Genome-scale technologies 2/ Algorithmic and statistical aspects of DNA sequencing

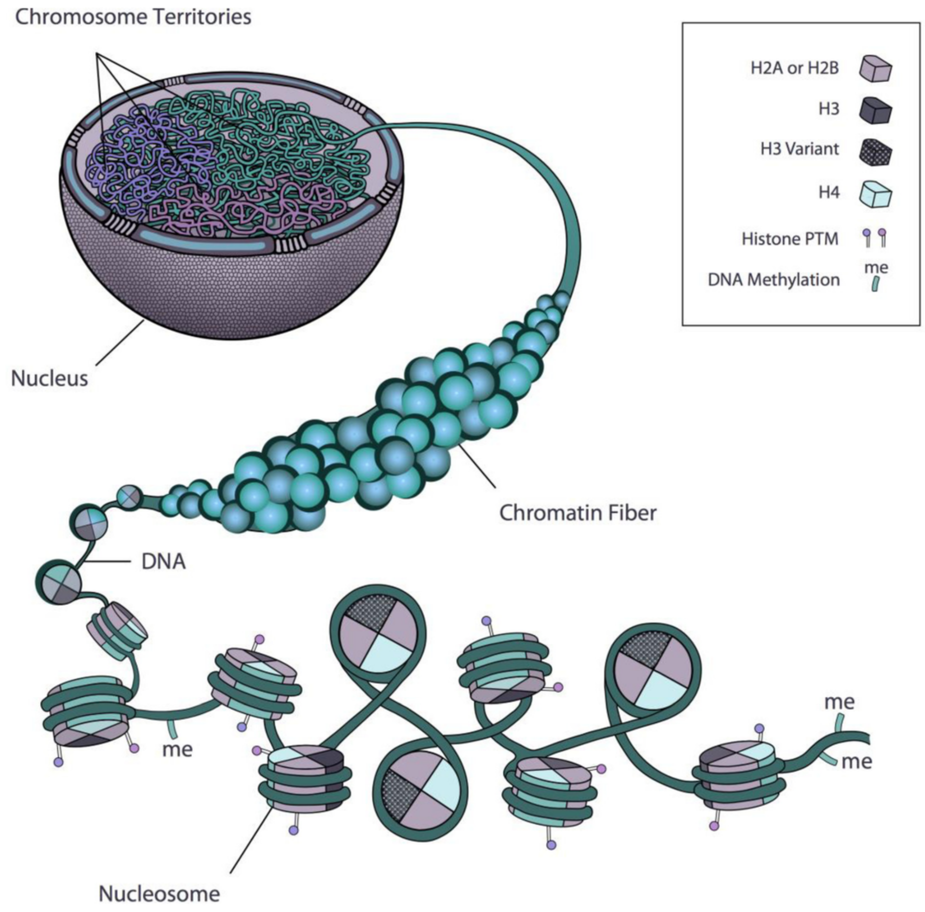
Studying chromatin with Hi-C

Ewa Szczurek
University of Warsaw, MIMUW

szczurek@mimuw.edu.pl

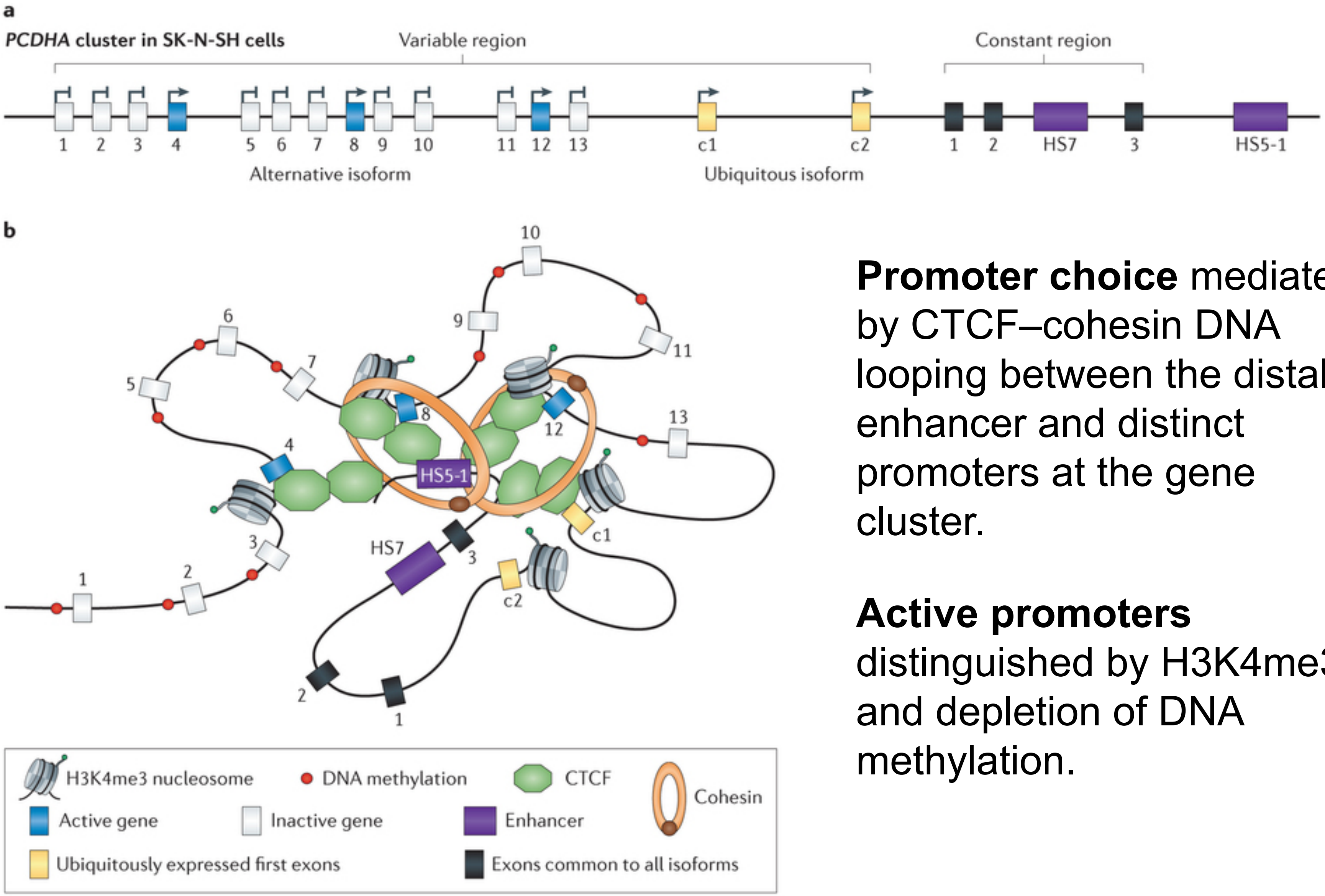
Chromatin organization

- **Compression**: 2 meters DNA → 10 micrometers nucleus
- **Accessibility:** for protein machineries that regulate:
    - Replication
    - Repair
    - Recombination
    - Gene expression



http://atlasgeneticsoncology.org/Educ/ChromatinEducEng.html Rosa and Shaw, *Biology* 2013, 2(4), 1378-1410

Impact on gene reg: far enhancers brought to promoters



Promoter choice mediated by CTCF–cohesin DNA looping between the distal enhancer and distinct promoters at the gene cluster.

Active promoters distinguished by H3K4me3 and depletion of DNA methylation.

Chin-Tong Ong & Victor G. Corces Nature Reviews Genetics 15, 234–246 (2014)

The project

- http://students.mimuw.edu.pl/~szczurek/TSG2_Project/project.html
- Report deadline: 20.01.2016
- Presentations: 26.01.2016

History: Chromosome Conformation Capture

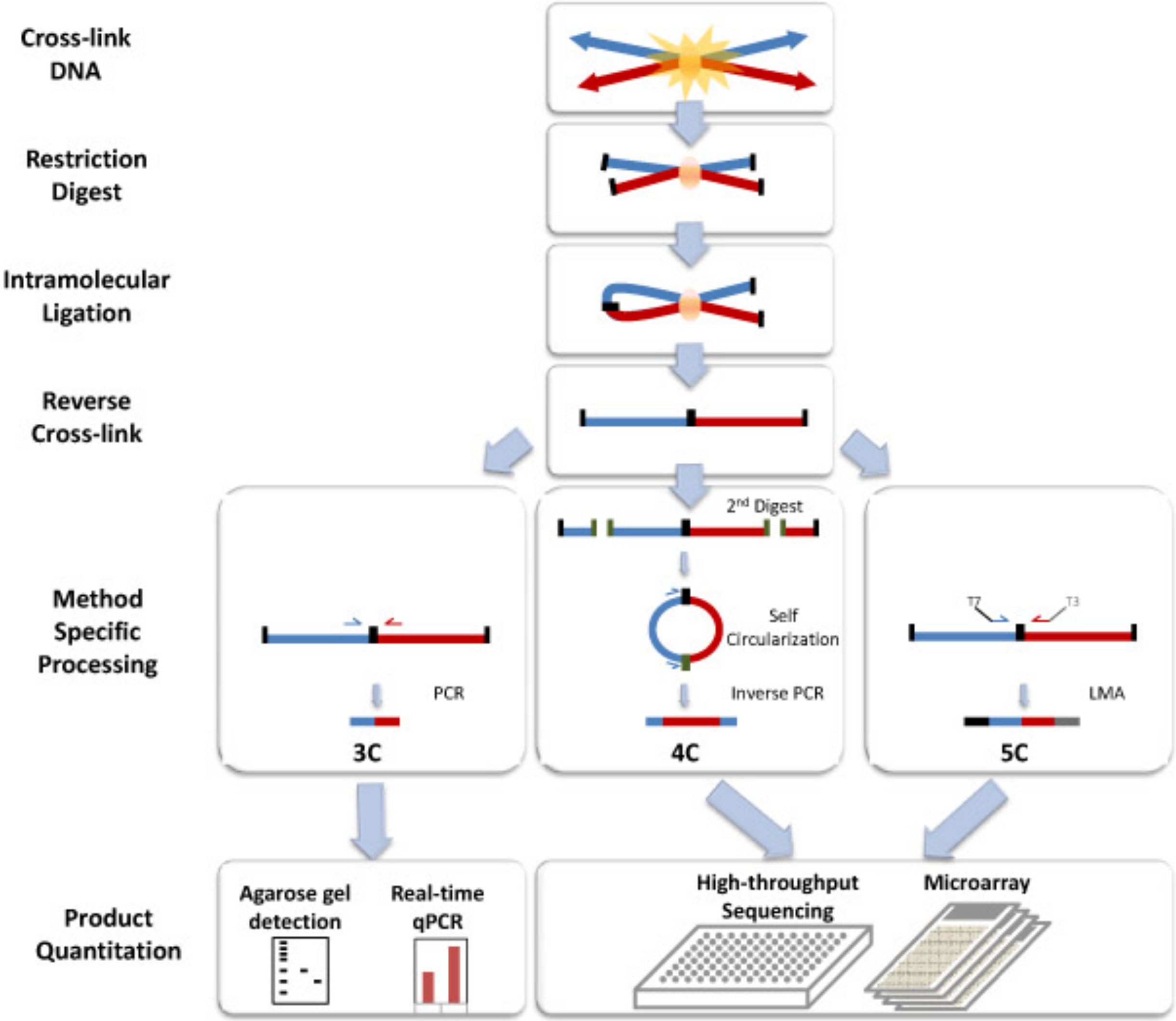
In the order of increasing throughput:

