA)
- Methods for characterizing genome interactions
 - Hi-C
 - ChIA-PET
 - HiChip
- Characterizing genomic interactions
 - Anchor based methods
 - Clustering based methods (CID)

Chromatin marks are important
biological state and can be predicted



Mary S. Gibbs (GNN)

Chromatin and Nucleosome Organization



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

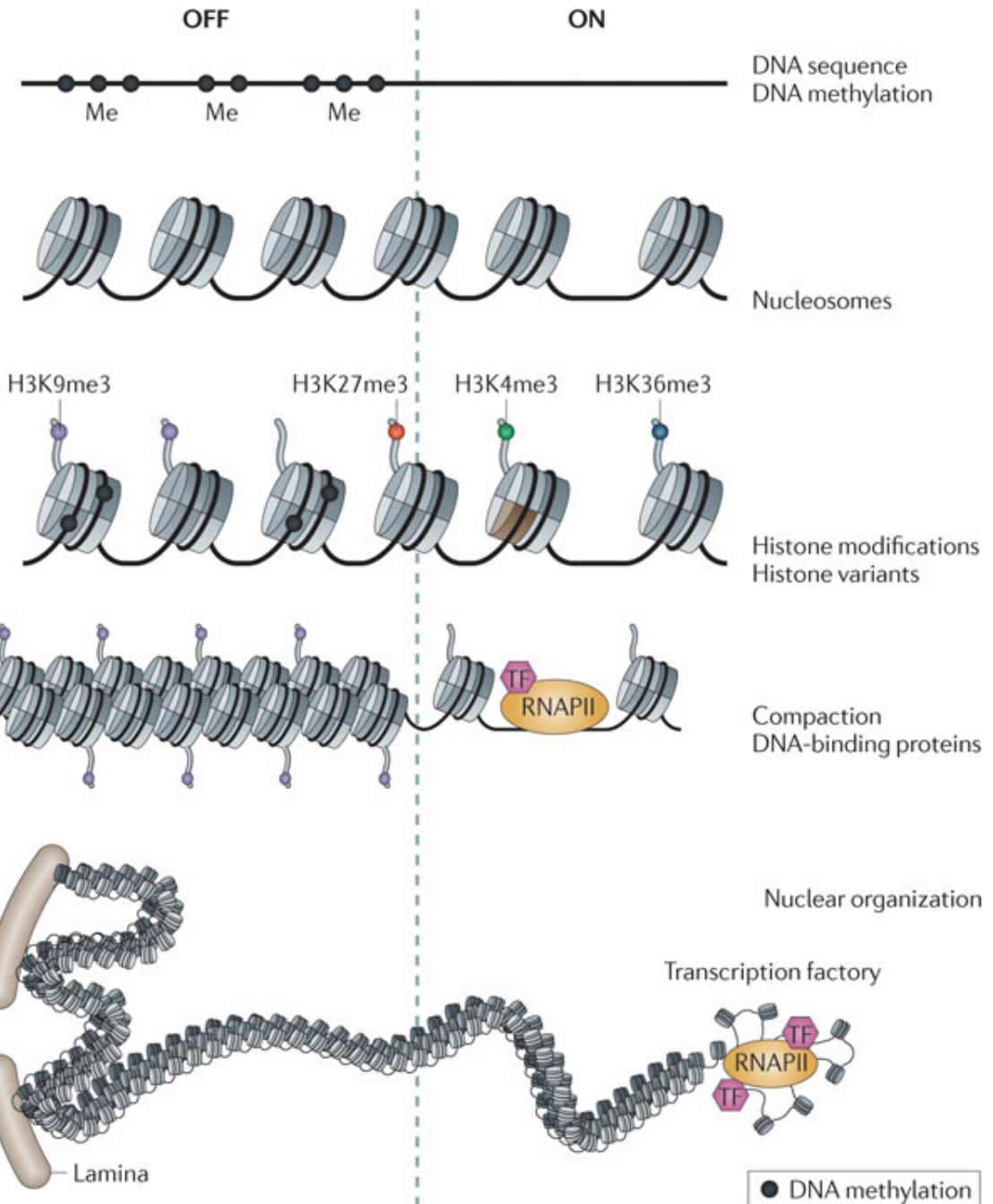
Histone variants

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

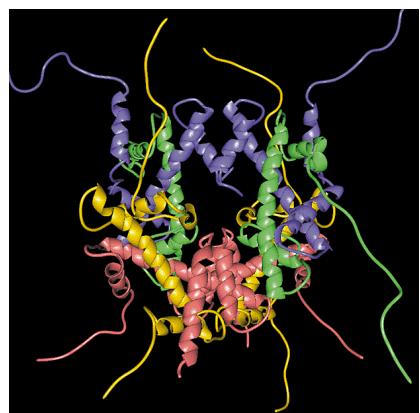
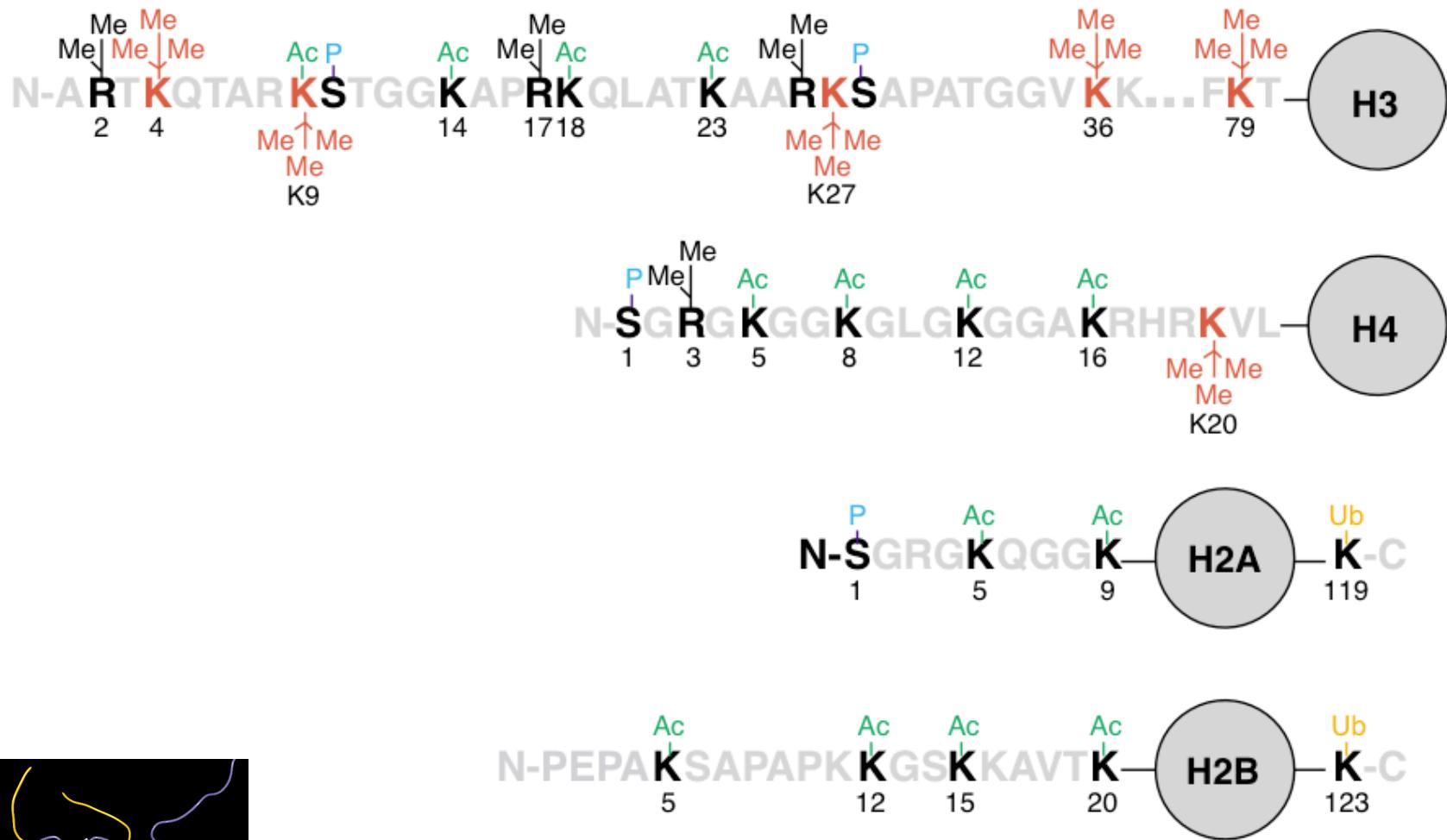
H2A variants: H2A.X - DNA damage
macroH2A - X chromosome
H2A.Z - transcribed regions

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

Both DNA methylation and chromatin marks contain important functional information

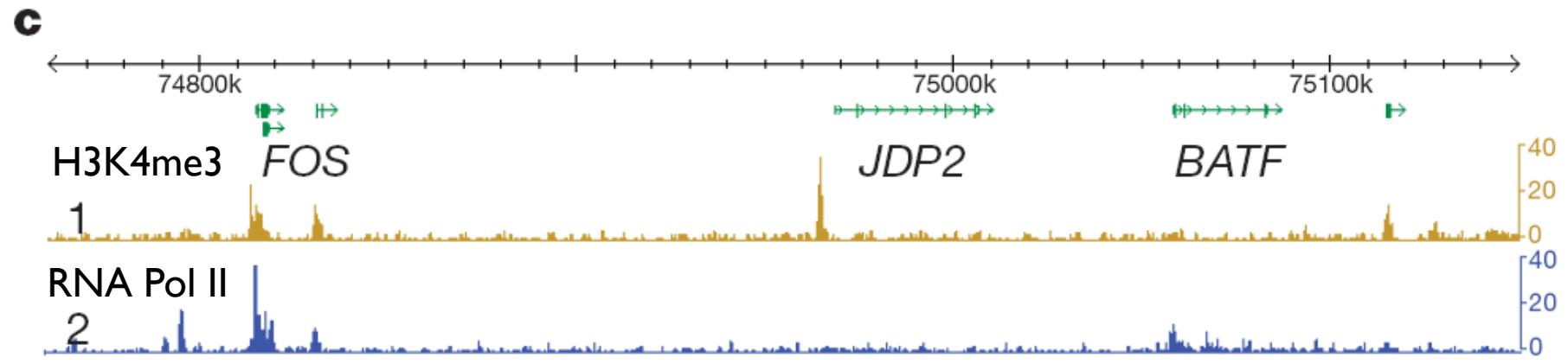


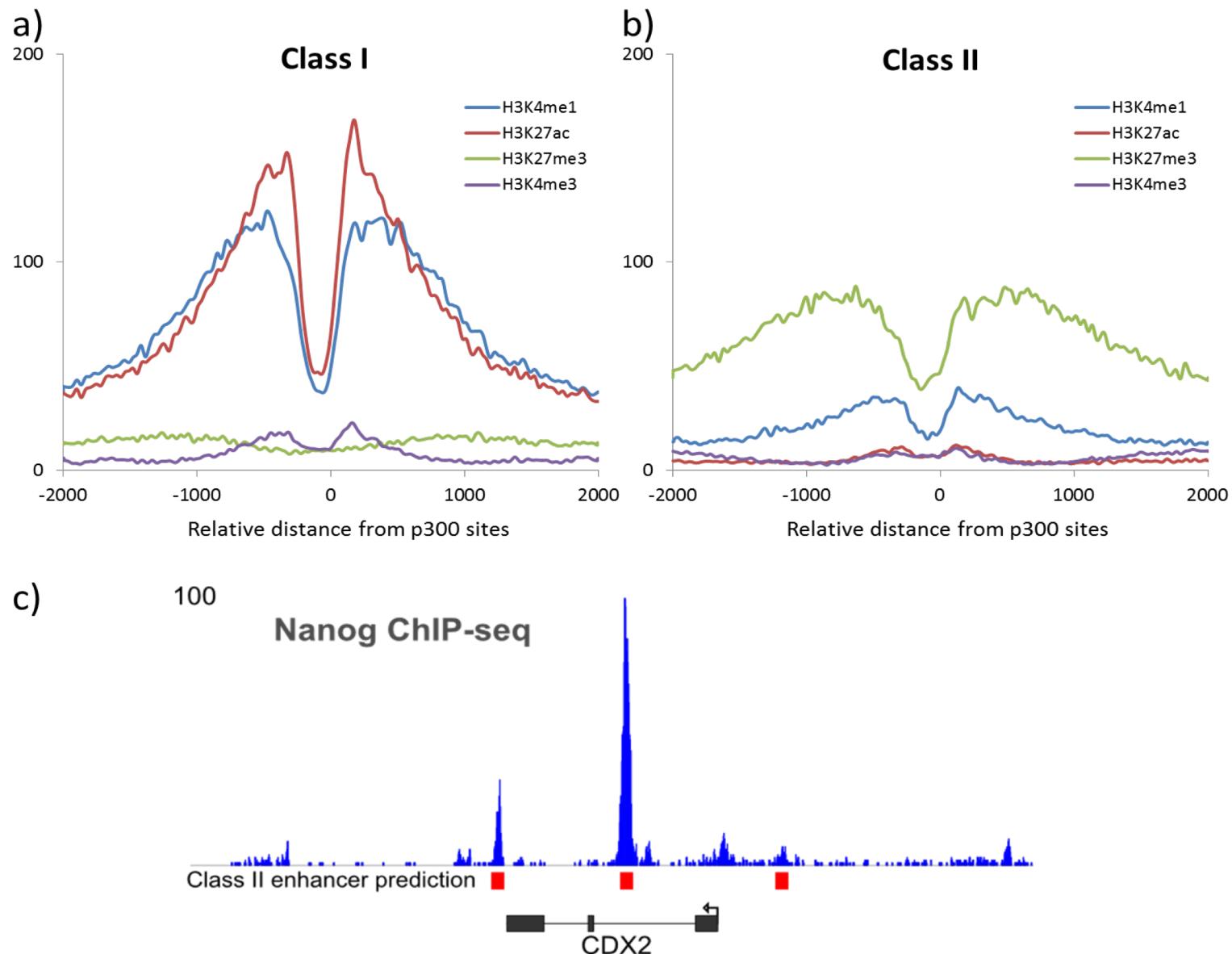
Histone Tail Modifications



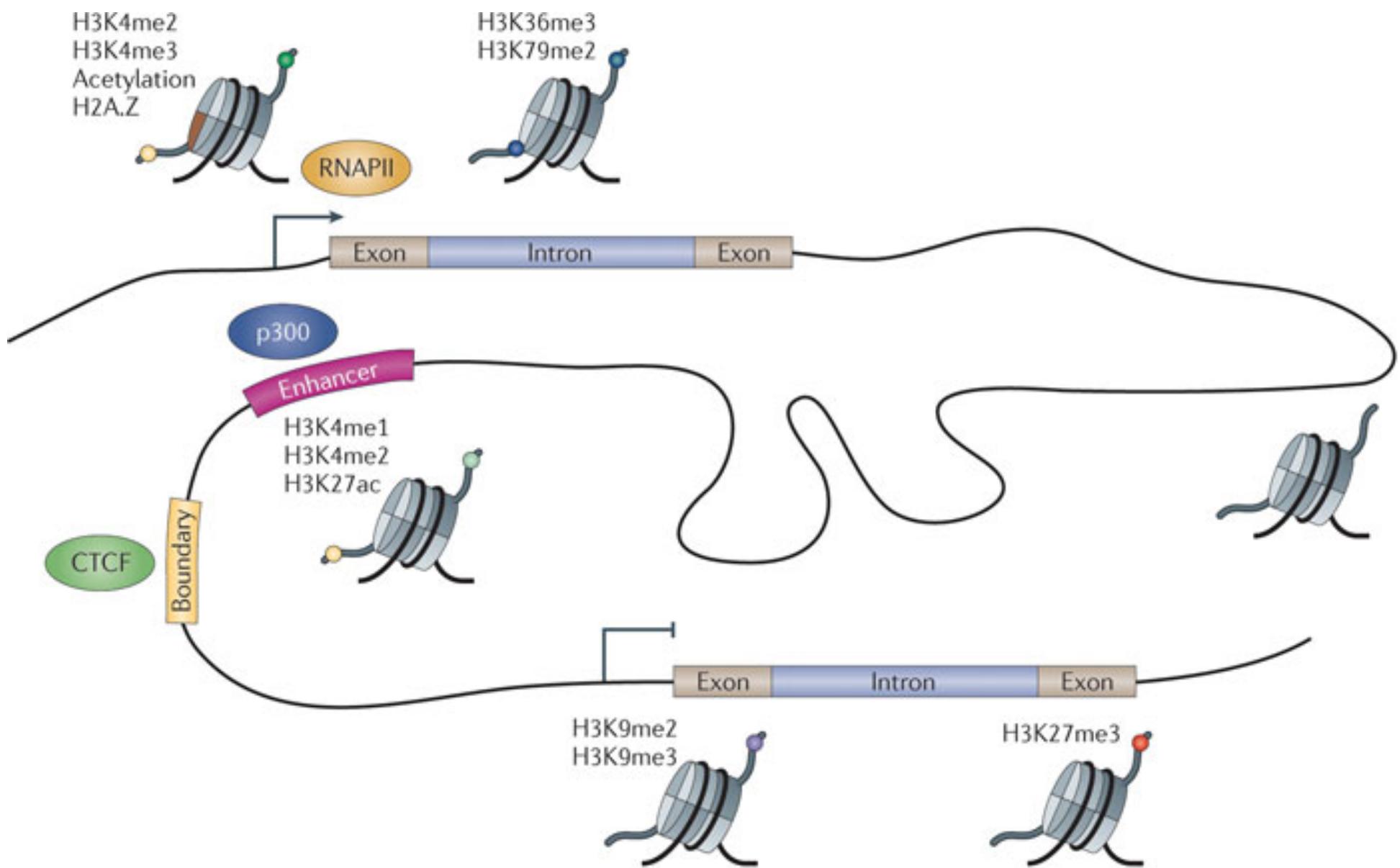
Sims III et al., 2003

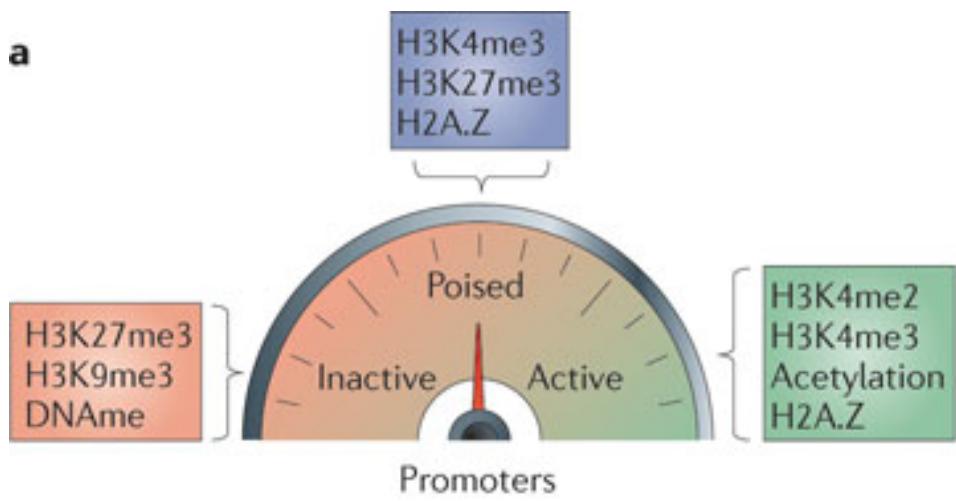
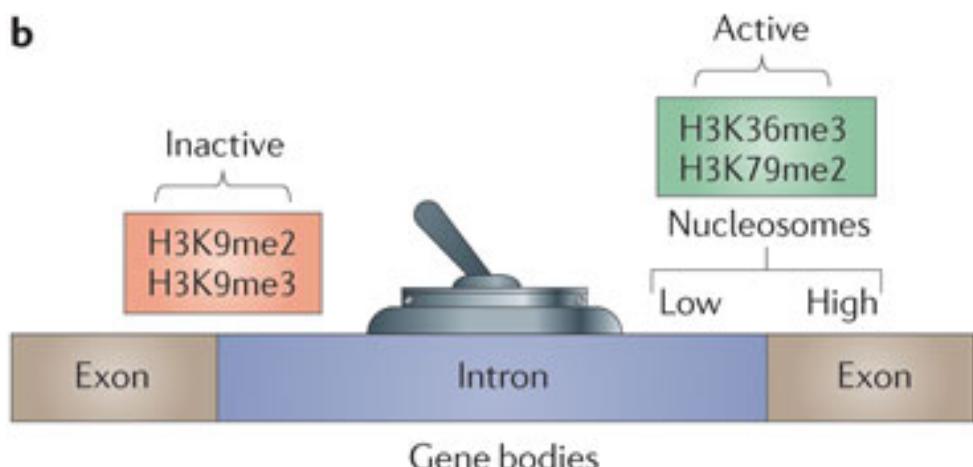
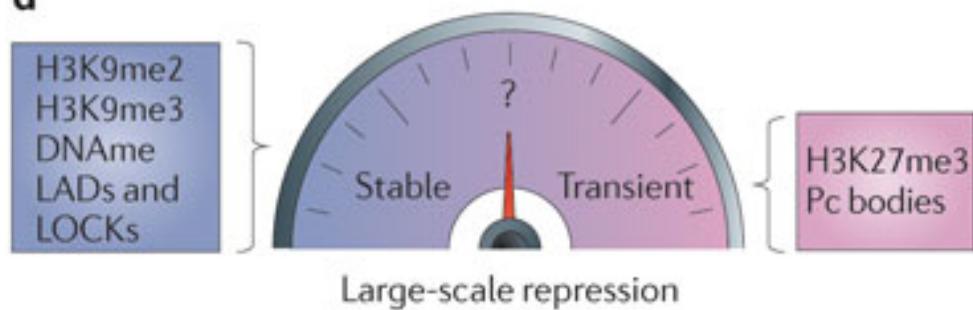
We can observe chromatin marks and other genome associated proteins using ChIP-seq



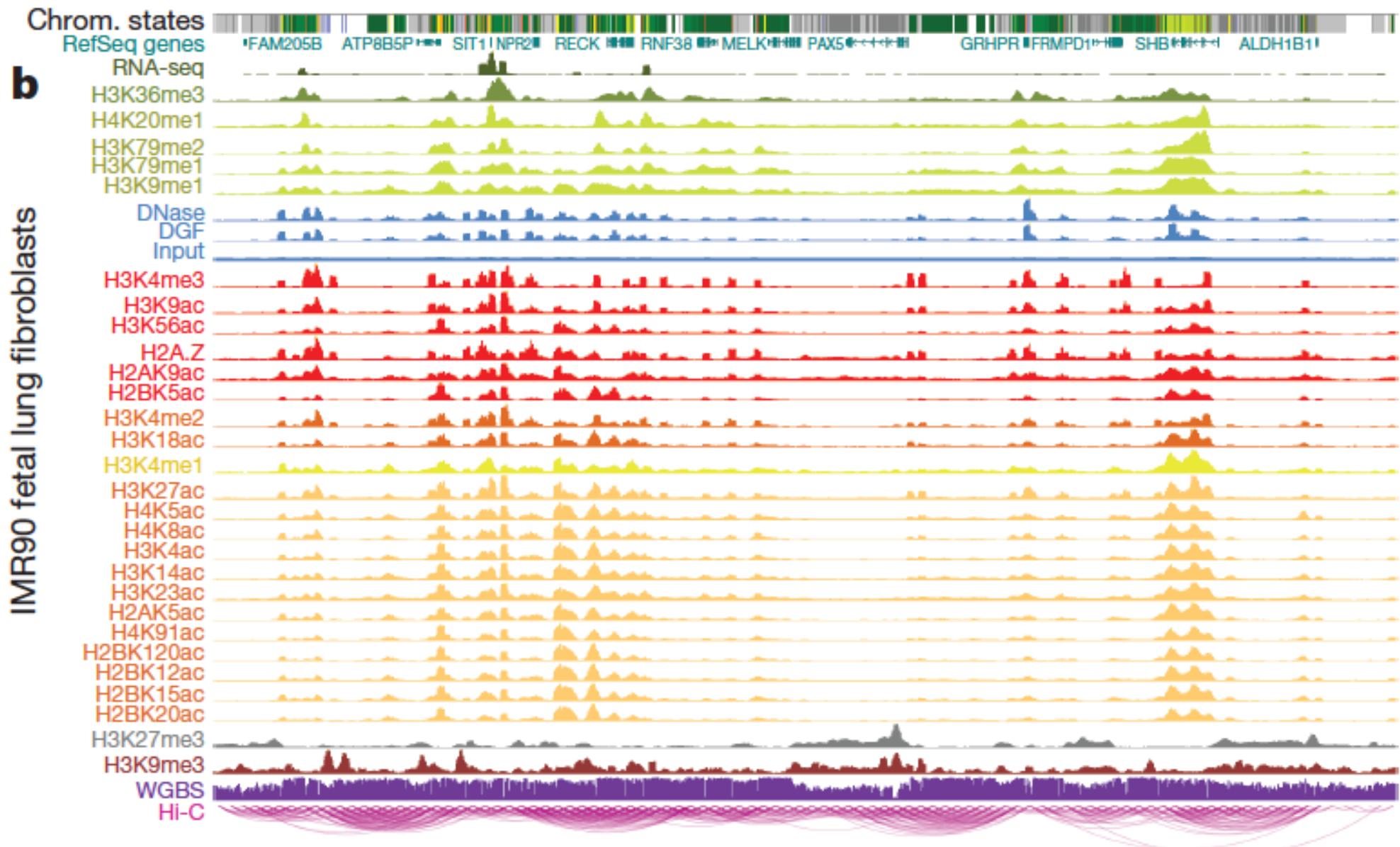


Detection of Class I (active) and Class II (poised) enhancers. a) b) hESC ChIP-seq read density profiles were generated for the indicated histone modifications centered on p300-bound regions in the top 1000 Class I and Class II enhancers, respectively. c) hESC Nanog ChIP-seq shows that Nanog binds at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



a**b****c****d**

Can we find latent state to explain observed marks?



Hidden Markov Models

Hidden state x in $[1 .. m]$

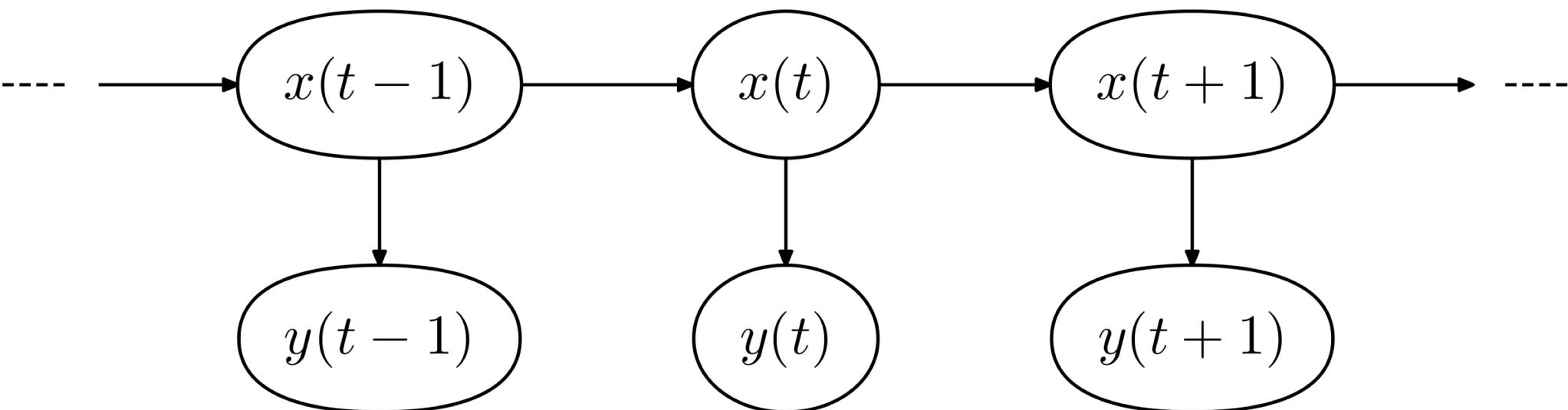
For example, m can 15

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 $P(x_{t+1} | x_t)$, $P(y_t | x_t)$



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

Hidden Markov Model (ChromHMM)

Divide genome into 200bp windows

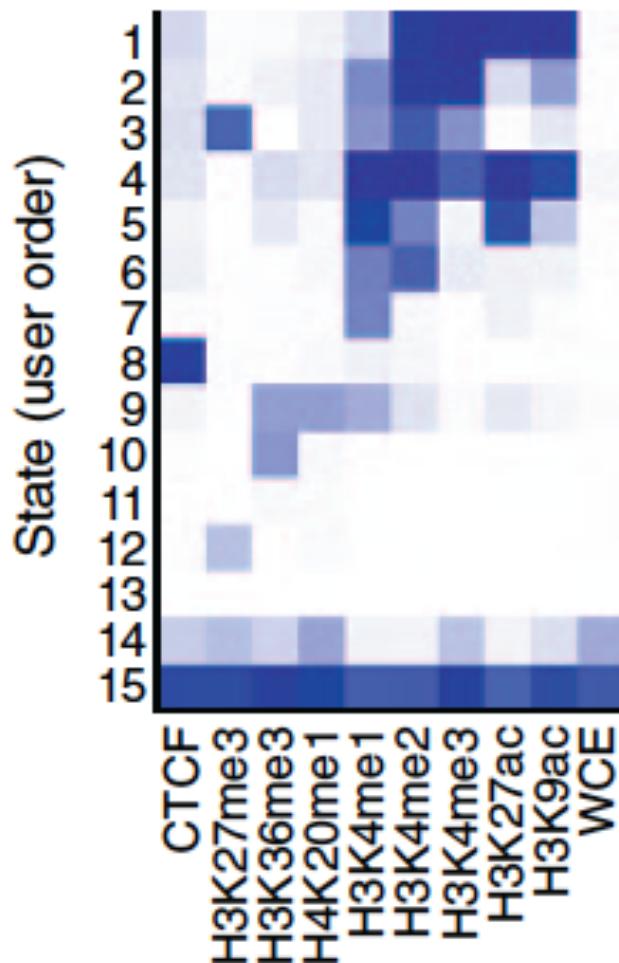
Hidden state for a 200bp window models what histone marks
are present in the window

Unsupervised – resulting states must be interpreted with
independent data

The number of states is fixed and is a modeling decision

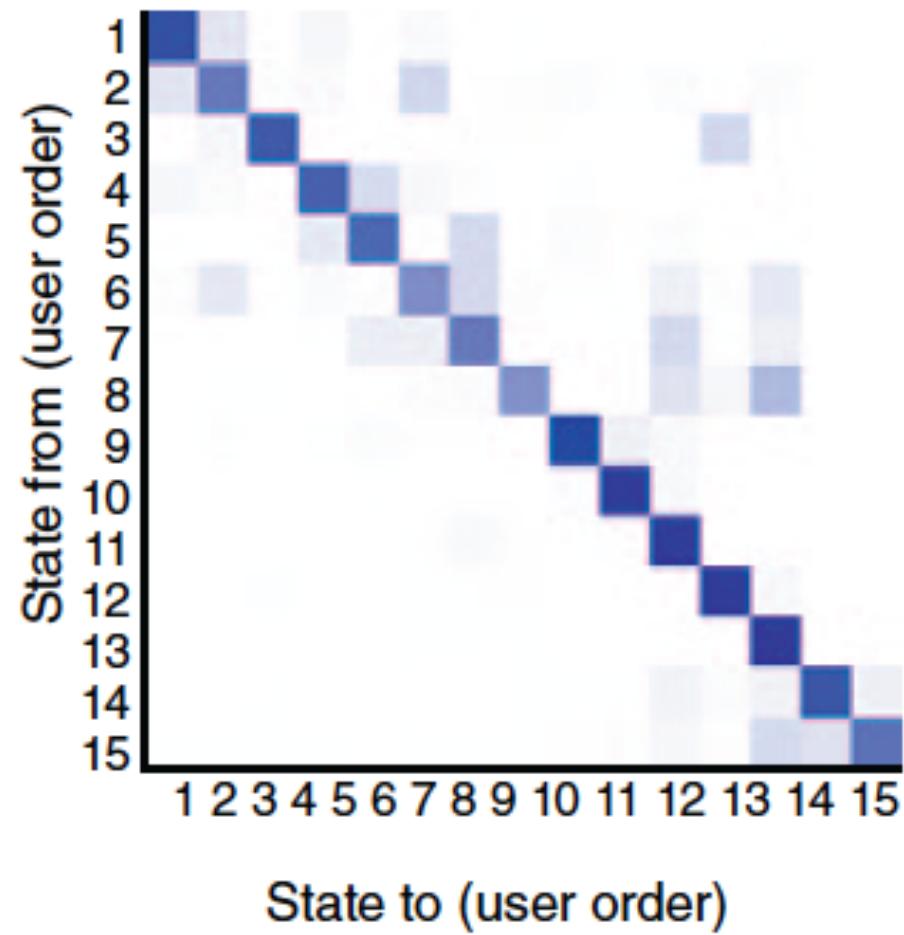
ChromHMM Model Parameter Visualization.