3C: Chromosome Conformation Capture

4C: Circularized 3C

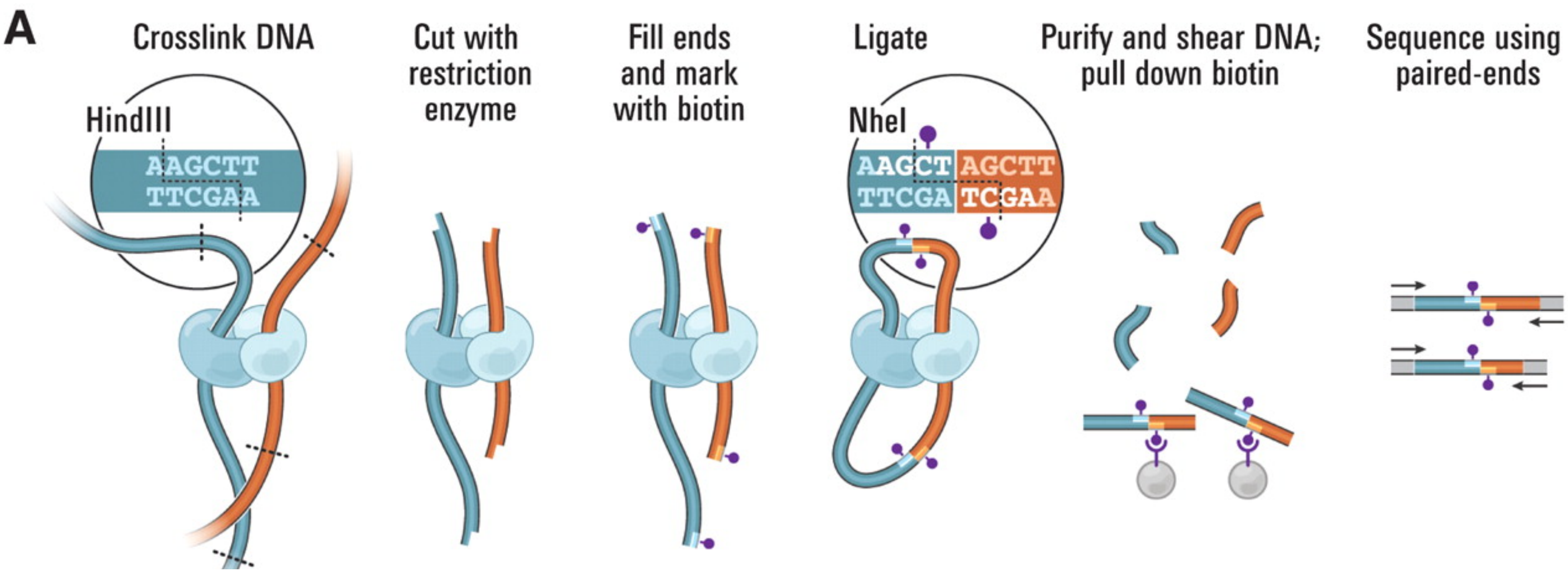
5C: Carbon Copy 3C

All require choosing a set of target loci and do not allow unbiased genomewide analysis.



Wikipedia

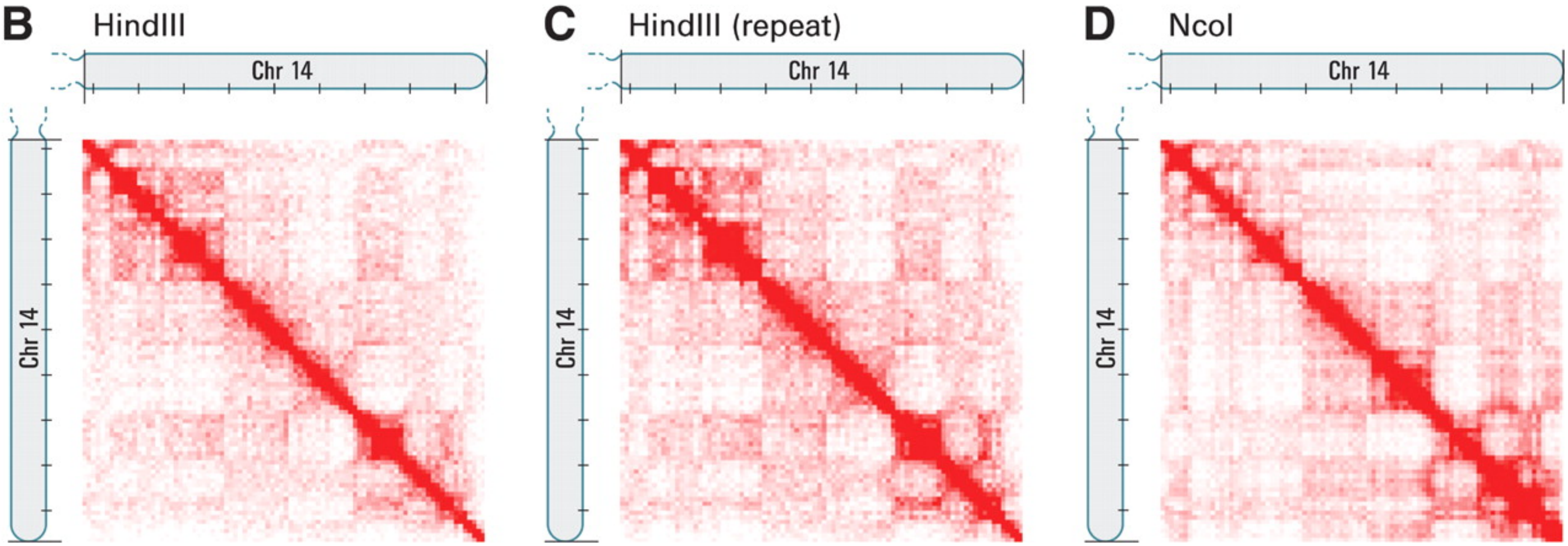
Now: Hi-C


A
Crosslink DNA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends
HindIII
AAGCTT
TTCGAA
NheI
AAGCTAGCTT
TTCGATCGAA

- DNA digested with a restriction enzyme that leaves a 5′ overhang;
- the 5′ overhang filled, including a biotinylated residue;
- the blunt-end fragments ligated (ligation of the cross-linked DNA)
- Resulting DNA sample: fragments that were originally **in close spatial proximity** in the nucleus, marked with biotin at the junction.
- Hi-C library: shearing the DNA and selecting the biotin-containing fragments with streptavidin beads.
- The library massively parallel DNA sequenced → a catalogue of interacting fragments