Emission parameters



$$P(y_t | x_t) \quad \text{Mark}$$

Transition parameters

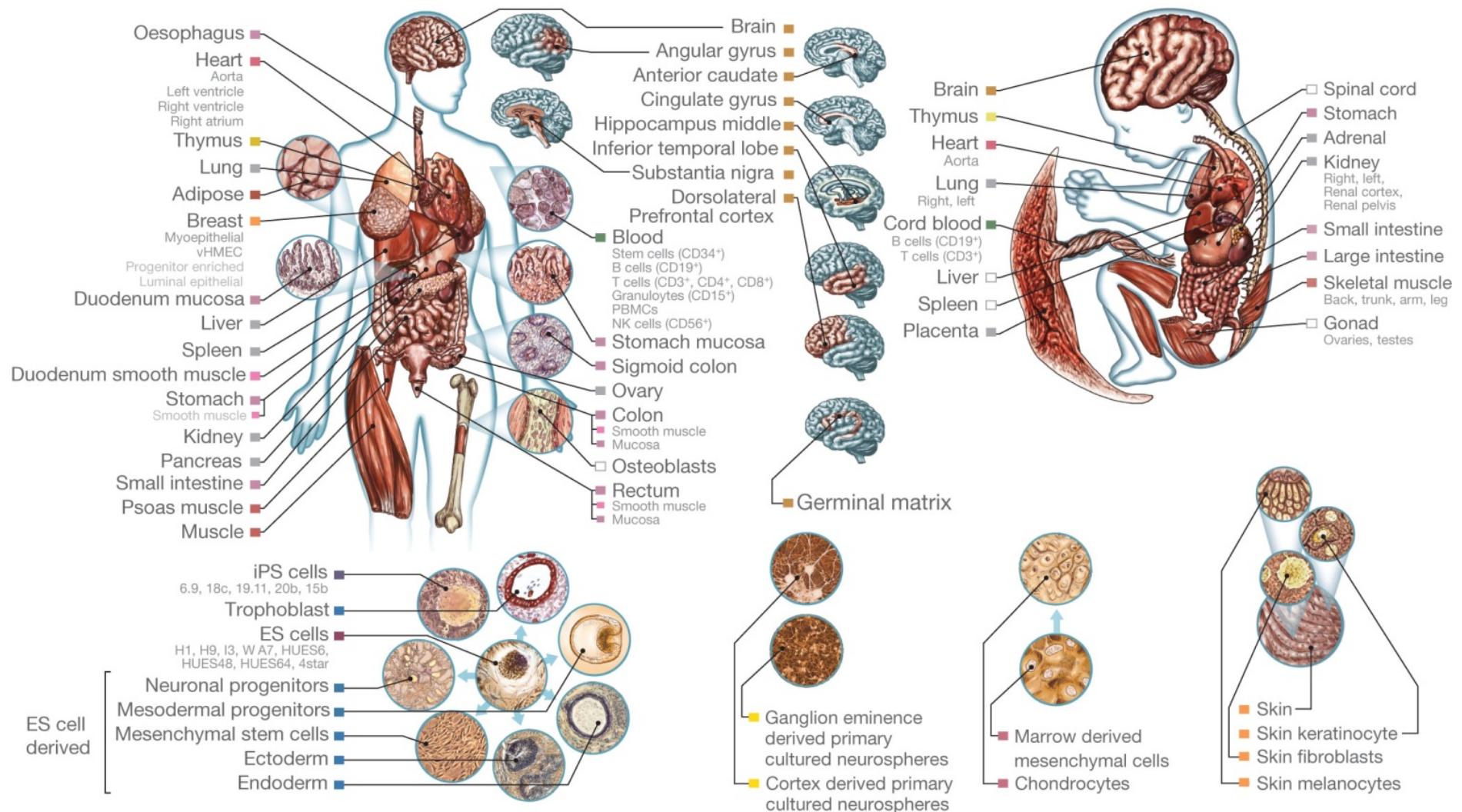


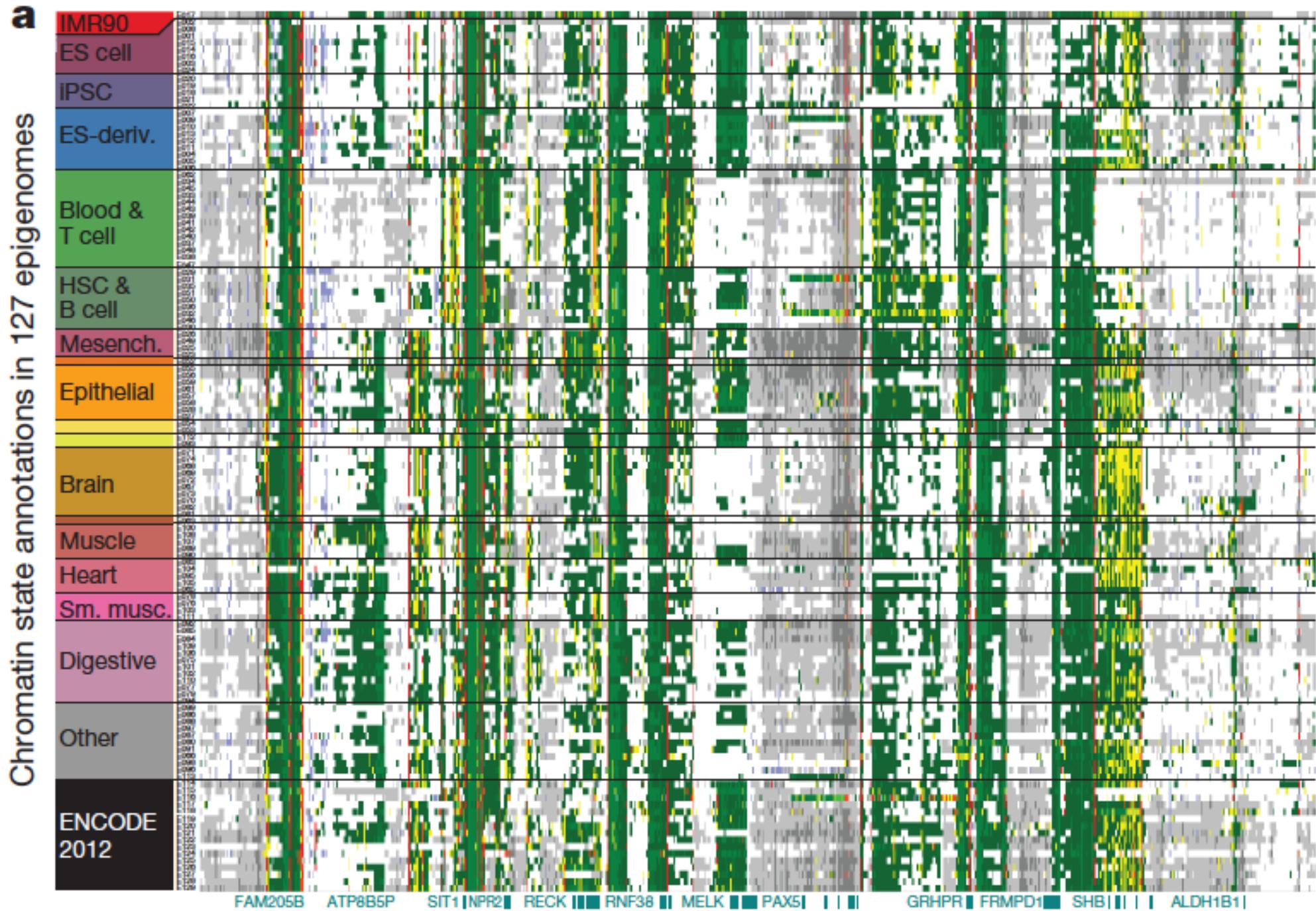
$$P(x_{t+1} | x_t)$$

ChromHMM segment based chromatin states



Tissues and cell types profiled in the Roadmap Epigenomics Consortium.

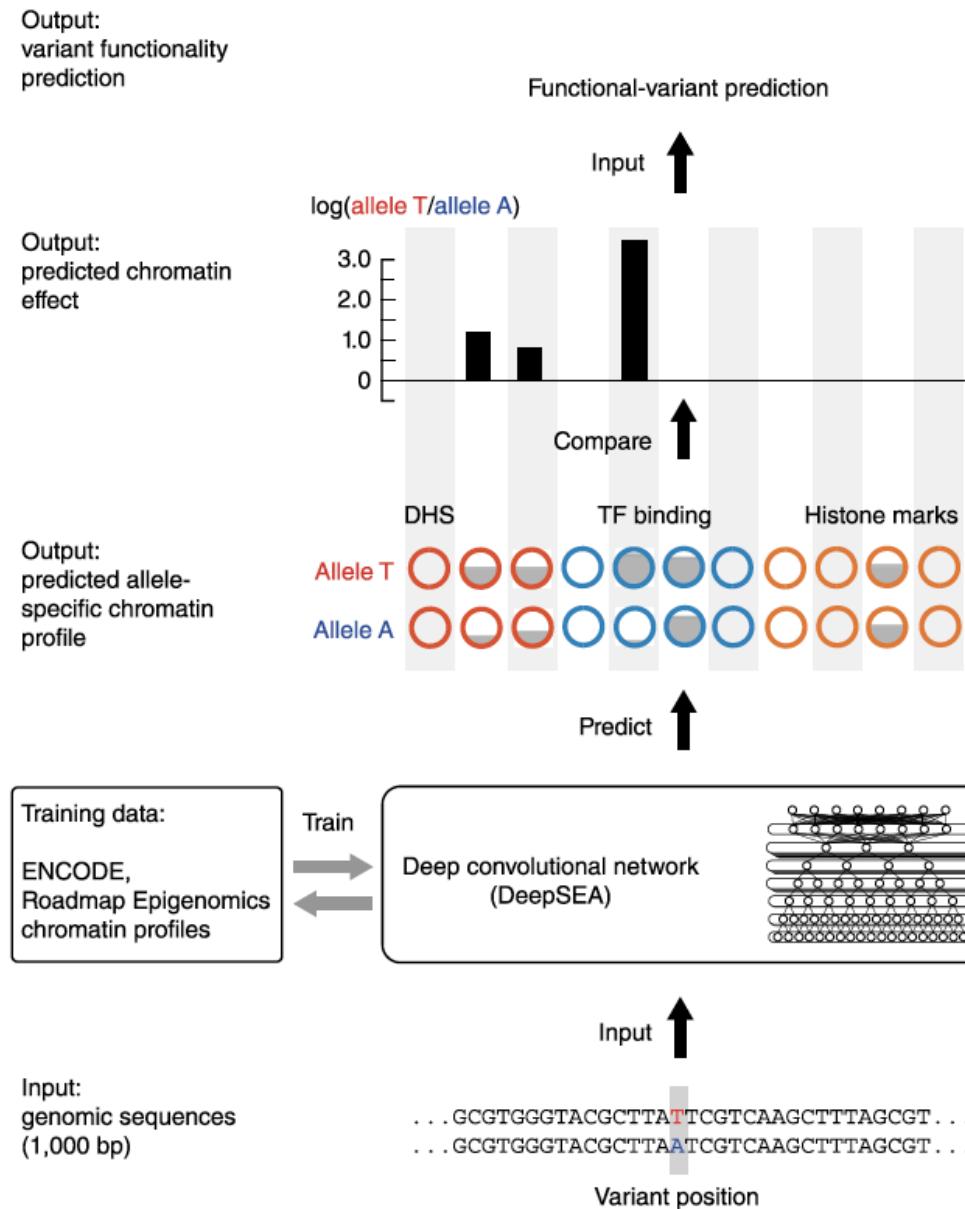




Can we predict chromatin state from
sequence?

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

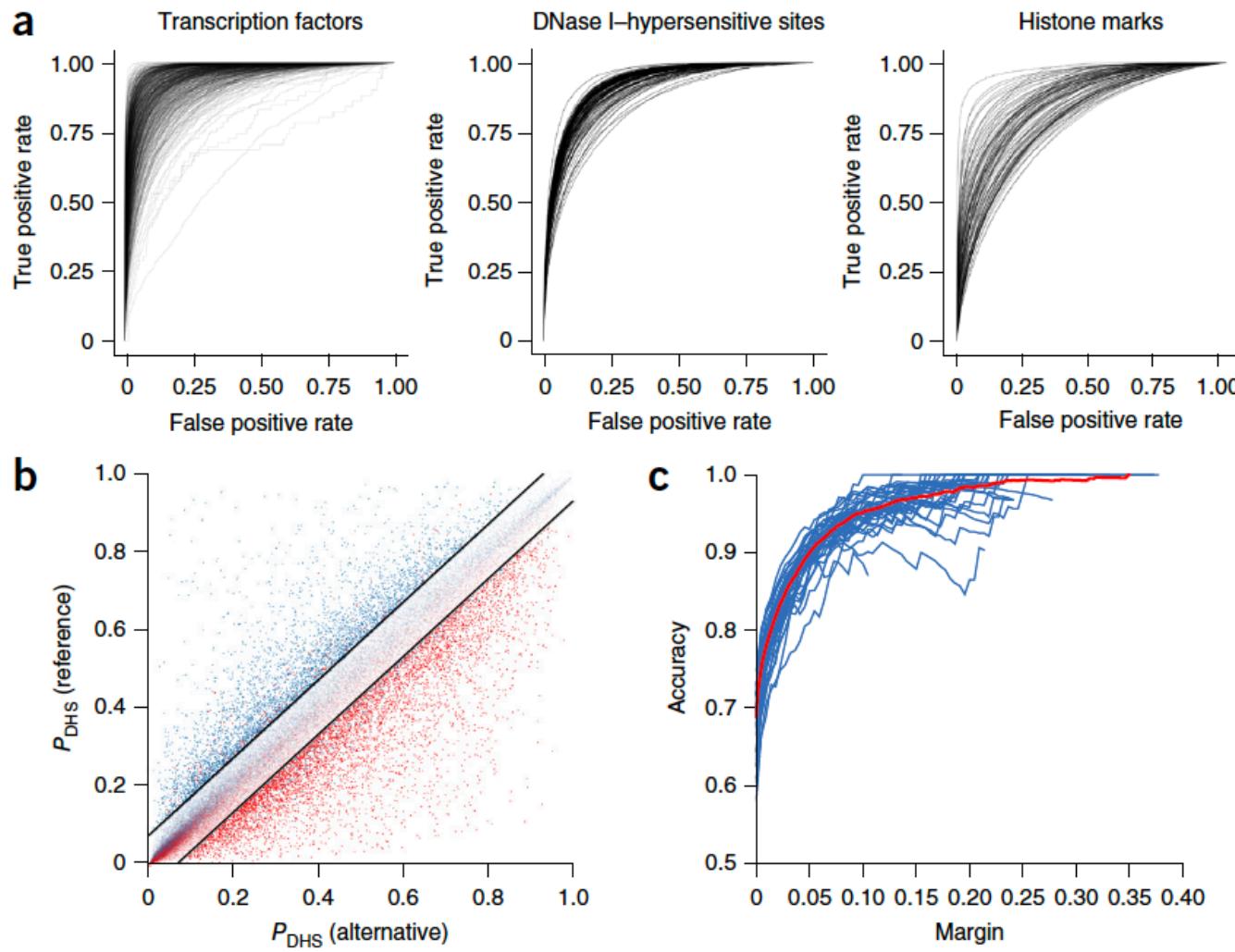


125 DNase features,
690 TF features, 104 histone features

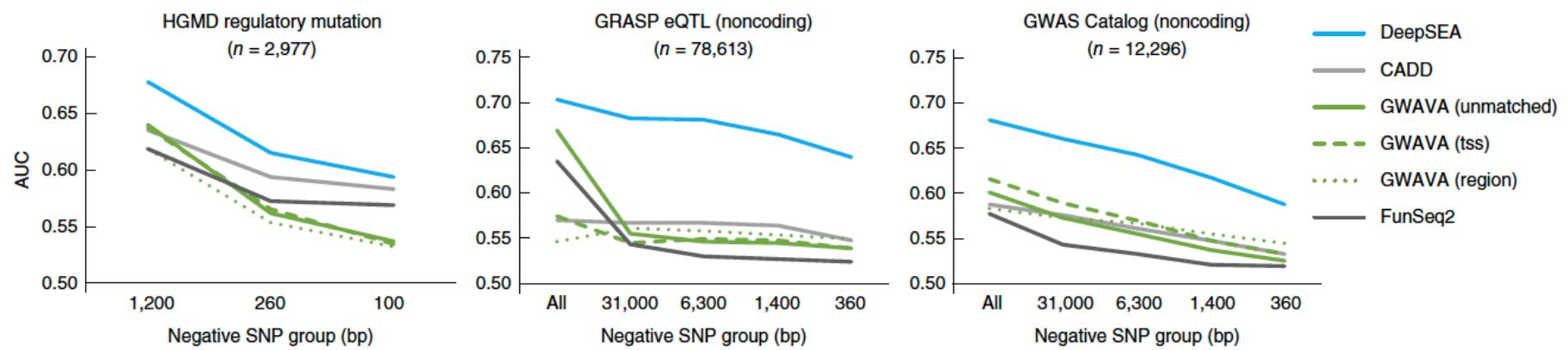
three convolution layers with 320, 480 and 960 kernels