Lieberman-Aiden et al. 2009

Hi-C produces a genome-wide contact matrix

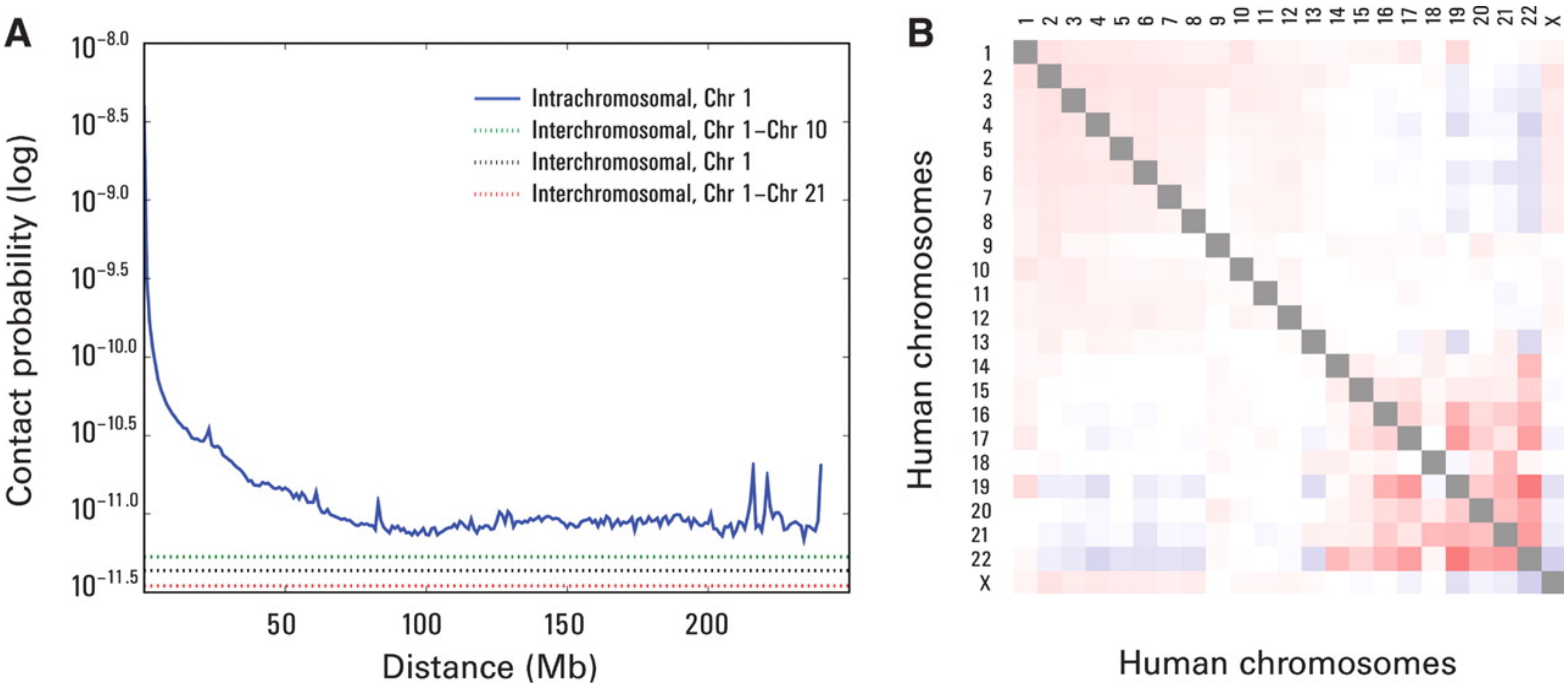


- Each pixel: all interactions between 1-Mb locuses
- Intensity: the total number of reads (0 to 50).
- Tick marks every 10 Mb.

C) a biological repeat using the same restriction enzyme
D) a different restriction enzyme

Lieberman-Aiden et al. 2009

The presence and organization of chromosome territories



(A) Contact prob. decreases with distance.

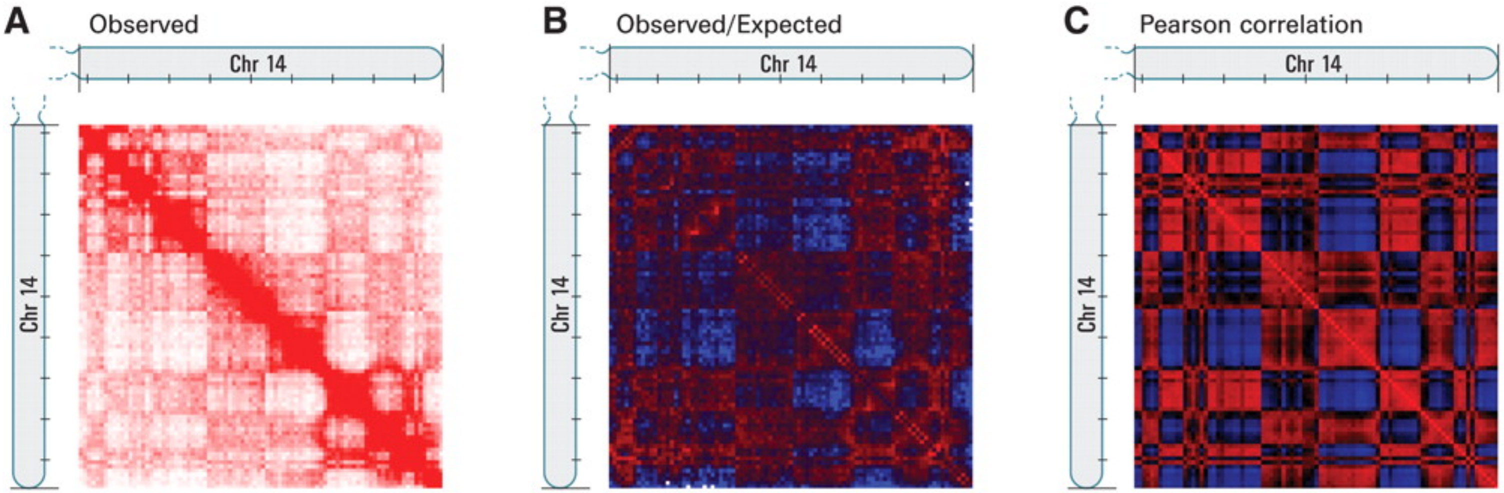
- Contacts more probable within than between chromosomes.

(B) Observed/expected number of interchromosomal contacts

- Red: enrichment, blue: depletion (range from 0.5 to 2).
- Small, gene-rich chromosomes interact more with one another, suggesting that they cluster together in the nucleus.

Lieberman-Aiden et al. 2009

Nucleus is segregated into to open & closed chromatin



(A) Substructure: intense diagonal, a constellation of large blocks

(B) Observed/expected matrix: each entry divided by the genome-wide average contact probability for loci at same genomic distance

- more (red) or less (blue) interactions than would be expected, given their genomic distance (range from 0.2 to 5).

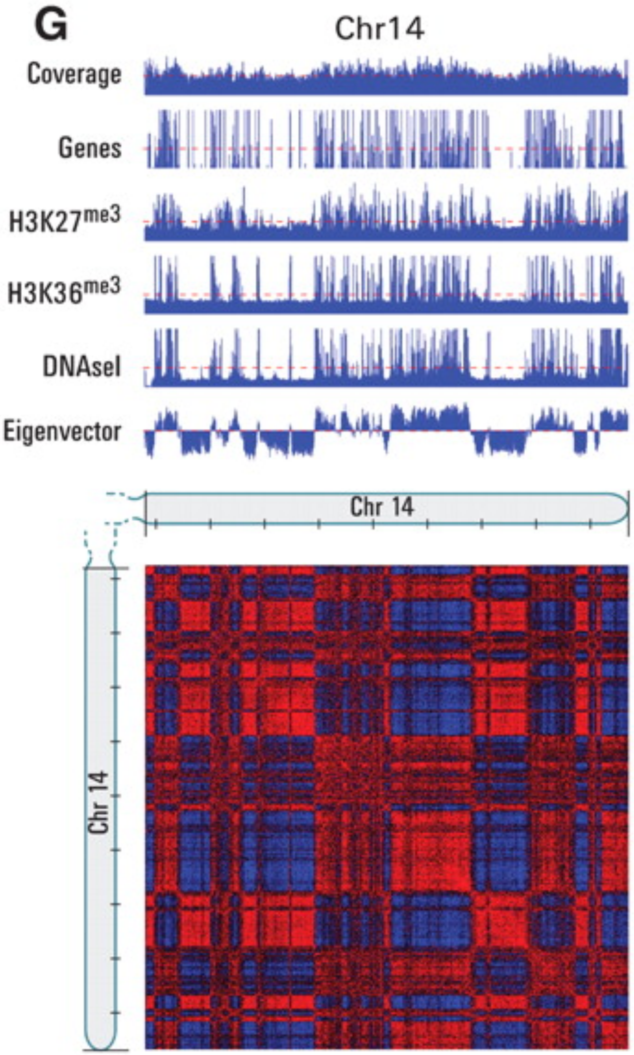
(C) Correlation matrix: entry *ij* = cor(row *i*, column *j*), from -1 (blue) to +1 (red)

- The pattern indicates two compartments within the chromosome
- Contacts within each compartment enriched and contacts between depleted

Lieberman-Aiden et al. 2009

The less packed compartment correlates with active DNA

- Less packed: more contacts (red)

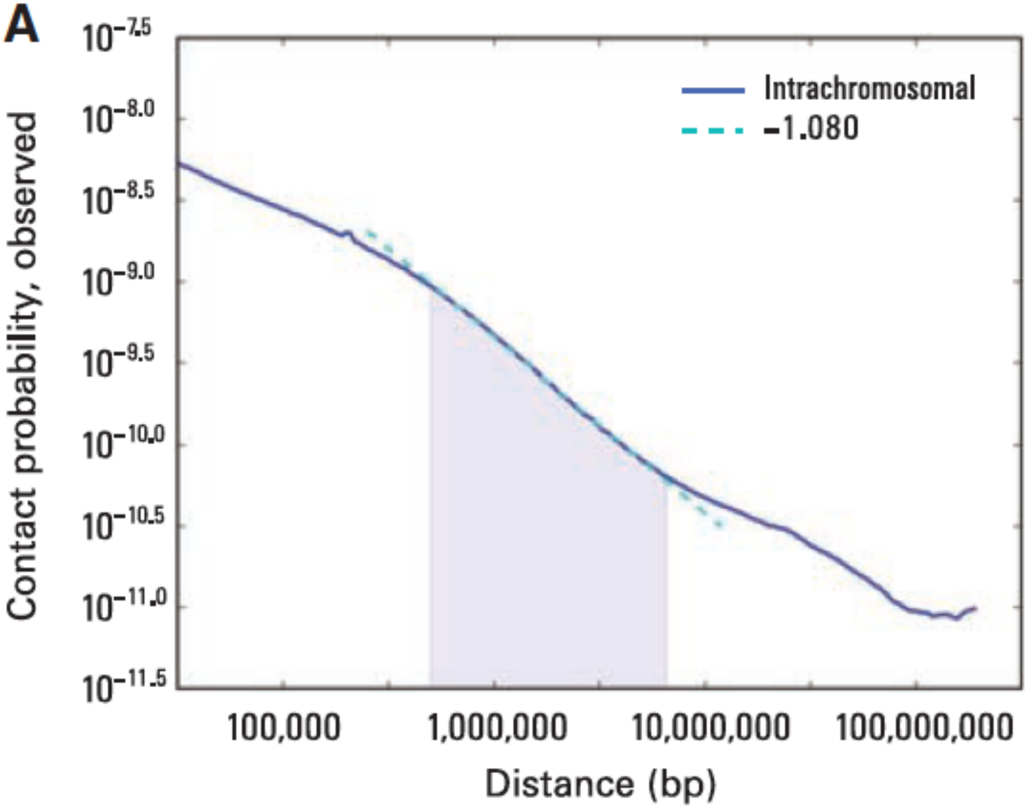


Lieberman-Aiden et al. 2009

Intrachromosomal contact prob *I* as a function of distance *s*

- Power law relation: $y = a\,x^{k}$
- Plotting power law on log – log scale gives a line: $Y = -k\,X + b$, where $Y = log(y)$, $X = log(x)$, $b=log(a)$