1000 bp window

DeepSea can predict differentially accessible regions based upon SNP value

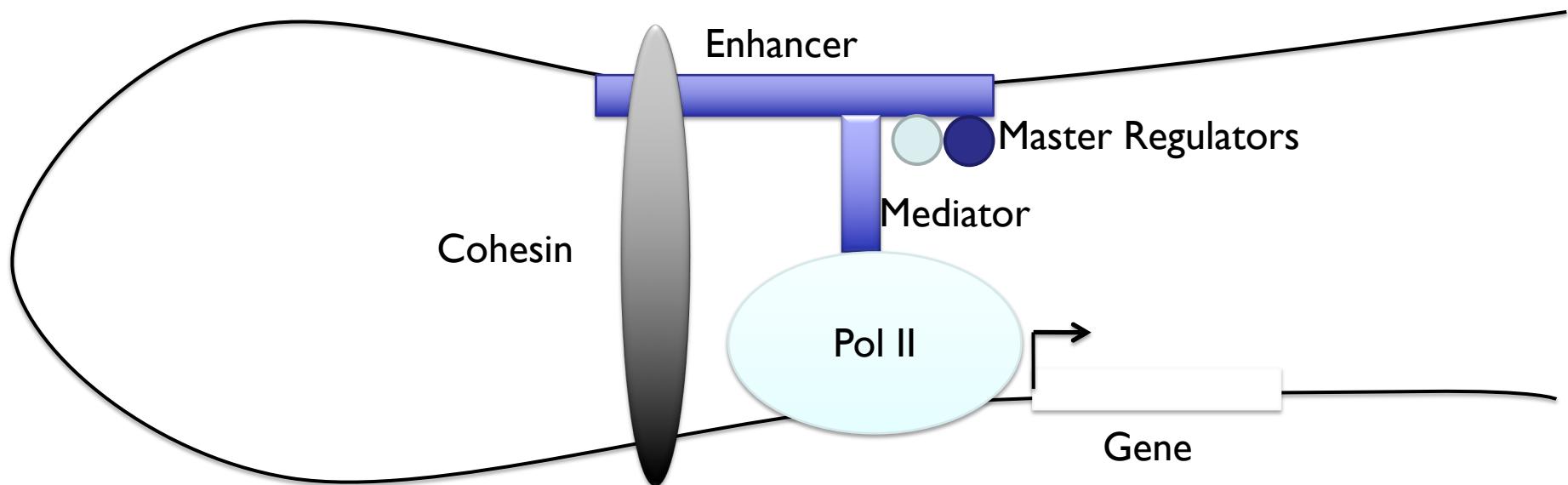


An ensemble logistic regression classifier based on DeepSea output can identify regulatory variants

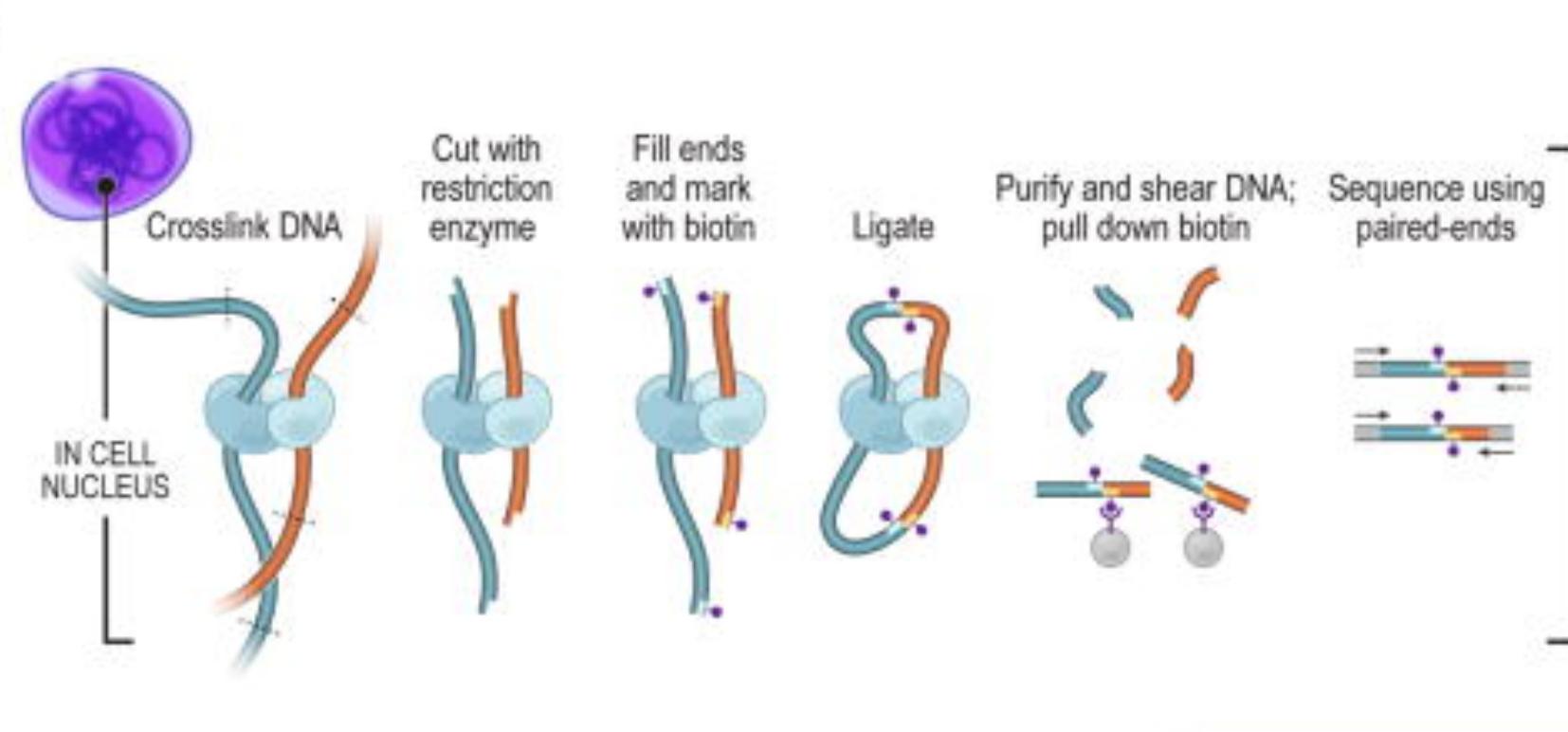


**HiC, HiChip, and ChIA-PET data
reveal distal genome interactions**

Enhancers regulate distal target genes by genome looping

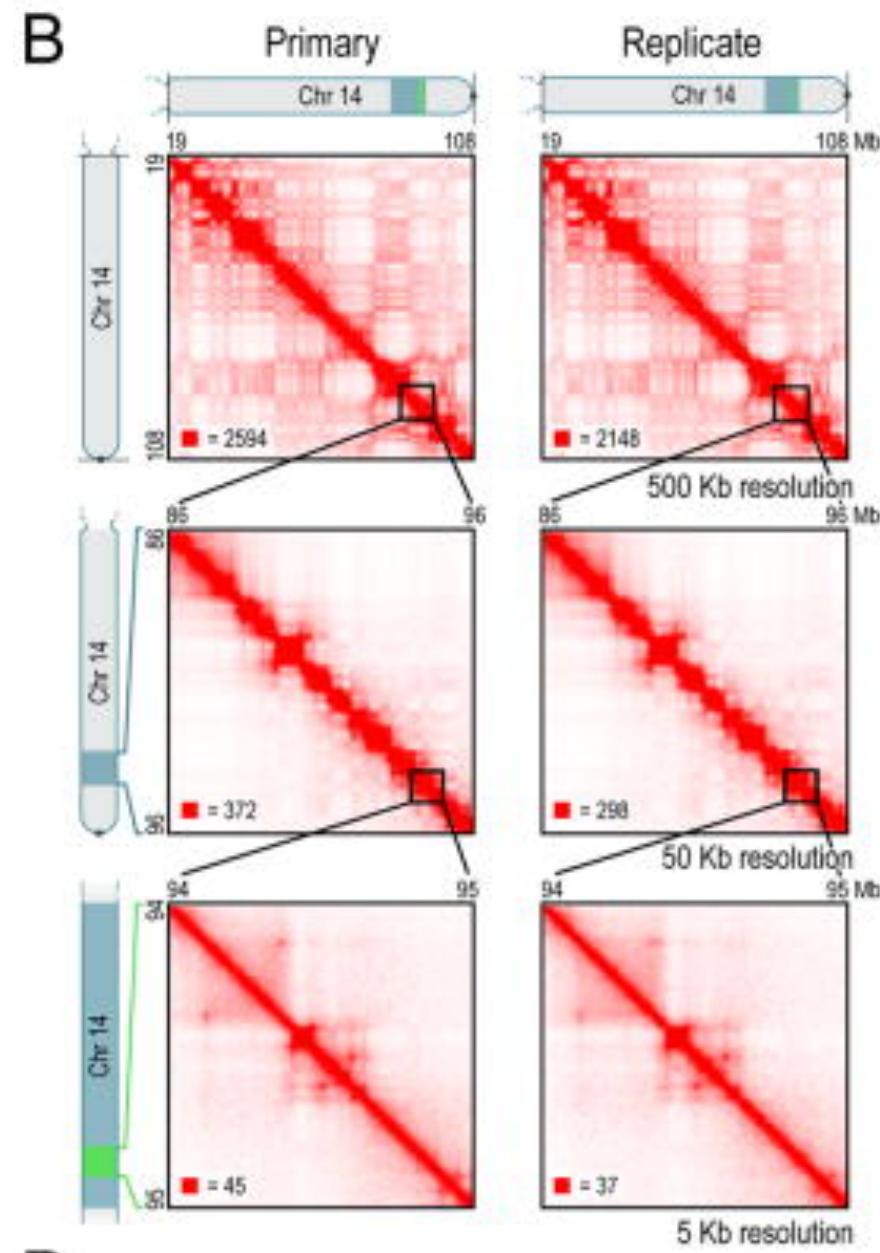


in situ HiC identifies proximal genomic contacts

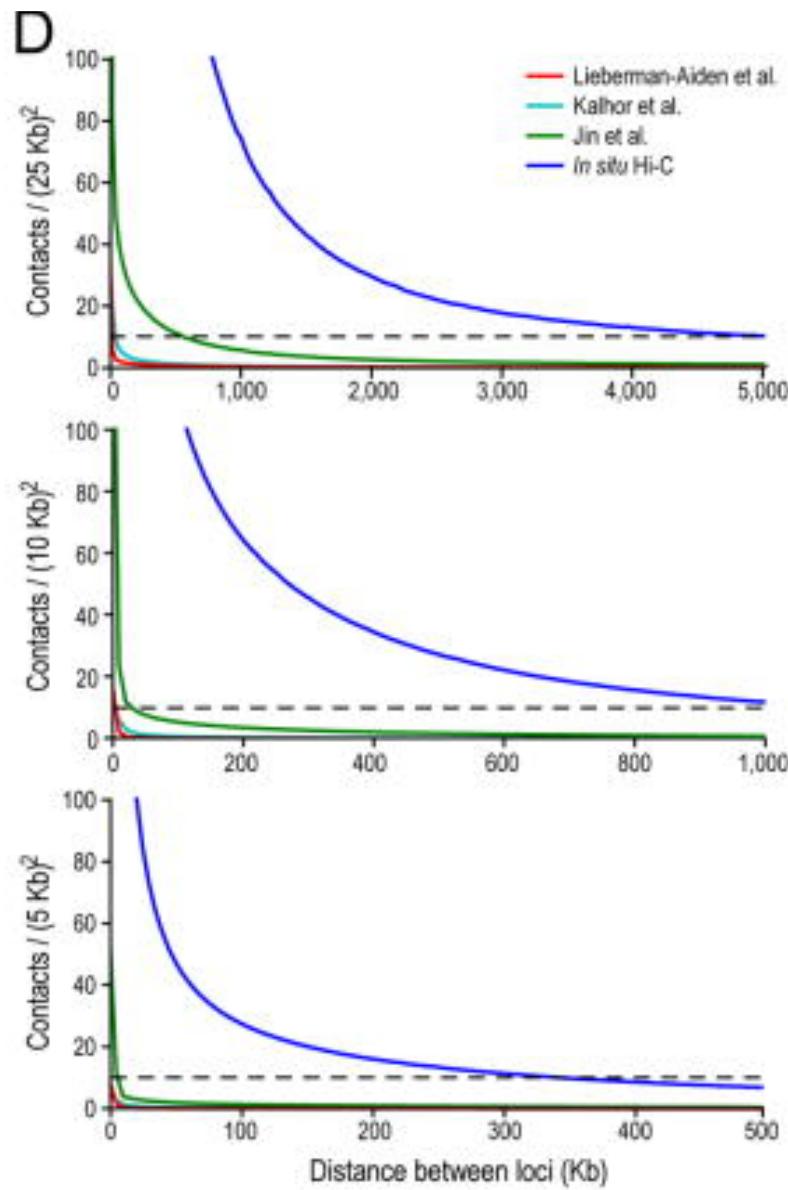


Cell. 2014 Dec 18; 159(7): 1665–1680.

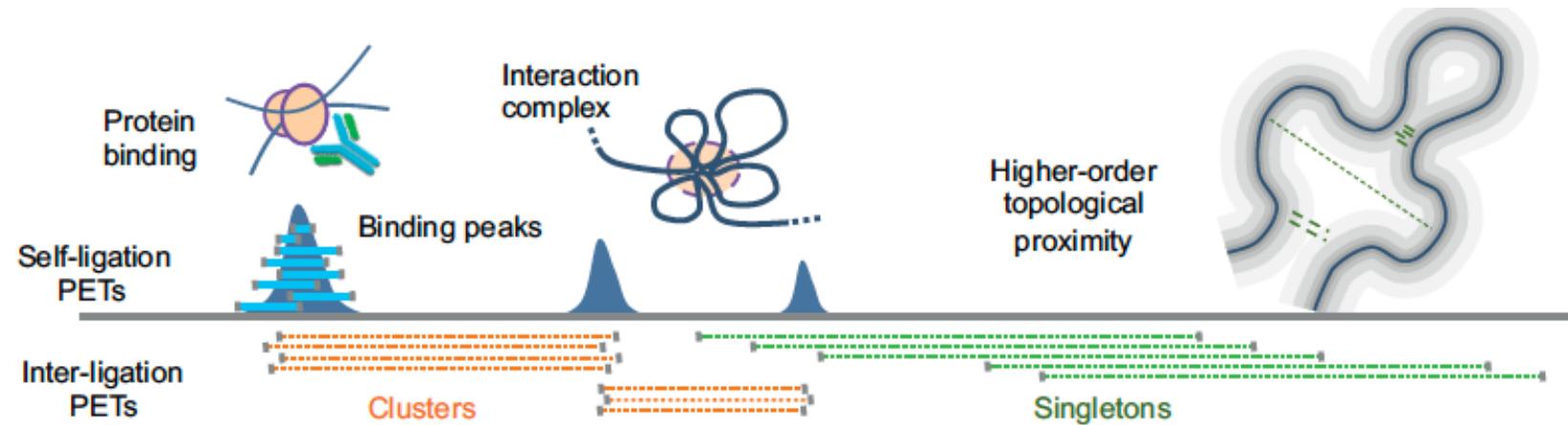
in situ HiC reveals interactions at 1 – 5 KB resolution



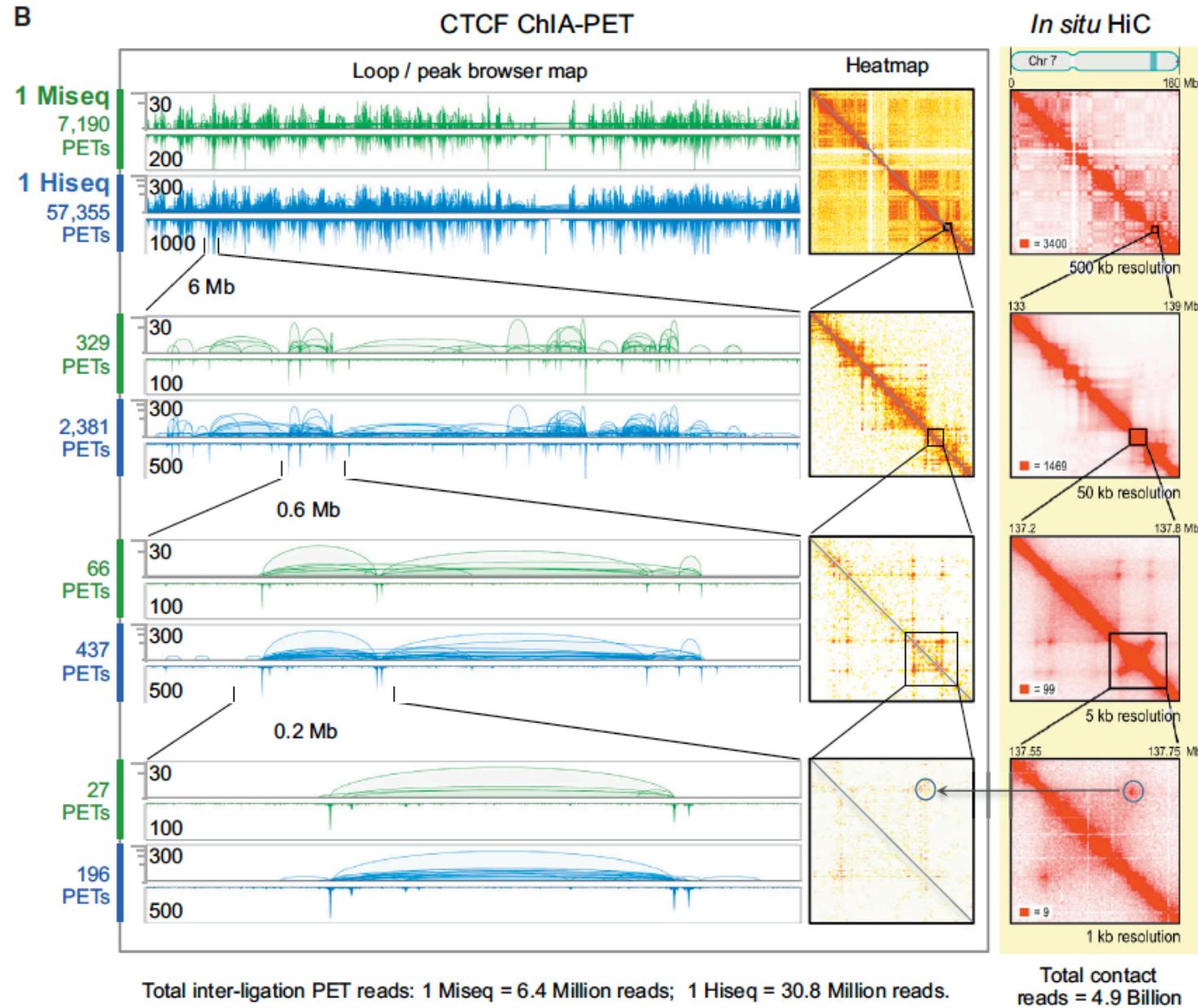
Observed interchromosomal interaction distances fall off exponentially