$I(s)$ plotted on log-log scale shows power law distribution with $k = -1$, $I(s) = s^{-1}$,between 500 kb and 7 Mb



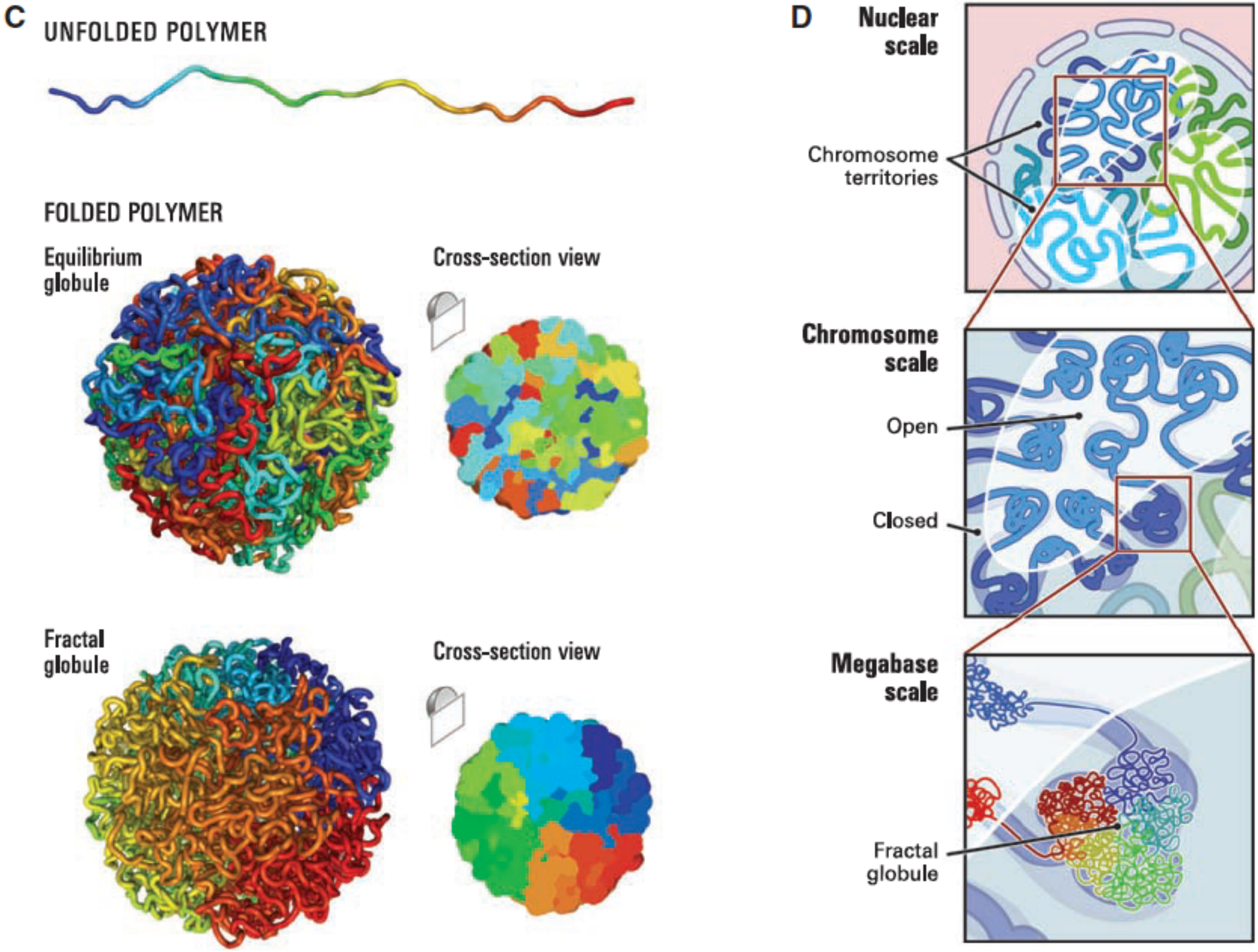
Lieberman-Aiden et al. 2009

Different models of chromatin organization

- DNA is a polymer: a large molecule composed of many repeated subunits.
- **“**equilibrium globule**”:** a compact configuration originally used to describe a polymer in a poor solvent at equilibrium. They
  - are highly knotted
  - have linear and spatial positions largely decorrelated after a few megabases
  - predict that contact probability will scale as $s^{-3/2}$
- **“**fractal globule”: highly compact, globule-of- globules-of-globules that densely fills 3D space without crossing itself. They:
  - lack knots
  - facilitate unfolding and refolding, e.g, during gene activation
  - contiguous regions of the genome form spatial sectors whose size corresponds to the length of the original region
  - predict that contact probability will scale as s^{-1}

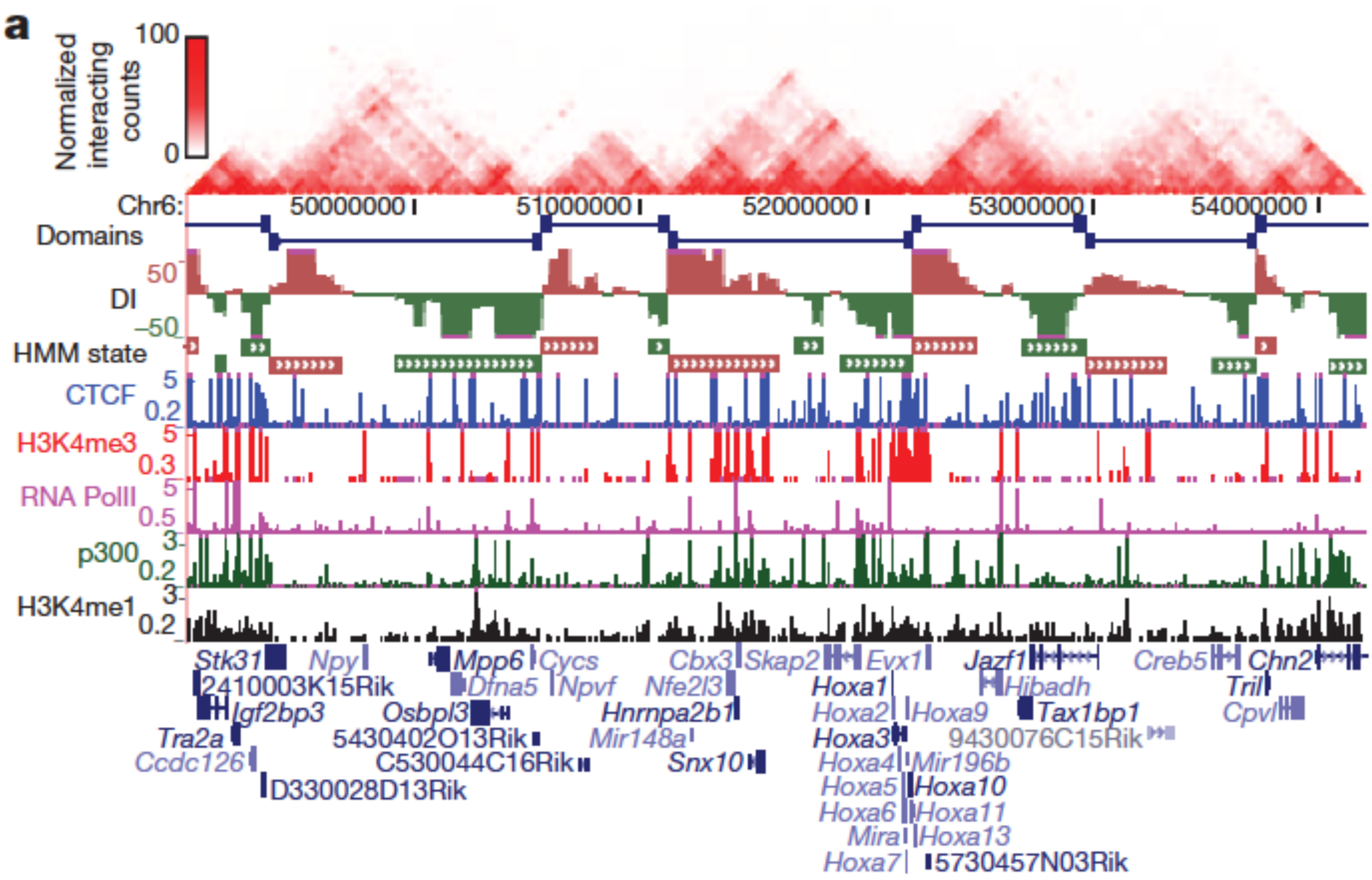
* Poor solvent: one in which the solute precipitates
Lieberman-Aiden et al. 2009

Chromatin is a fractal globule



Lieberman-Aiden et al. 2009

Topological association domains (TADs)