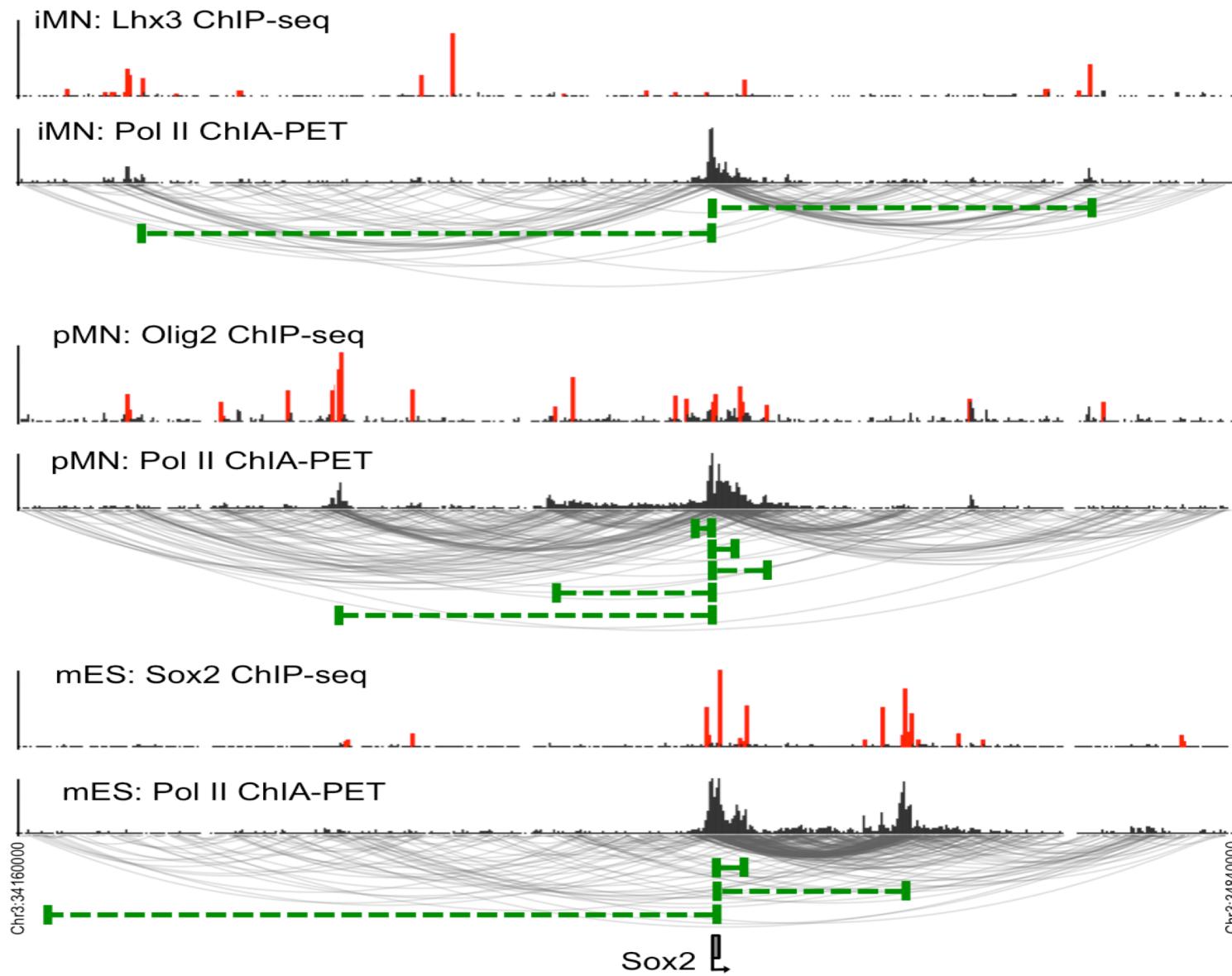
ChIA-PET identifies protein mediated interactions and improves resolution for those events



ChIA-PET data are consistent with HiC data



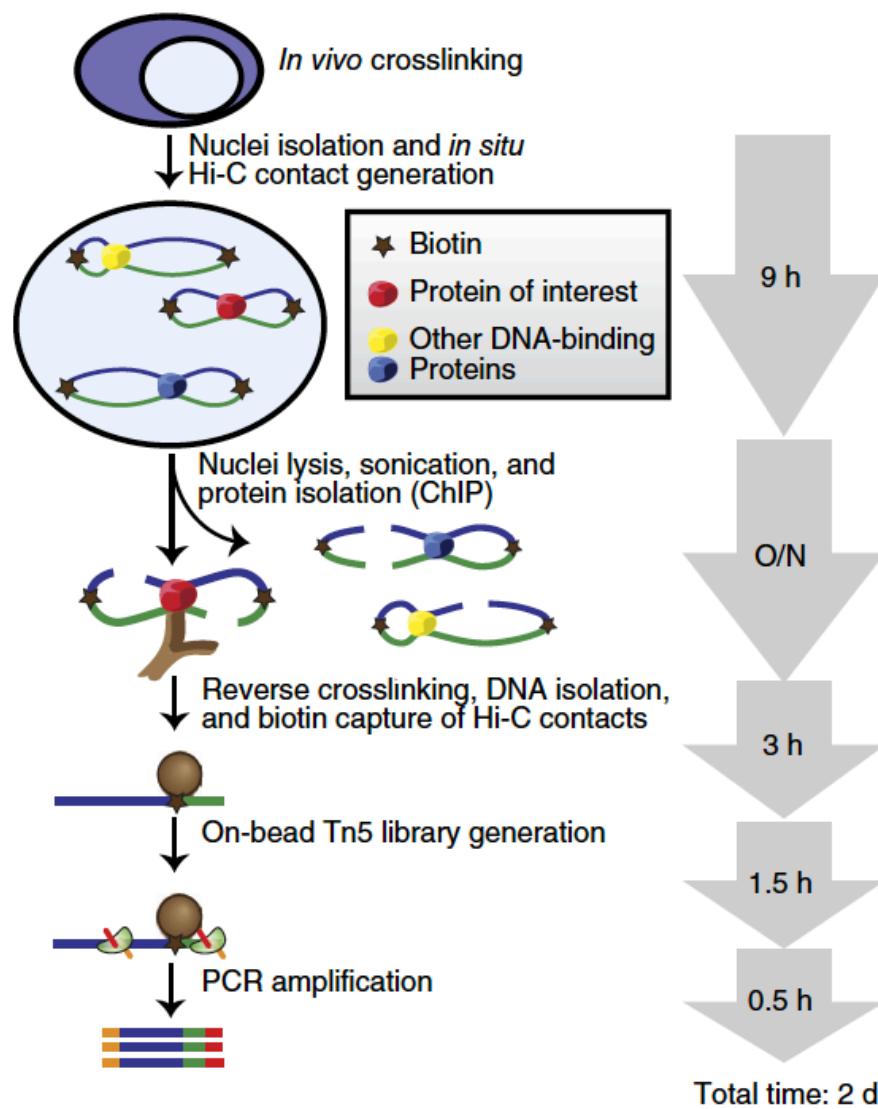
ChIA-PET discovered enhancer linkages



Issues with ChIA-PET

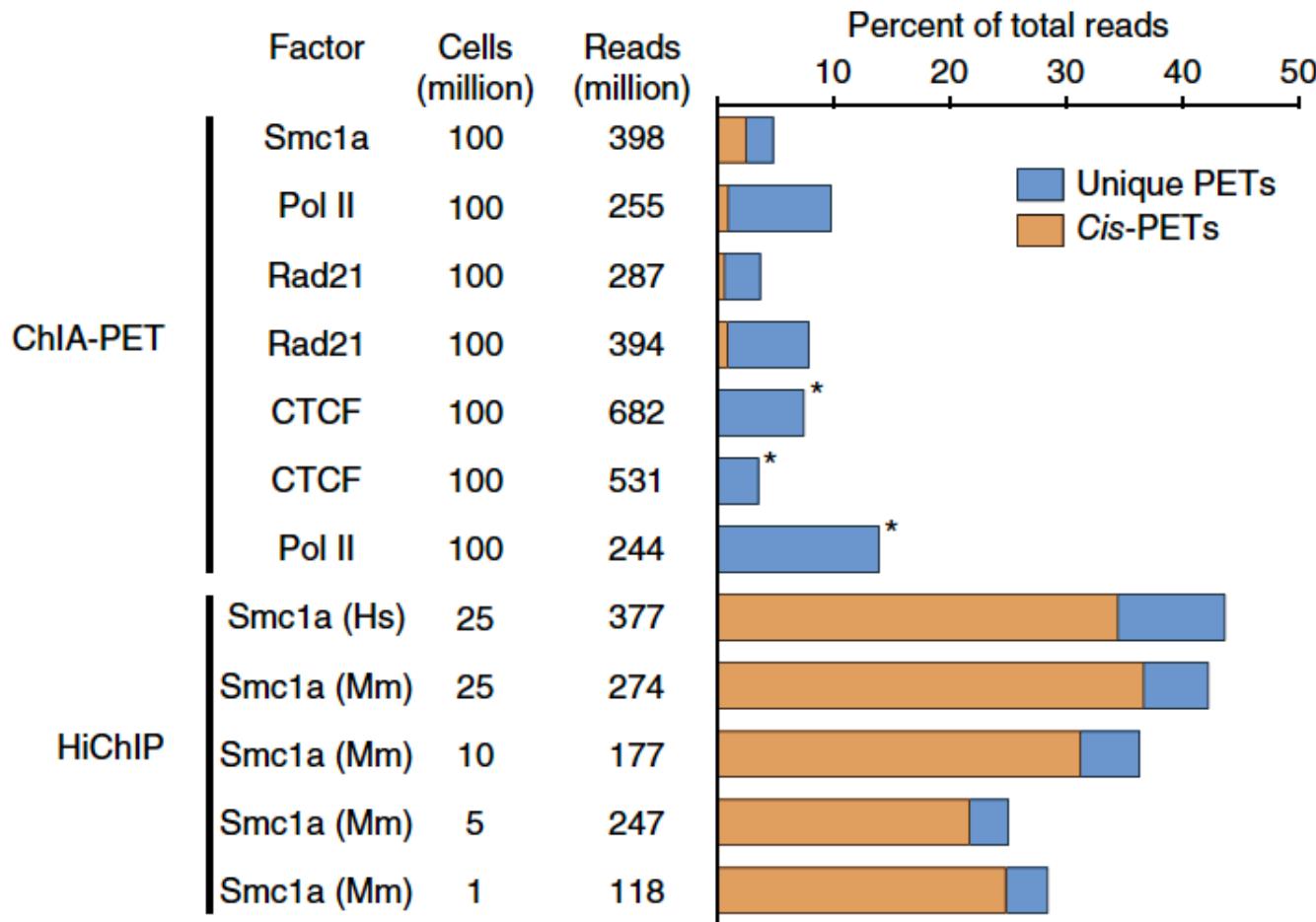
- 1.** High false negative rate. Libraries produced are not complex enough to permit further discovery by additional sequencing.
- 2.** Specific to a protein (RNA Polymerase II in our example)
- 3.** Hi-C and derivatives may solve these problems eventually

HiChIP identifies protein mediated interactions

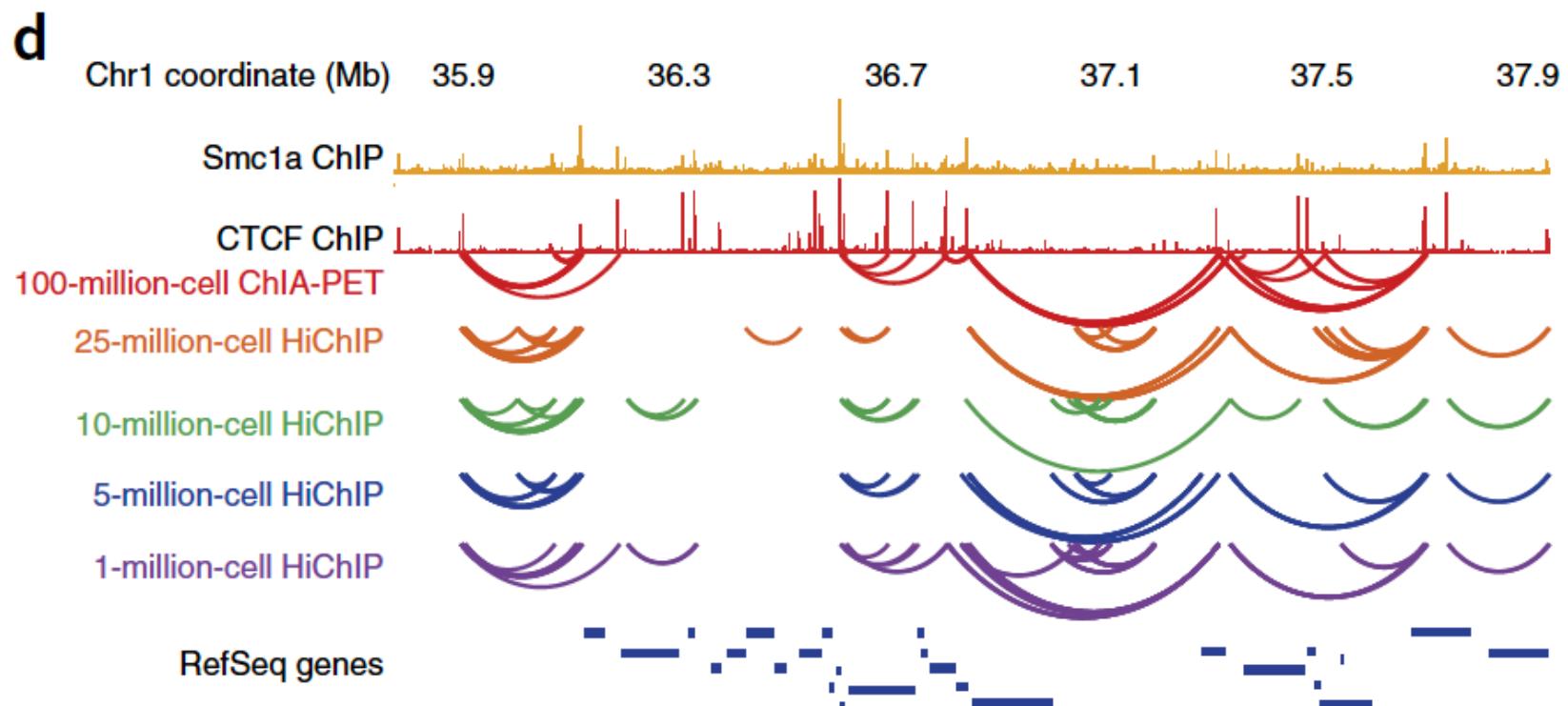


HiChIP is more sensitive than ChIA-PET

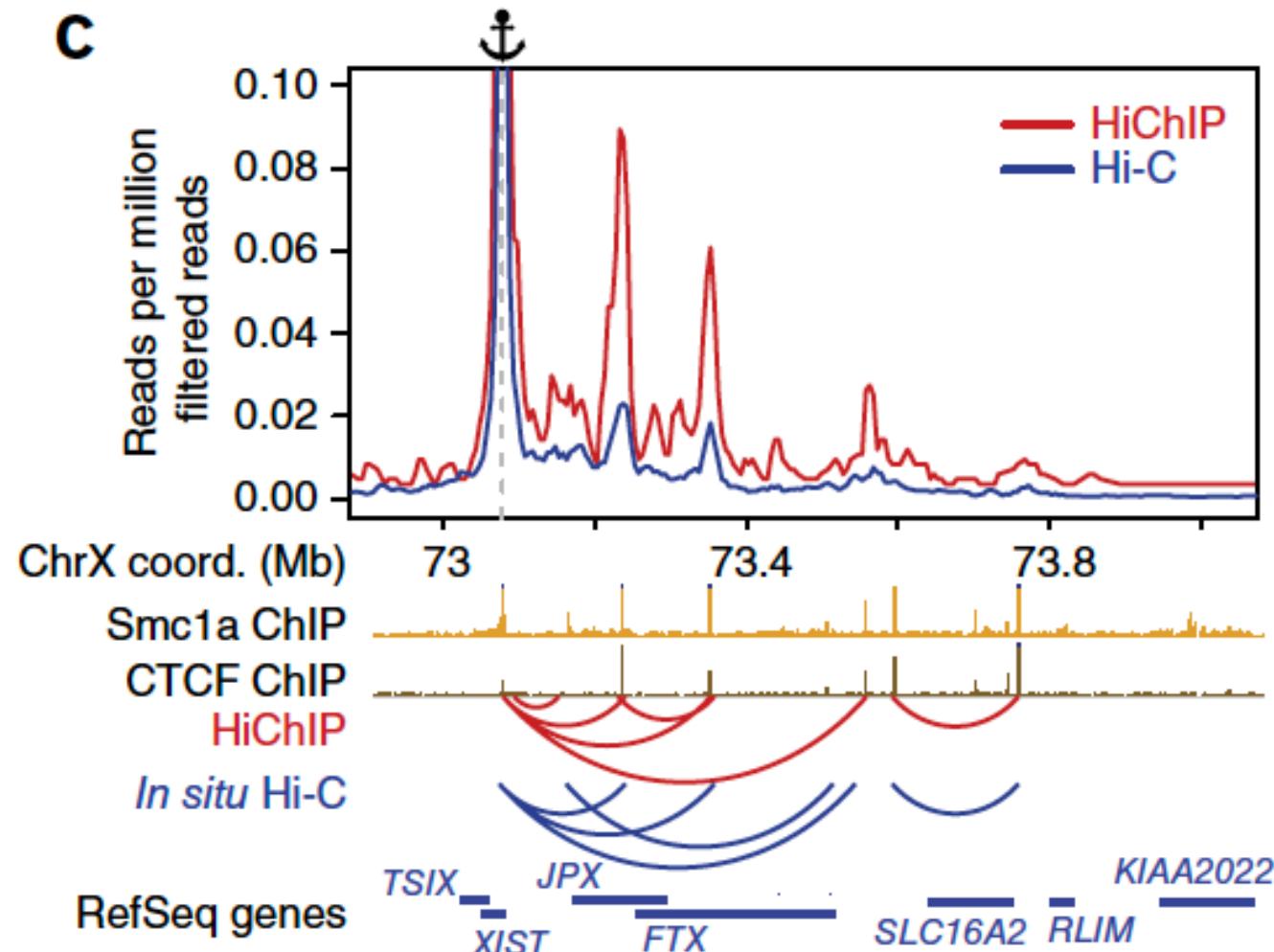
b



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

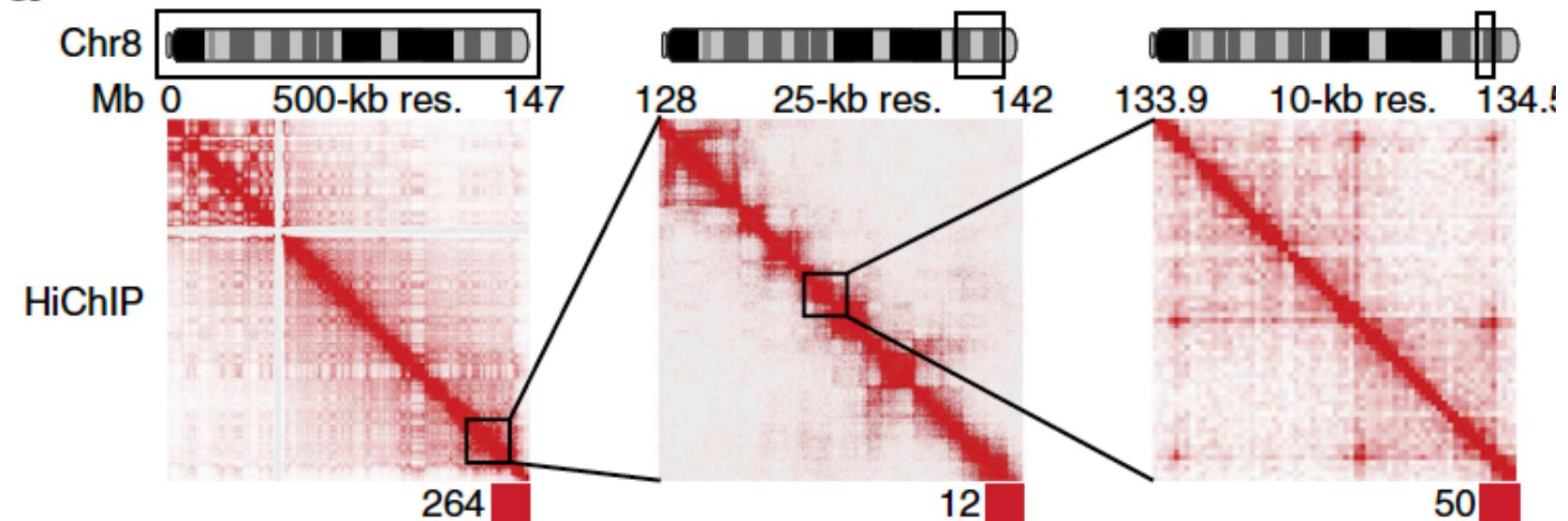


XIST promoter interactions show more support from HiChIP than Hi-C



HiChIP (Smc1a) is more sensitive than HiC

a



b

Discovering interactions

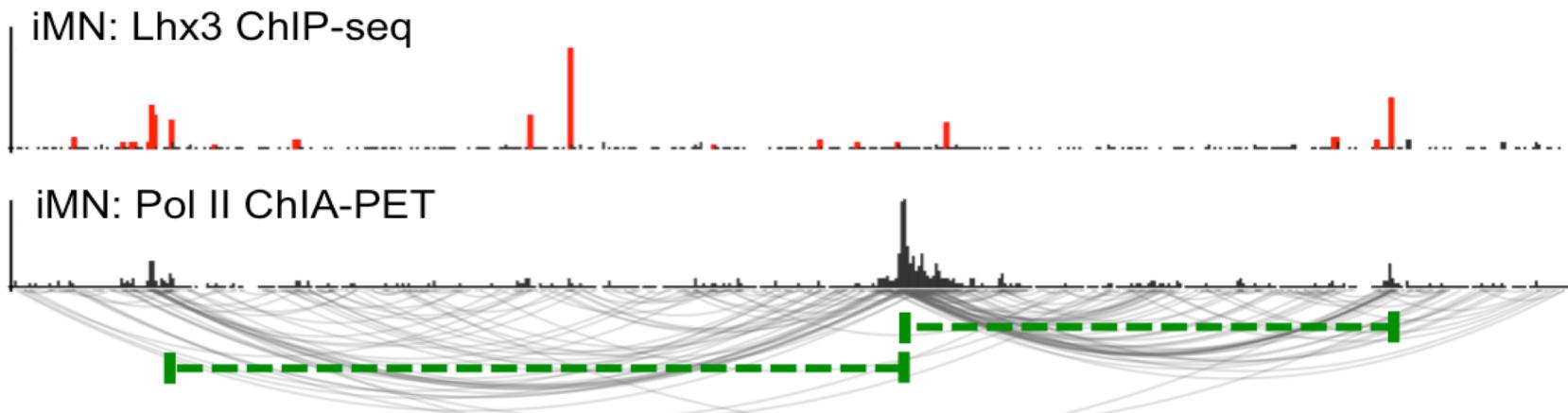
Method 1: Discover anchors using ChIP-seq methods
Given anchors, what is the chance of observing an interaction by chance?

N total ends

$I_{a,b}$ interactions observed

c_a ends

c_b ends



What is the chance of observing an interaction by chance?

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

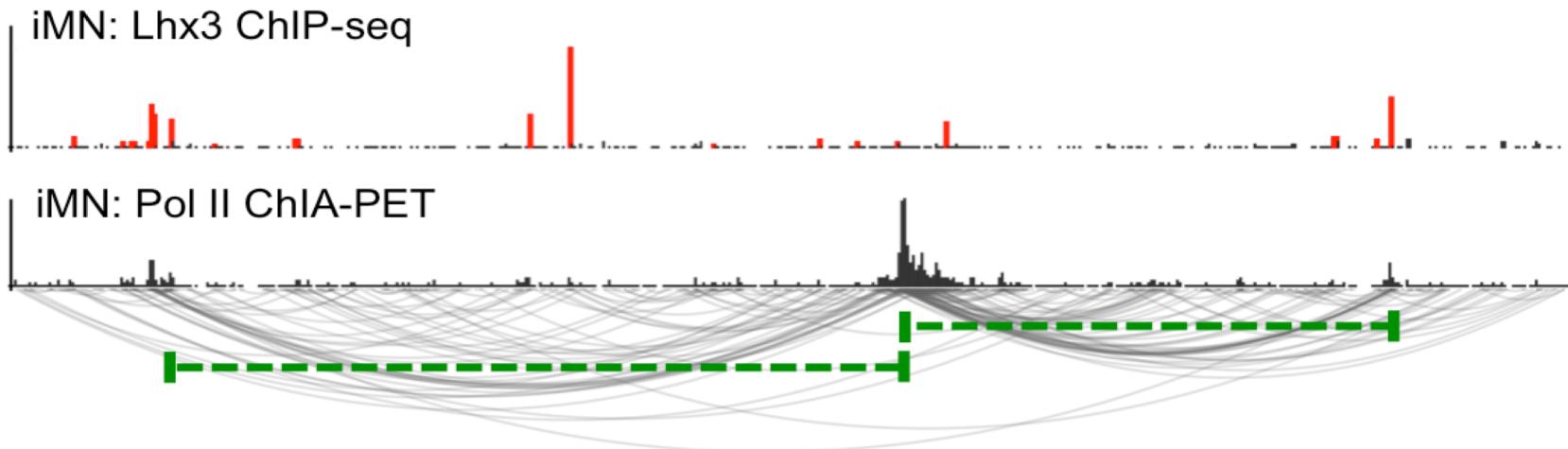
N total ends

$I_{a,b}$ interactions observed

c_a ends

c_b ends

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



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):

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

Using this model, we predict ~ 1100 total events

Approximate closed form solution for total number of events

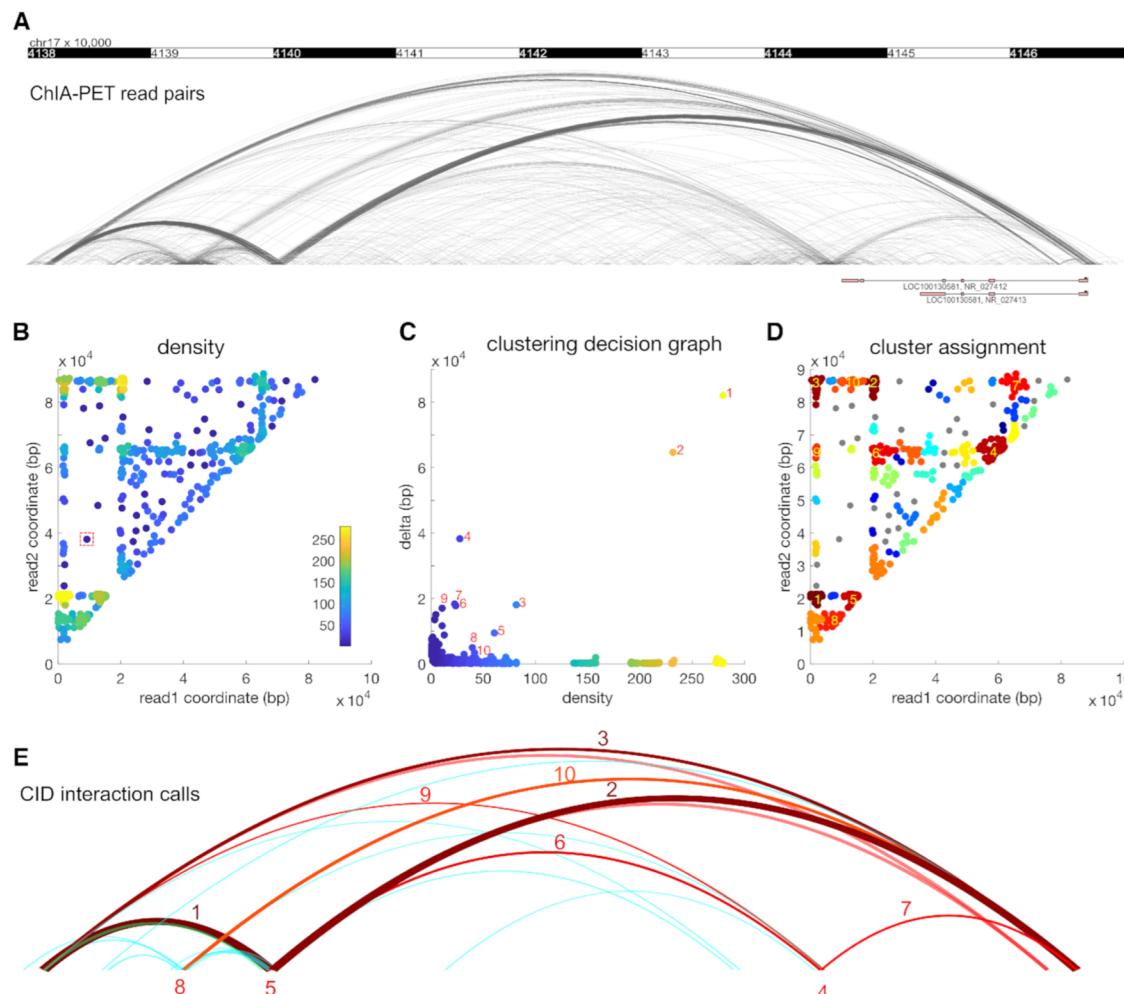
The ML estimate of N is approximately:

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

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

$$\begin{aligned} P(X = k; N, m, n) &\approx \text{Binomial}\left(X = k; n = n, p = \frac{m}{N}\right) \\ &\approx \text{Normal}\left(X = k; \mu = \frac{mn}{N}, \sigma^2 = \frac{mn}{N}\left(1 - \frac{m}{N}\right)\right) \end{aligned}$$

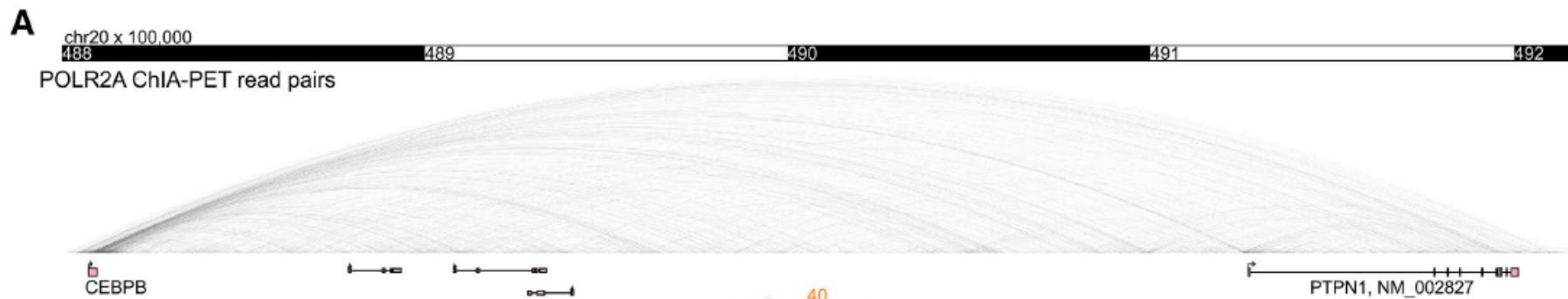
Method 2: CID uses density-based clustering to discover chromatin interactions



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



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

<https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz051/5319126>

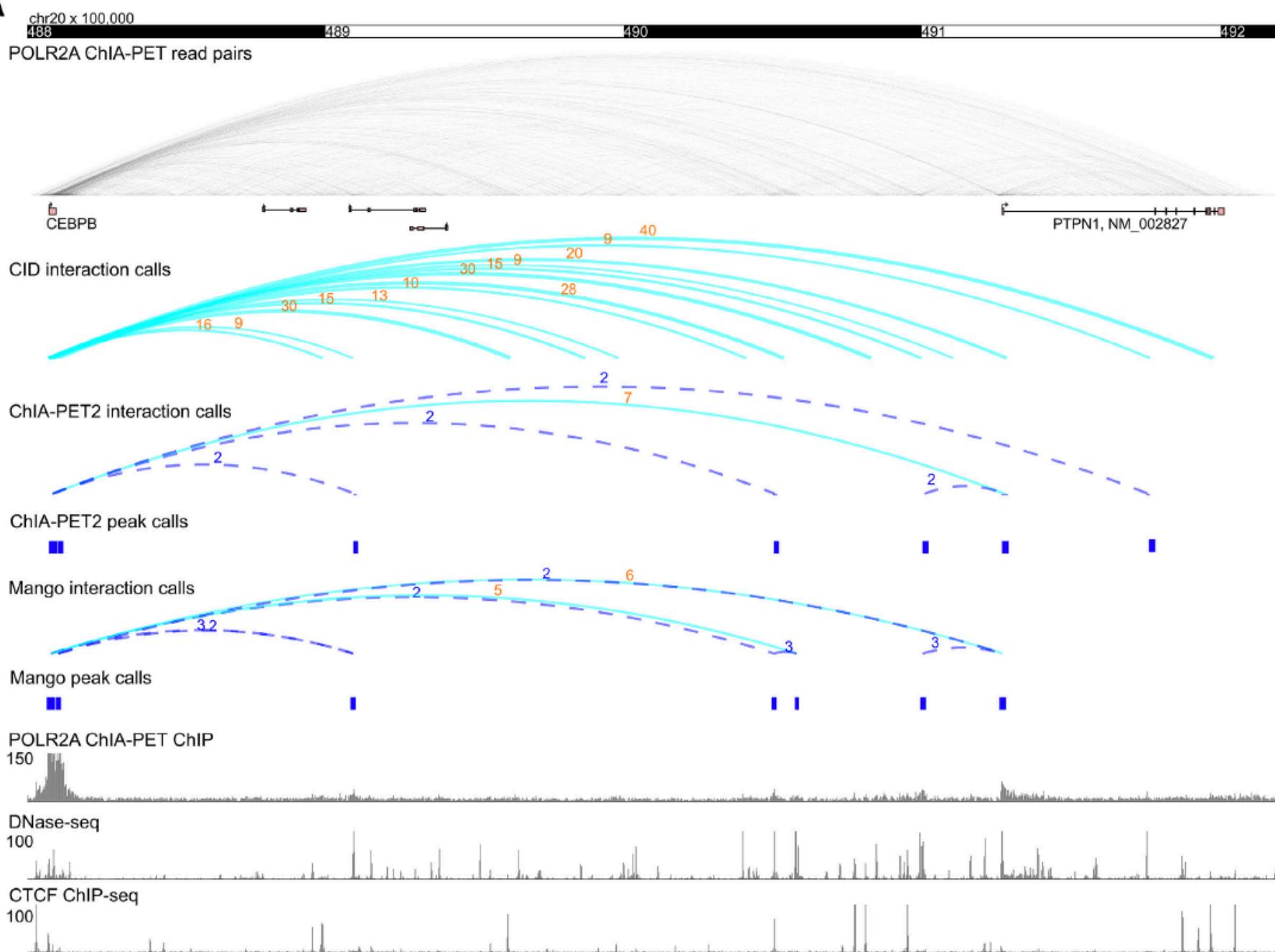
We use a three-component mixture model to describe conditional distribution of PET-count from all the PET clusters. One component represents true interaction PET cluster (TiPC), and the other two for random collision PET cluster (RcPC) and random ligation PET cluster (RIPC), respectively.