Dixon et al. 2012

Directionality index

- A : number of reads that map from a given 40kb bin to the upstream 2Mb (upstream mapping bias)
- B : no. of reads that map from the same 40kb bin to the downstream 2Mb (downstream mapping bias)
- E = (A + B)/2 (average of A and B)

$$DI = \left(\frac{B - A}{|B - A|}\right)\left(\frac{(A - E)^2}{E} + \frac{(B - E)^2}{E}\right)$$

- Useful to detect boundaries of TADs: more biased bins have a higher magnitude of DI.
- A HMM model to infer the “true” biases in the data

Dixon et al. 2012

Markov chain

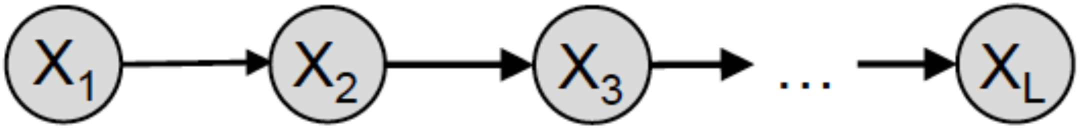
- Let $\{X_1, \ldots, X_L\}$ be discrete r. v. with common state space [K] = {1, …, K}.
- We always have the factorization

$$\begin{aligned} P(x_1, \ldots, x_L) &= P(x_1, \ldots, x_{L-1}) P(x_L \mid x_{L-1}, \ldots, x_1) \\ &= P(x_1, \ldots, x_{L-2}) P(x_{L-1} \mid x_{L-2}, \ldots, x_1) P(x_L \mid x_{L-1}, \ldots, x_1) \\ & \quad \ldots \\ &= P(x_1) P(x_2 \mid x_1) P(x_3 \mid x_2, x_1) \ldots P(x_L \mid x_{L-1}, \ldots, x_1) \end{aligned}$$

- $\{X_n\}$ is a Markov chain if the **Markov property** holds, i.e., if

$$P(X_n \mid X_{n-1}, \ldots, X_1) = P(X_n \mid X_{n-1})$$

for all n = 2, …, L.



$$X_{n+1} \perp X_{n-1} \mid X_n$$

Transition matrix

- A Markov chain $\{X_n\}$ is homogeneous, if

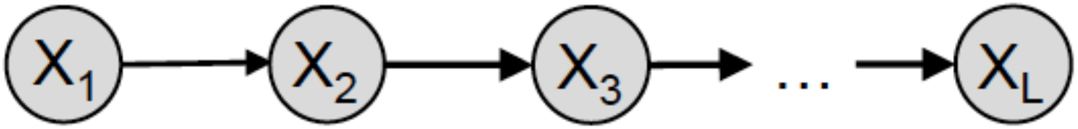
$$P(X_n \mid X_{n-1}) = P(X_2 \mid X_1) \quad \text{for all } n \geq 2$$

- A homogeneous Markov chain is determined by
  - the initial state distribution $\Pi \in \Delta_{K-1}$ defined by

$$\Pi_k = P(X_1 = k)$$

  - and the K × K transition matrix $T = (T_{kl})$ given by

$$T_{kl} = P(X_{n+1} = l \mid X_n = k)$$

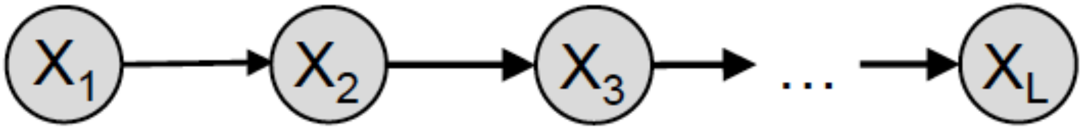


Markov chain model

- The probability of an observation x = (x_1, …, x_L) in the Markov chain model MC(Π, T) is

$$P(X = x) \;=\; P(X_1 = x_1) \prod_{n=1}^{L-1} P(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

$$\;=\; \Pi_{x_1} \prod_{n=1}^{L-1} T_{x_n, x_{n+1}}$$

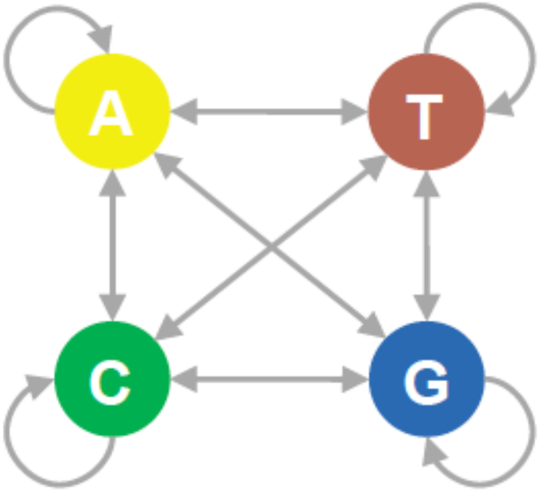


HMM for Hi C

- Hidden states: “Upstream Bias”, “Downstream Bias” or “No Bias”
- Y = {Y1,…,Yn} : observed directionality index, modeled as mixtures of Gaussians
- Q = {Q1,…,Qn} : the true hidden directionality biases
- M={M1,…,Mn}: mixtures
- $P(Y_t = y_t \mid Q_t = i, M_t = m) = N(y_t;\mu_{i,m}, \Sigma_{i,m})$
- $P(M_t =m \mid Q_t = i) = C(i,m)$,
  where C encodes the mixture weights for each state i.
- Baum-Welch algorithm [EM] to compute maximum likelihood estimates

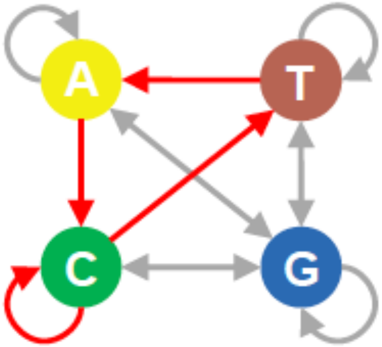
Dixon et al. 2012

DNA example



$$\Pi = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{pmatrix} .3 \\ .4 \\ .2 \\ .1 \end{pmatrix} \qquad T = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & .3 & .1 & .3 & .3 \\ C & .4 & .1 & .1 & .4 \\ G & .3 & .2 & .2 & .3 \\ T & .3 & .2 & .1 & .4 \end{array}$$

- We consider DNA sequences $x \in \{A,C,G,T\}^*$ as observations of a homogeneous Markov chain $\{X_i\}$.
- For example, P(ACCTA) = 0.3 · 0.1 · 0.1 · 0.4 · 0.3



CpG islands

- CpG islands are stretches of mammalian genomes enriched for the dinucleotide CG, typically 300 to 3,000 bases long.

- CG tends to mutate to CT, so in general $P(CG) < P(C)P(G)$

- But in promoter regions, this effect is suppressed and hence CpG islands are more common.

How can we find CpG islands in a genome?

...ACTTCGCGCGCCGATGCCACTGCACATGCATGCATCGCGCCGCGCGACAGACTTACG...

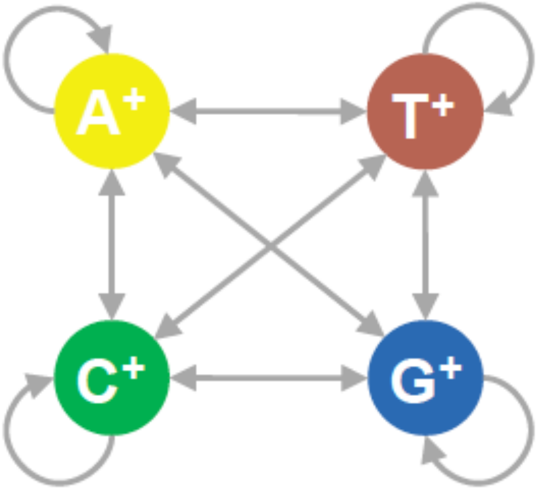
Annotating genomic sequences

```
...----+++++++++--------------------+++++++++++++-----------...
...ACTTCGCGCGCCGATGCCACTGCACATGCATGCATCGCGCCGCGCGACAGACTTACG...
```

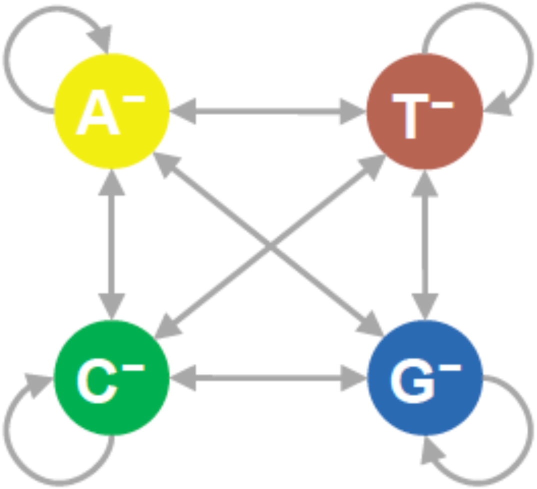
Two Markov chain models

...----+++++++++----------------------+++++++++++++-----------...

...ACTTCGCGCGCCGATGCCACTGCACATGCATGCATCGCGCGCCGCGCGACAGACTTACG...



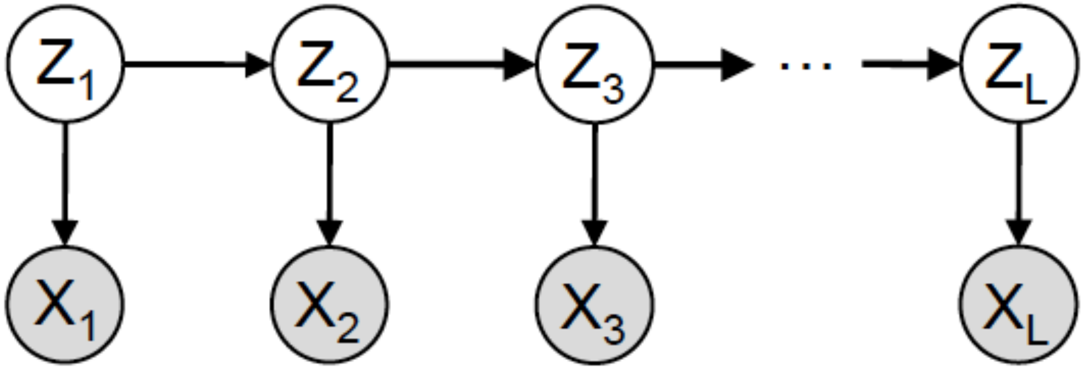
CpG island



Non-CpG island

Hidden Markov model (HMM)

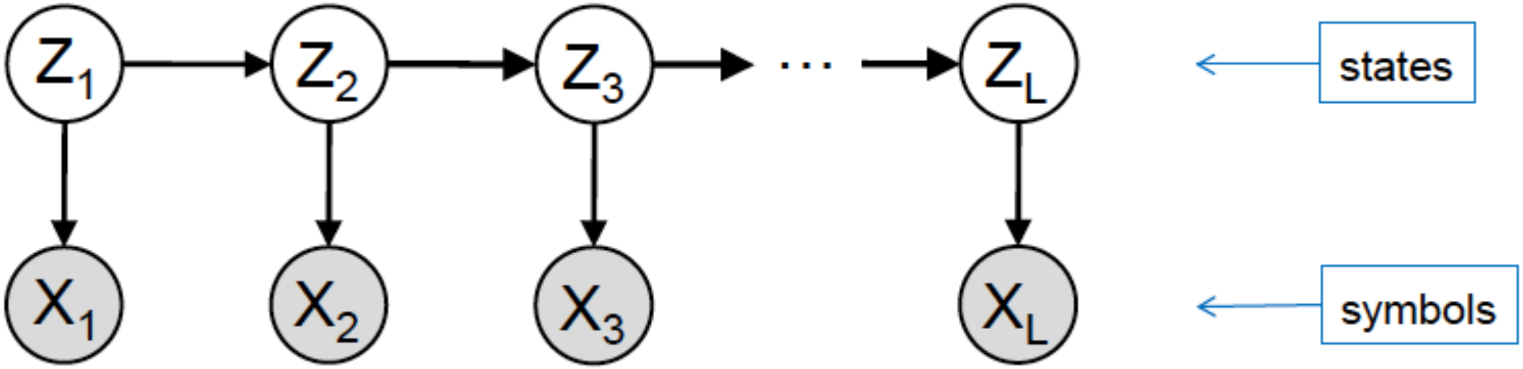
- Hidden (non-observable) random variables $\{Z_n\}$ form a homogeneous Markov chain (the annotation).
  - For example, Z_n indicates whether sequence position n belongs to a CpG island or not, $Z_n \in \{+, -\}$.
- Observed random variables $X_n \in \{A,C,G,T\}$ result from hidden states emitting symbols.



$$Z_{n+1} \perp Z_{n-1} \mid Z_n$$

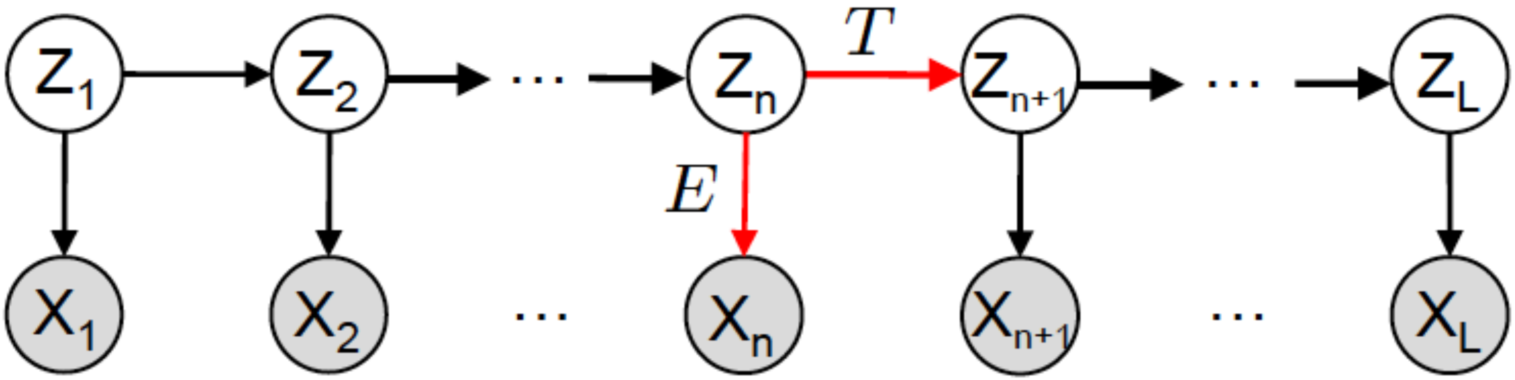
$$X_n \perp X_m \mid Z_n, \; m \neq n$$

Definitions



- Initial state probabilities: $\Pi_k = P(Z_1 = k)$
- Transition probabilities: $T_{kl} = P(Z_n = l \mid Z_{n-1} = k)$
- Emission probabilities: $E_{kx} = P(X_n = x \mid Z_n = k)$

Joint probability

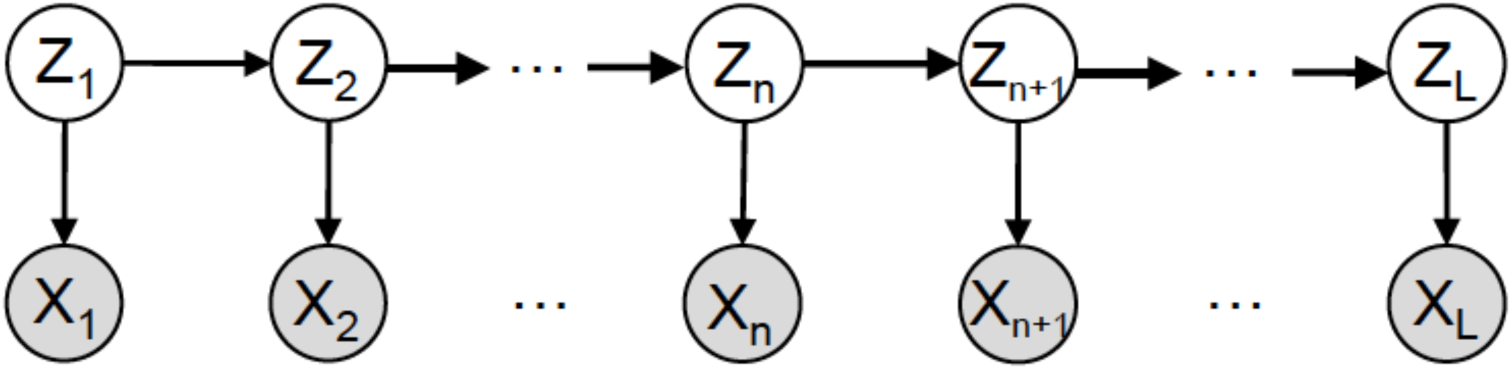


$$P(X,Z) = P(Z_1) \prod_{n=1}^{L} P(X_n \mid Z_n) P(Z_{n+1} \mid Z_n)$$

$$= \Pi_{Z_1} \prod_{n=1}^{L} E_{Z_n, X_n} T_{Z_n, Z_{n+1}}$$

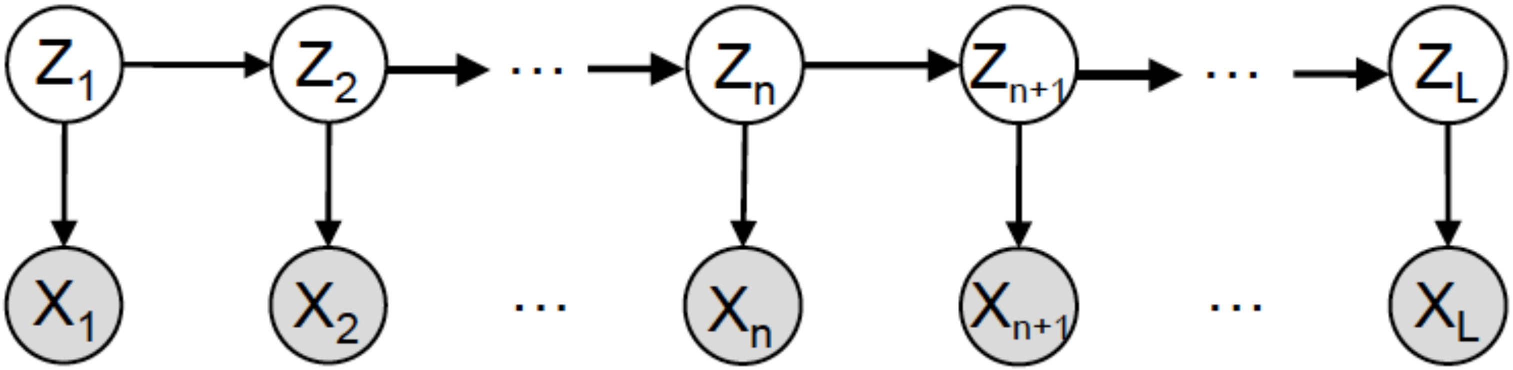
where $P(Z_{L+1} \mid Z_L) = T_{Z_L, Z_{L+1}} \equiv 1$

State path



- We observe the DNA sequence X, but we are interested in the hidden states Z of the Markov chain (the *annotation*).
- Each **z** = (z_1, …, z_L) is called a state path. There are K^L possible paths, where K is the number of (hidden) states.
- Different state path can give rise to the same sequence of observed symbols, but with different probabilities.