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

<https://academic.oup.com/bioinformatics/article/31/23/3832/208584>

Cluster interaction origins

A



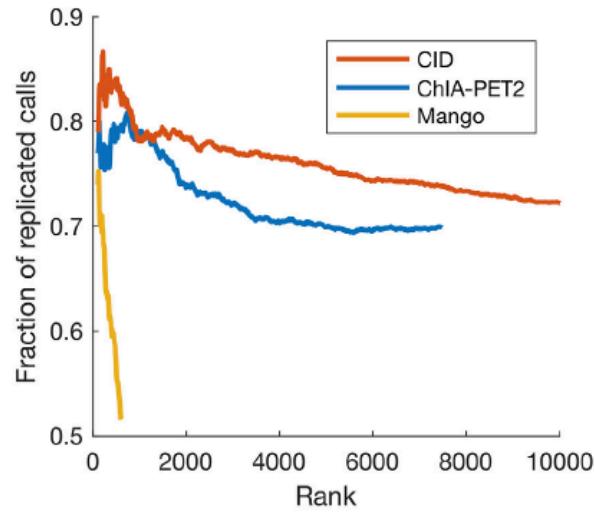
Jaccard coefficient – measure of set similarity

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

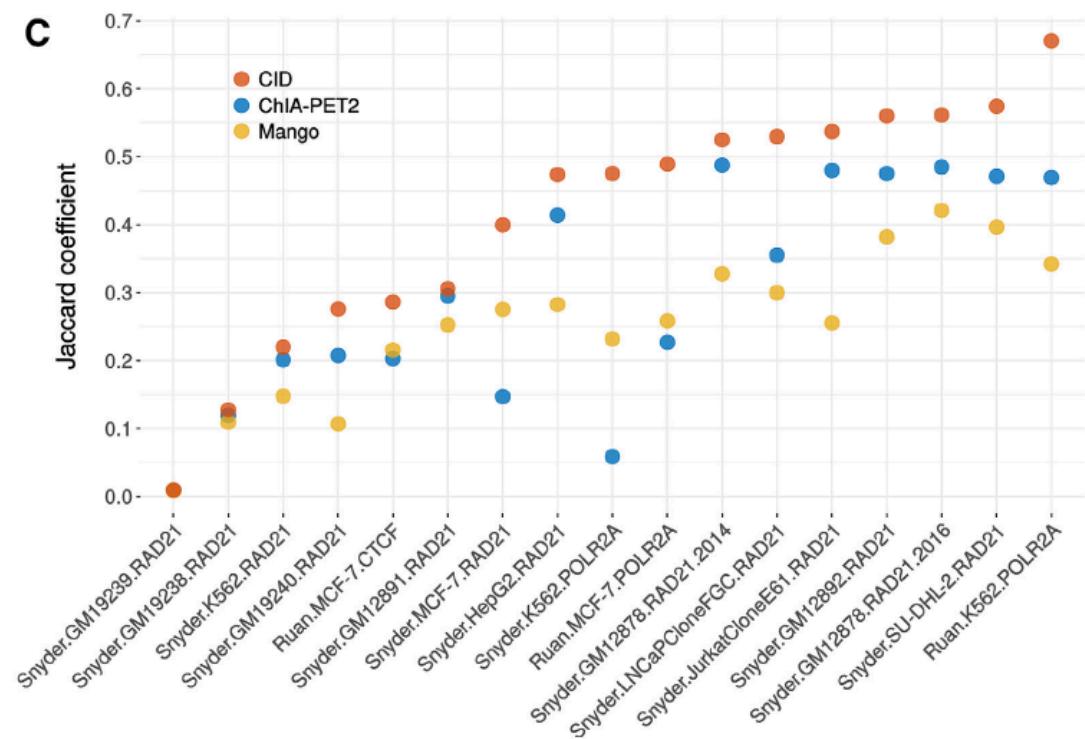
$$0 \leq J(A, B) \leq 1.$$

CID is more reproducible and sensitive

B

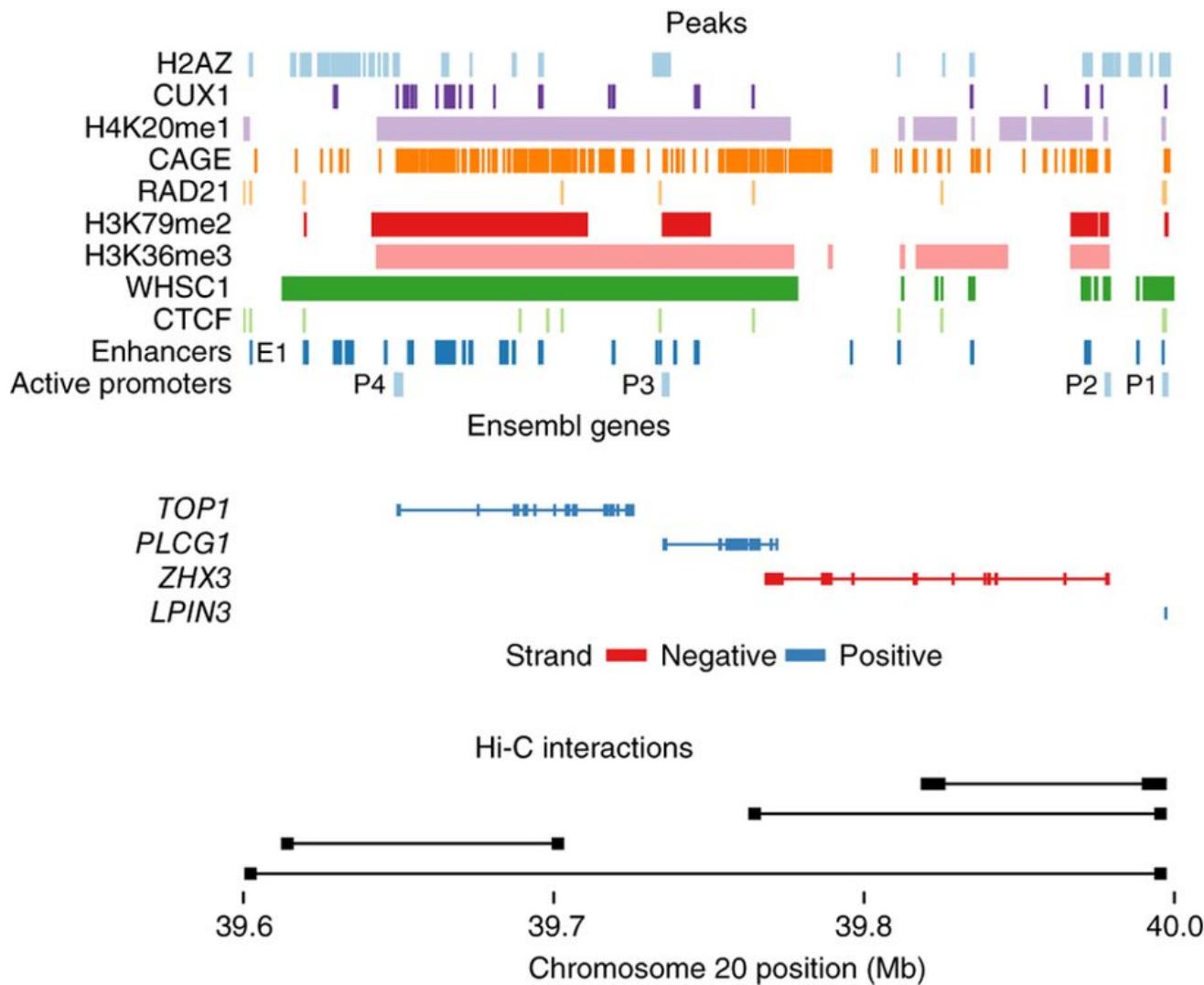


C



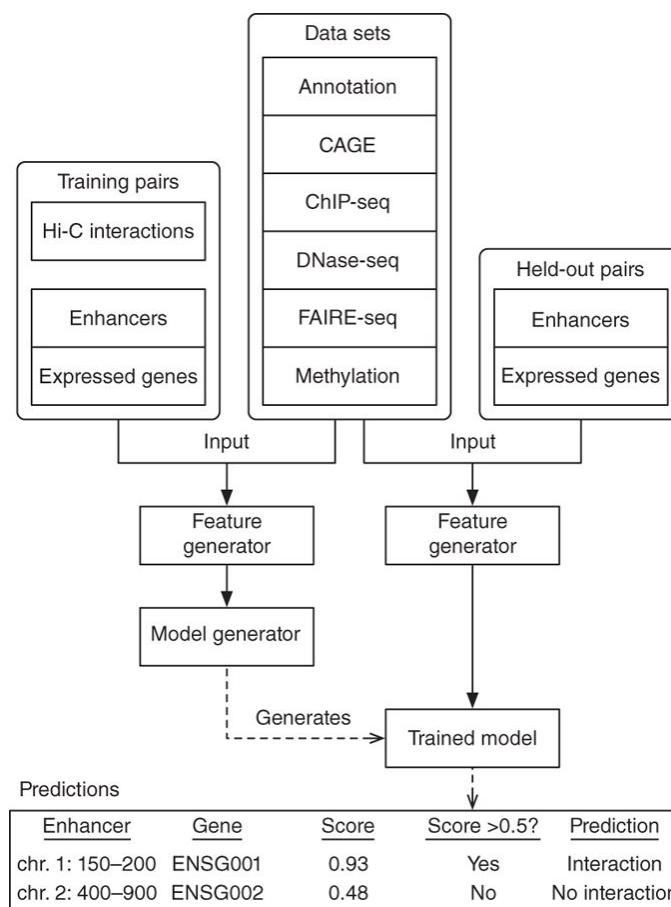
How can we predict interacting
enhancers and promoters?

TargetFinder uses multiple data types to predict HiC interactions

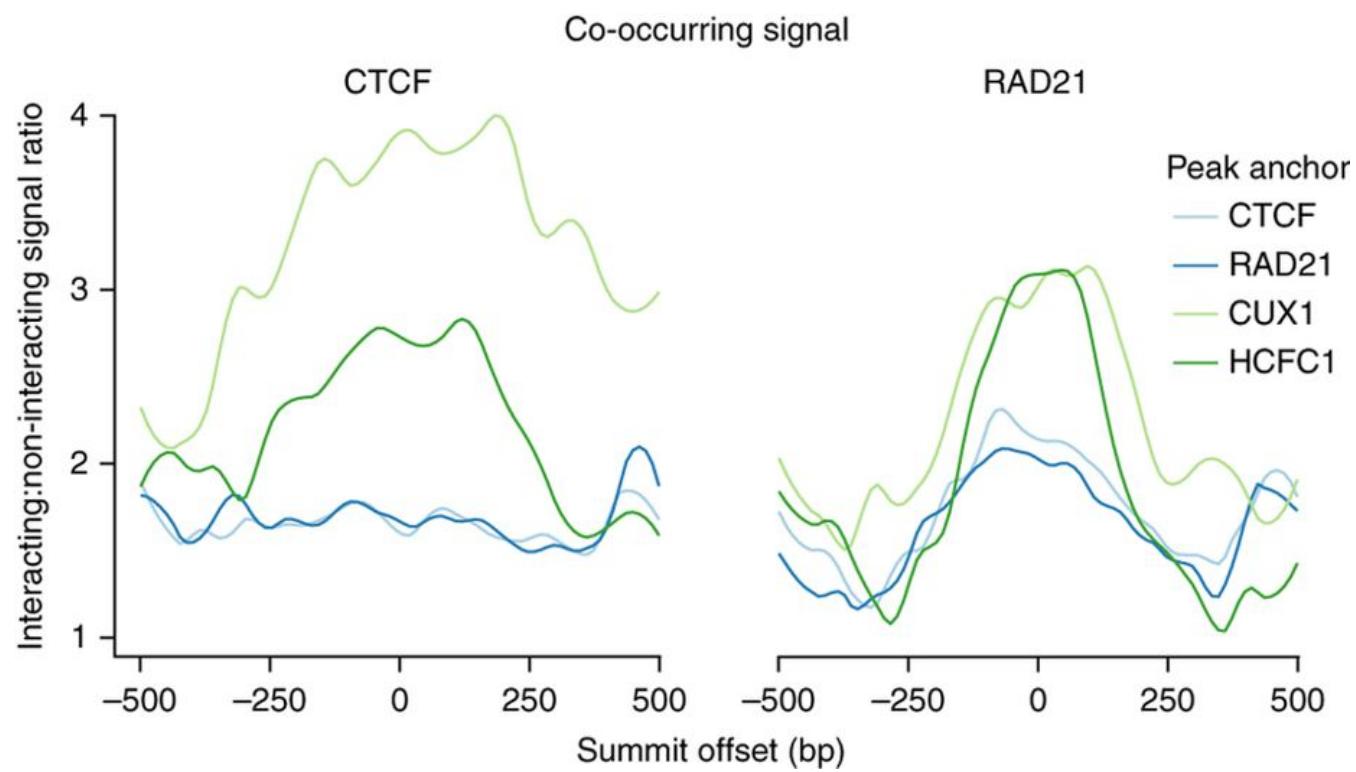


TargetFinder Training Data

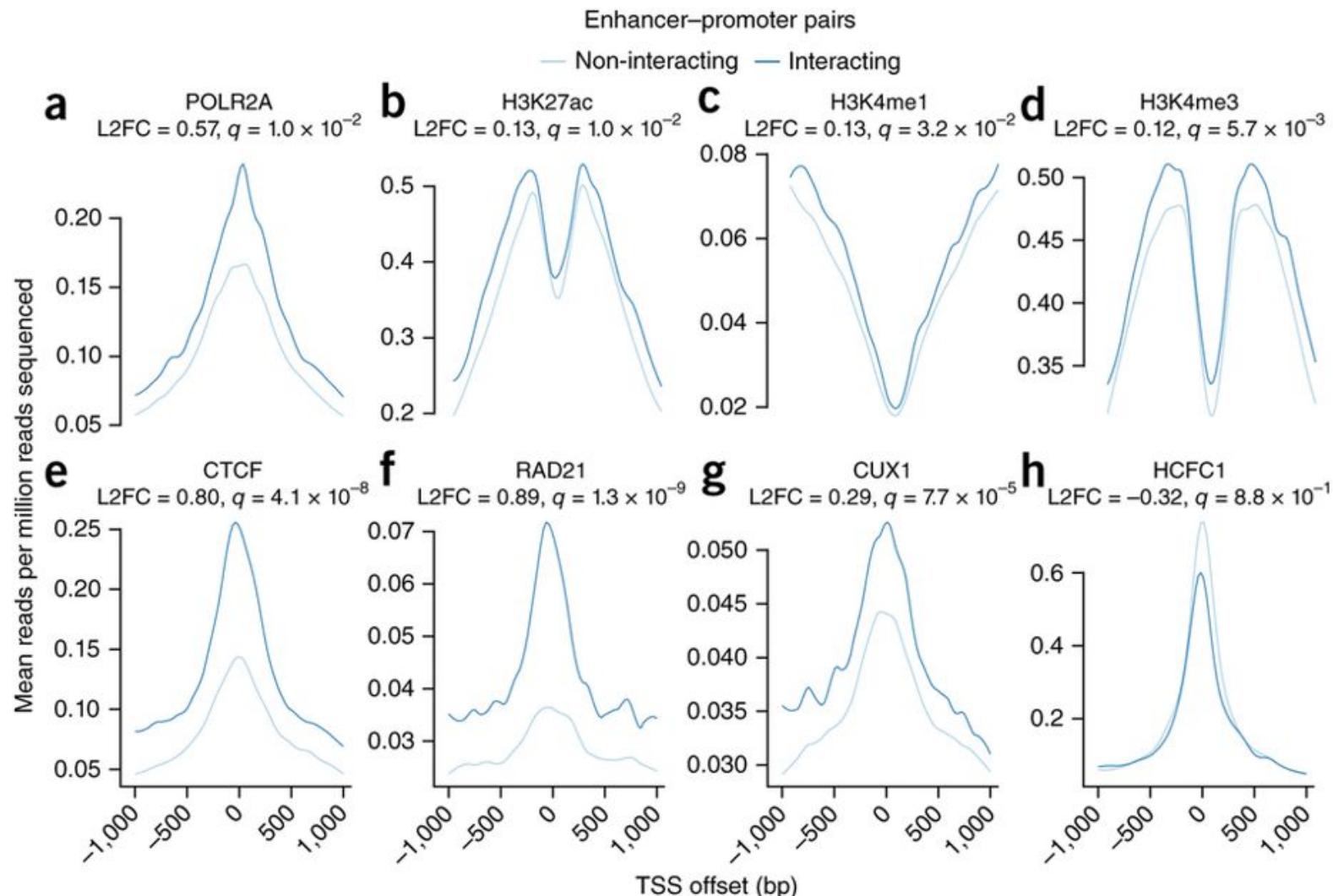
Cell Line	Promoters	Enhancers	Interacting Pairs	Non-Interacting Pairs
K562	8196	82806	1977	39500
GM12878	8453	100036	2113	42200
HeLa-S3	7794	103460	1740	34800
HUVEC	8180	65358	1524	30400
IMR90	5253	108996	1254	25000
NHEK	5254	144302	1291	25600