Decoding

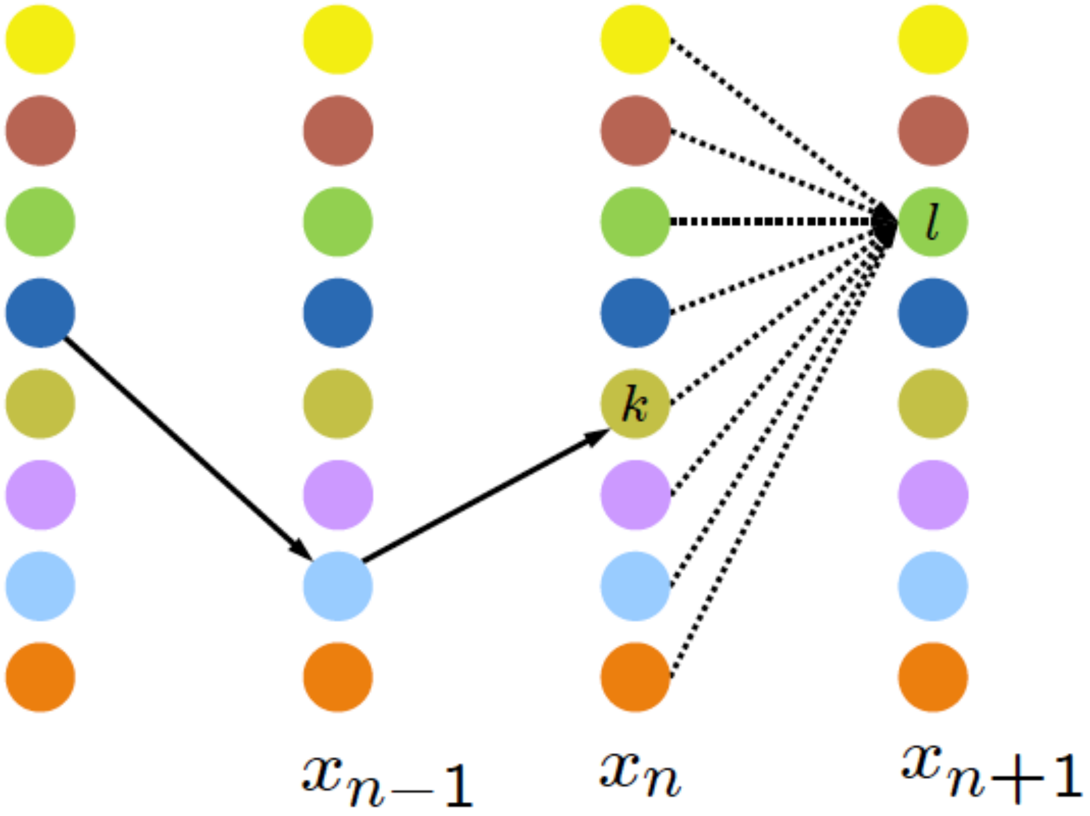


- For given parameters, the decoding problem is to find the most probable state path **z** for a given observation x:

$$z^* = \underset{z}{\operatorname{argmax}}\, P(X = x, Z = z)$$

Viterbi algorithm: basic idea

- Define $v_k(n)$ as the probability of z* ending in state k with observation x_n
- If $v_k(n)$ is known for all states k, then $v_l(n+1)$ is obtained by maximizing over all states:



$$v_l(n+1) = E_{l,x_{n+1}} \max_k v_k(n) T_{kl}$$

Viterbi algorithm

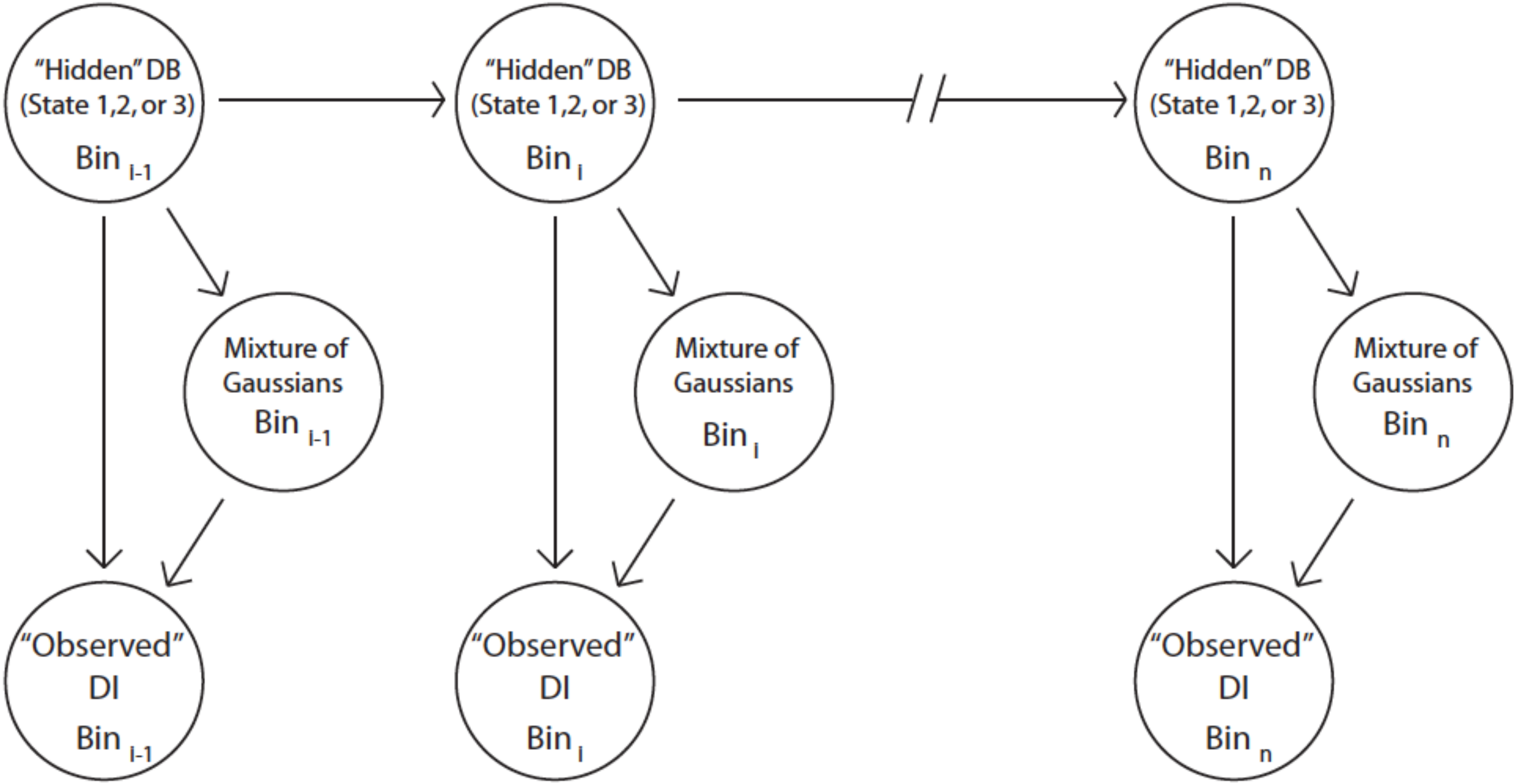
- Initialization:
  - $v_0(0) = 1$
  - $v_k(0) = 0$ for all $k > 1$
- Recursion: for $n = 1, \ldots, L$,
  - $v_l(n) = E_{lx_n} \max_k v_k(n-1)T_{kl}$ for all $l = 1, \ldots, K$
  - $ptr_n(l) = \text{argmax}_k \, v_k(n-1)T_{kl}$ for all $l = 1, \ldots, K$
- Termination (assuming an end state):
  - $P(x, z^*) = \max_k v_k(L)T_{k0}$
  - $z^*_L = \text{argmax}_k \, v_k(L)T_{k0}$
- Traceback: for $n = L, \ldots, 1$,
  - $z^*_{n-1} = ptr_n(z^*_n)$

- Dynamic programming, $O(LK^2)$ despite K^L paths!

Summary

- Markov chains can model temporal or spatial (linear) dependencies.
- HMMs consist of a hidden state space with a Markov chain structure emitting observable symbols.
- HMMs are frequently used for genome annotation, for example, CpG islands, gene finding, etc.
- The Viterbi algorithm computes the most probable state path and the forward and backward algorithms the likelihood in an efficient way.
- Parameter estimation can be performed using the EM algorithm (Baum-Welch algorithm).