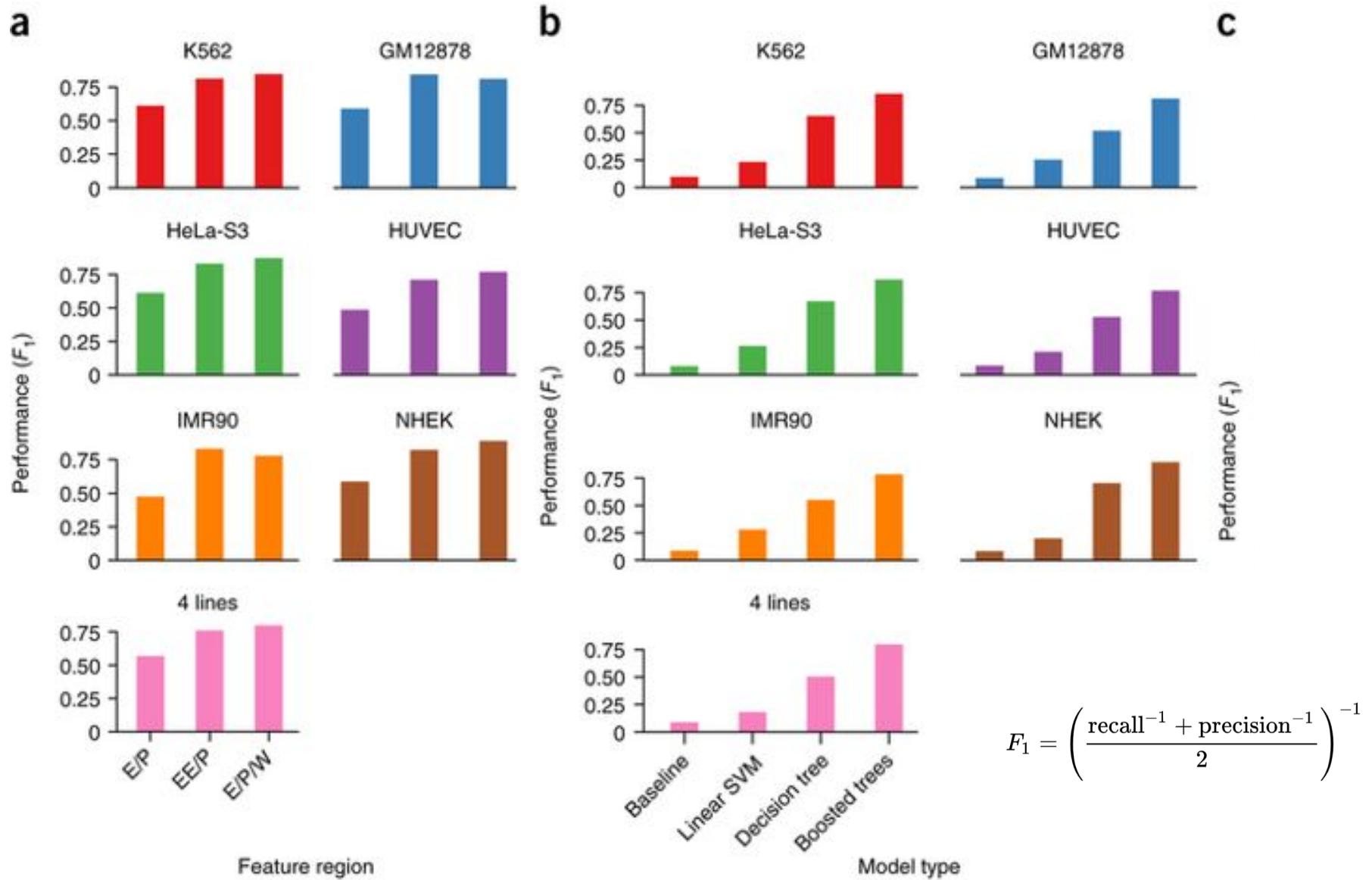
TargetFinder – Ratio of the CTCF and RAD21 ChIP-seq signals occurring within interacting enhancers and non-interacting enhancers



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

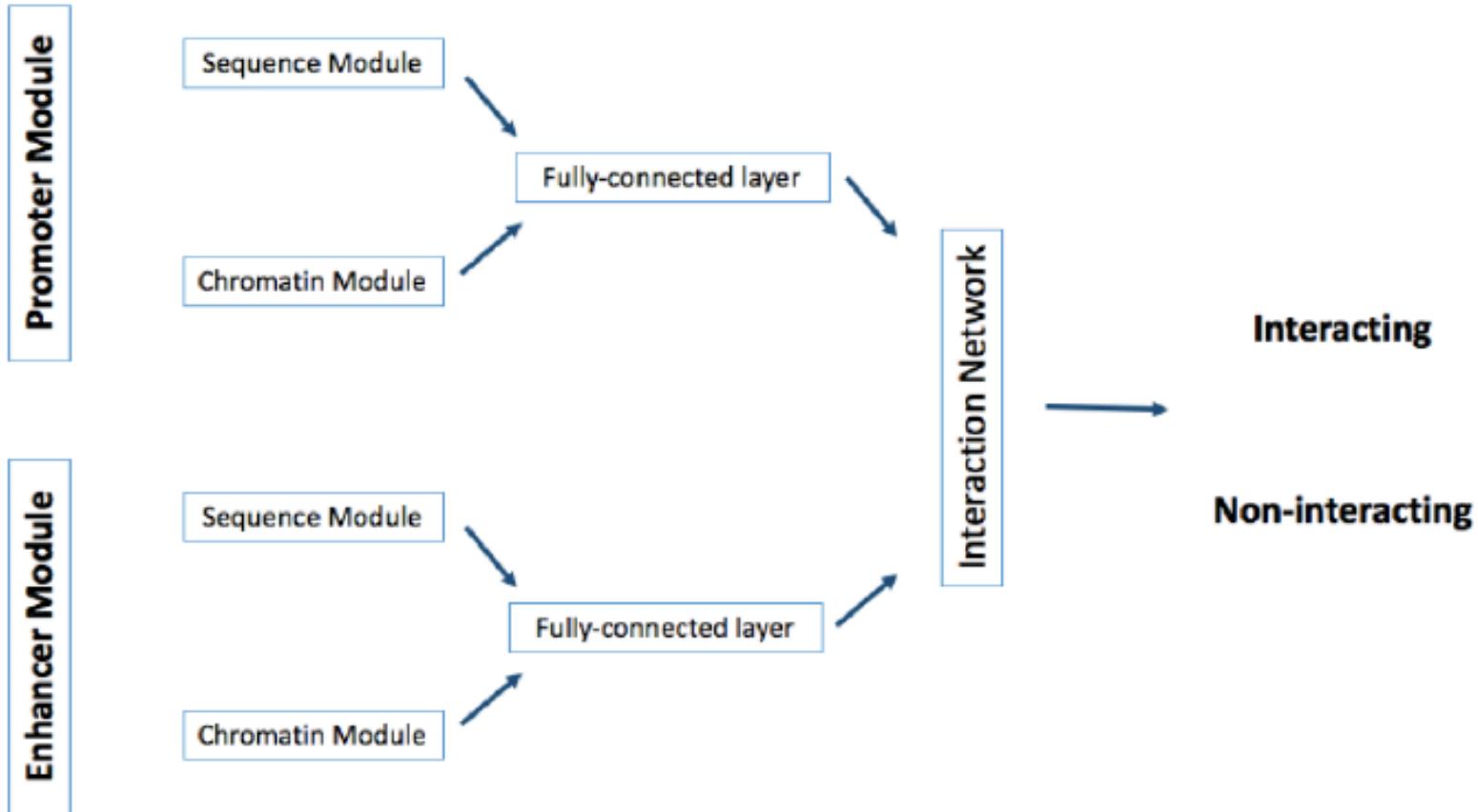


TargetFinder – Performance



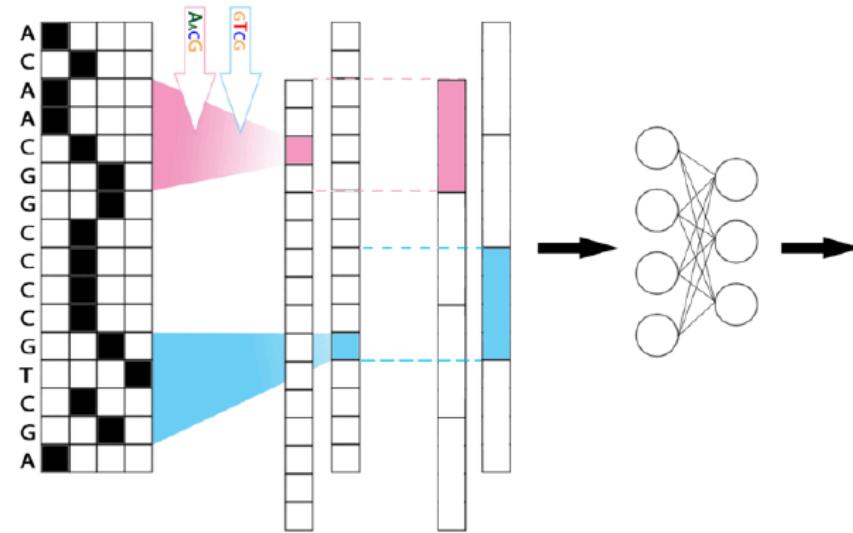
Features for enhancers and promoters only (E/P), extended enhancers and promoters (EE/P), and enhancers and promoters plus the windows between them (E/P/W)

Deep learning network for predicting enhancer-promoter interactions

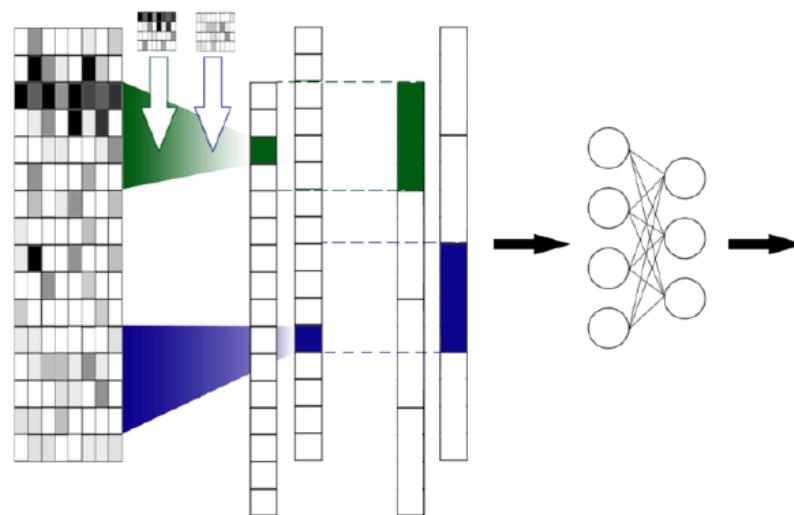


Sequence and chromatin anchor networks outputs are concatenated

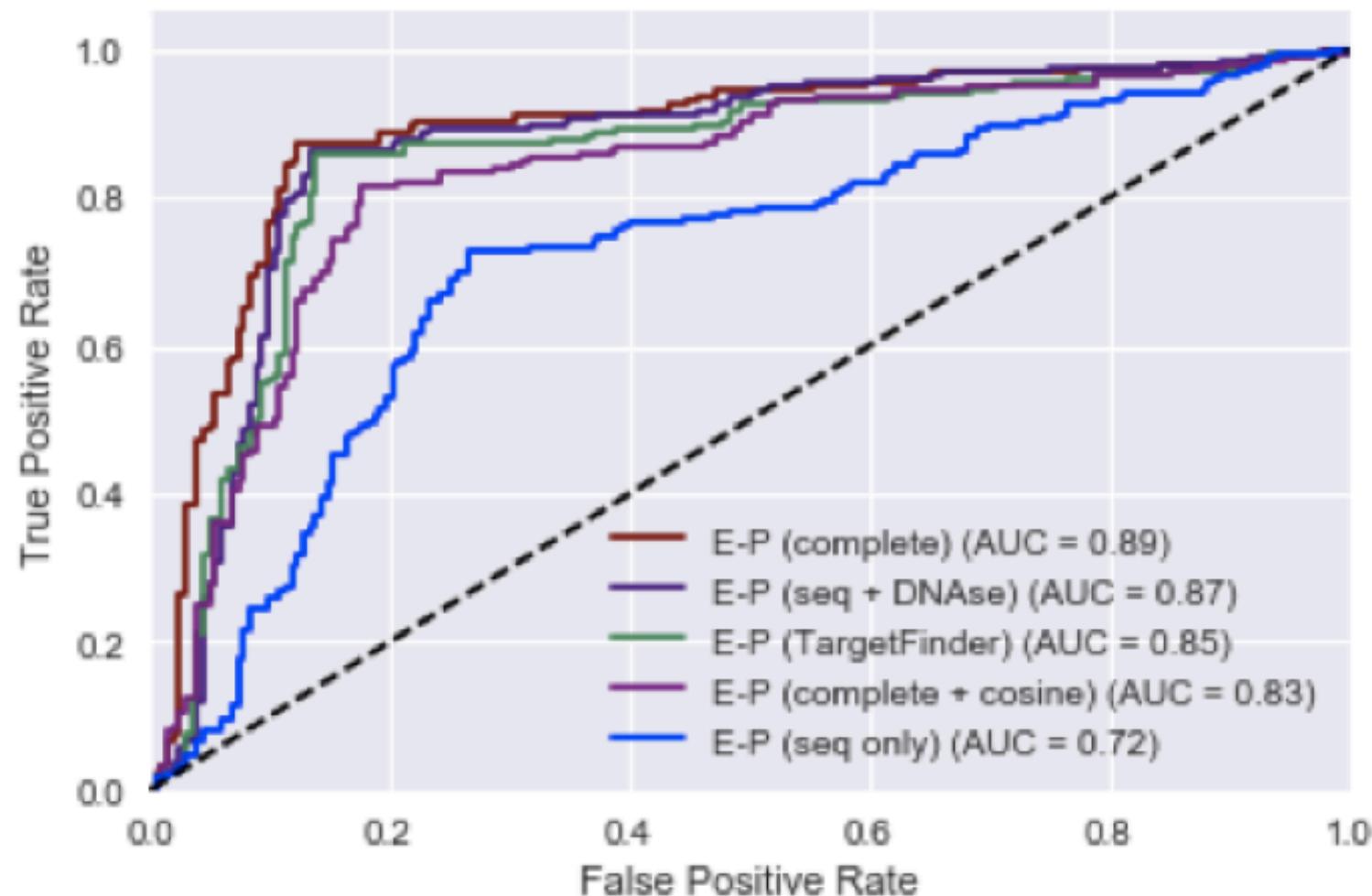
Sequence -
2kb
sequence
windows



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



Enhancer promoter prediction performance with varying feature sets



FIN - Thank You

Allowing for false positive events

- What if some events in each replicate are false positives?
Then we will overestimate the total event count
- We can assume that overlapping (shared) events are true positives and that $(1 - f)$ of the remaining events are false negatives, where f is the true positive rate (TPR)
- This approximation lets us update m and n and apply the same model:

$$m' = (1 - f)(m - k) + k$$

$$n' = (1 - f)(n - k) + k$$

A higher true positive rate estimates more total events with a fixed overlap

- Replicate A had 3811 events, replicate B had 1384 events
- The overlap was 533 events
- Likelihood plots versus N for several true positive rates (TPR):