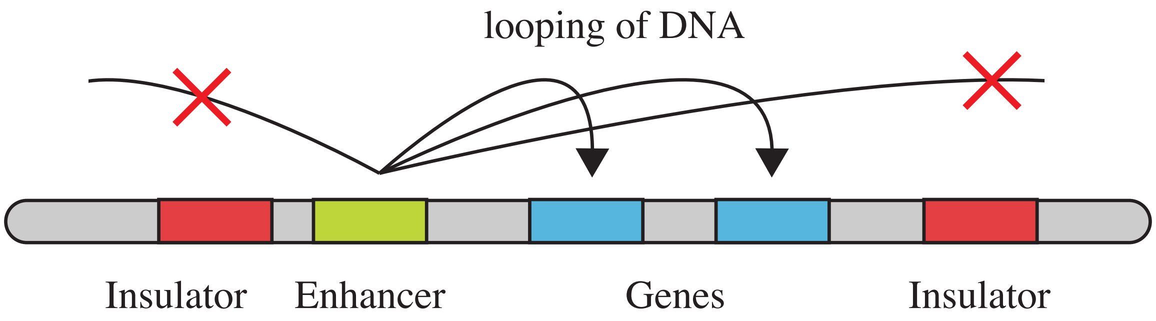
HMM for Hi C

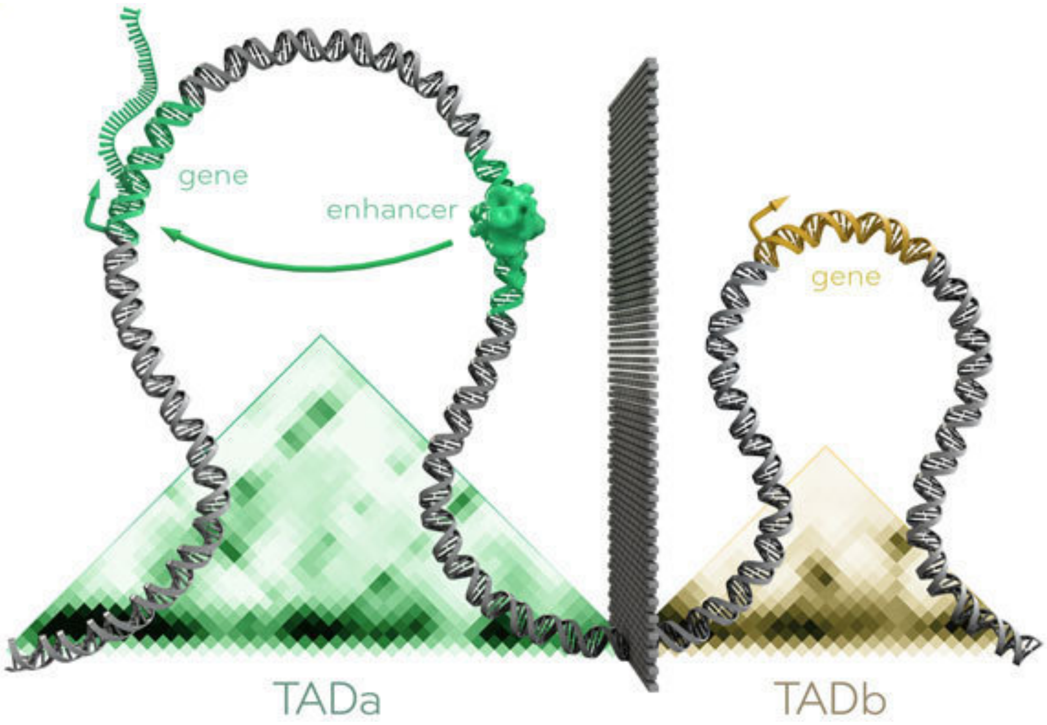

"Hidden" DB (State 1,2, or 3) Bin $_{i-1}$
"Hidden" DB (State 1,2, or 3) Bin $_{i}$
"Hidden" DB (State 1,2, or 3) Bin $_{n}$
Mixture of Gaussians Bin $_{i-1}$
Mixture of Gaussians Bin $_{i}$
Mixture of Gaussians Bin $_{n}$
"Observed" DI Bin $_{i-1}$
"Observed" DI Bin $_{i}$
"Observed" DI Bin $_{n}$

Dixon et al. 2012

Boundaries of TADs ~ insulator (barrier) elements

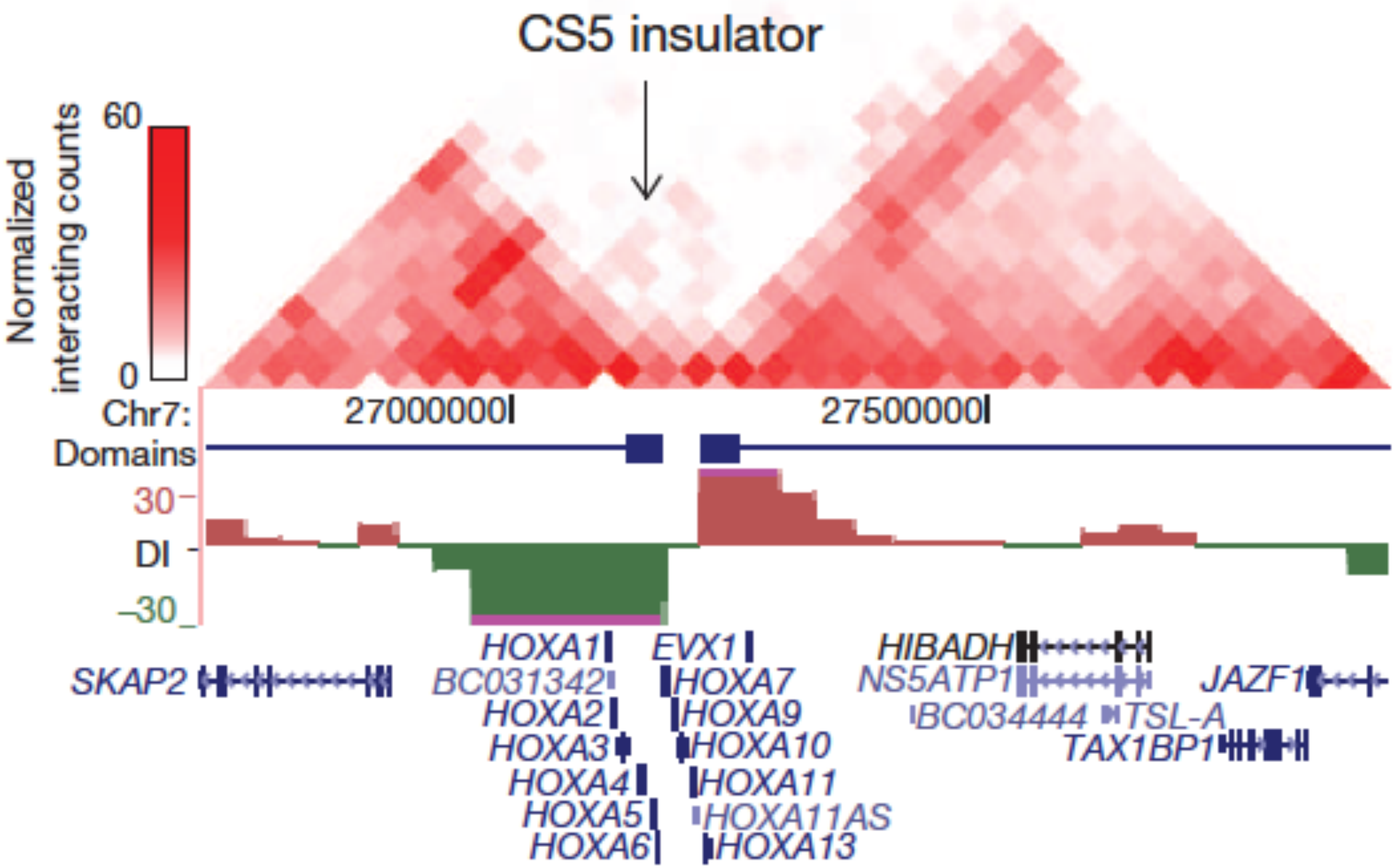
- Insulator: genetic boundary element that blocks the interaction between enhancers and promoters.





Boundaries of TADs ~ insulator (barrier) elements

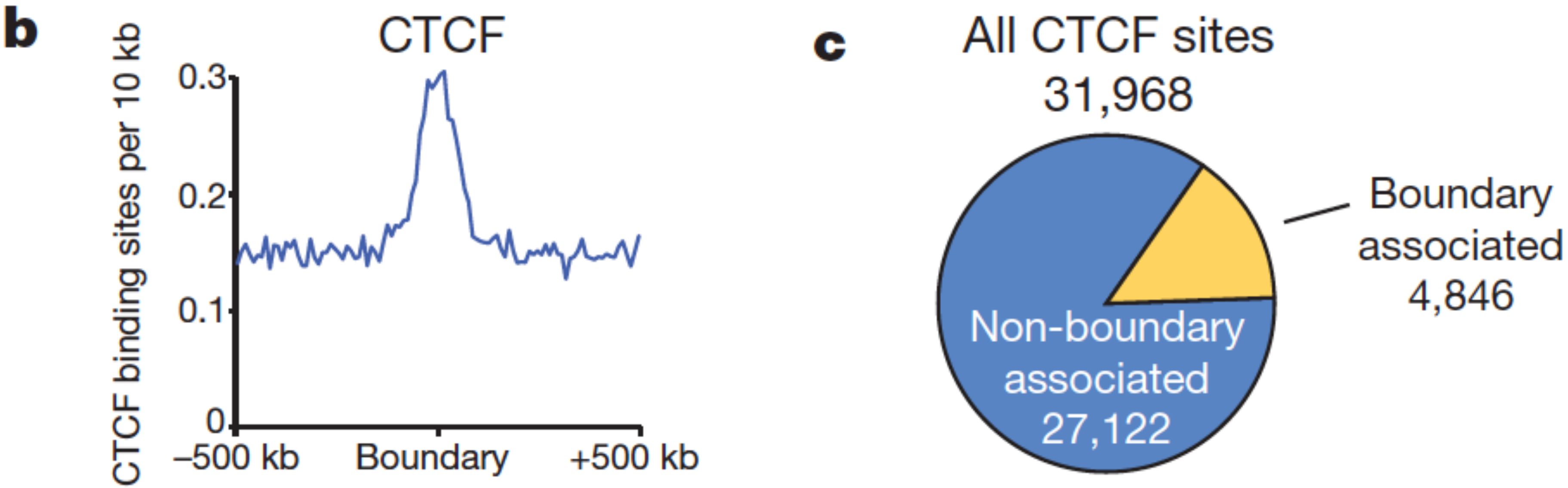
- Insulator: genetic boundary element that blocks the interaction between enhancers and promoters.
- Eg. The Hoxa locus



Dixon et al. 2012

Boundaries of TADs ~ insulator (barrier) elements

- Many known insulator or barrier elements bound by the zincfinger-containing protein CTCF
- Strong enrichment of CTCF at the topological boundary regions
- CTCF binds also outside of the boundary regions
- How to show enrichment?

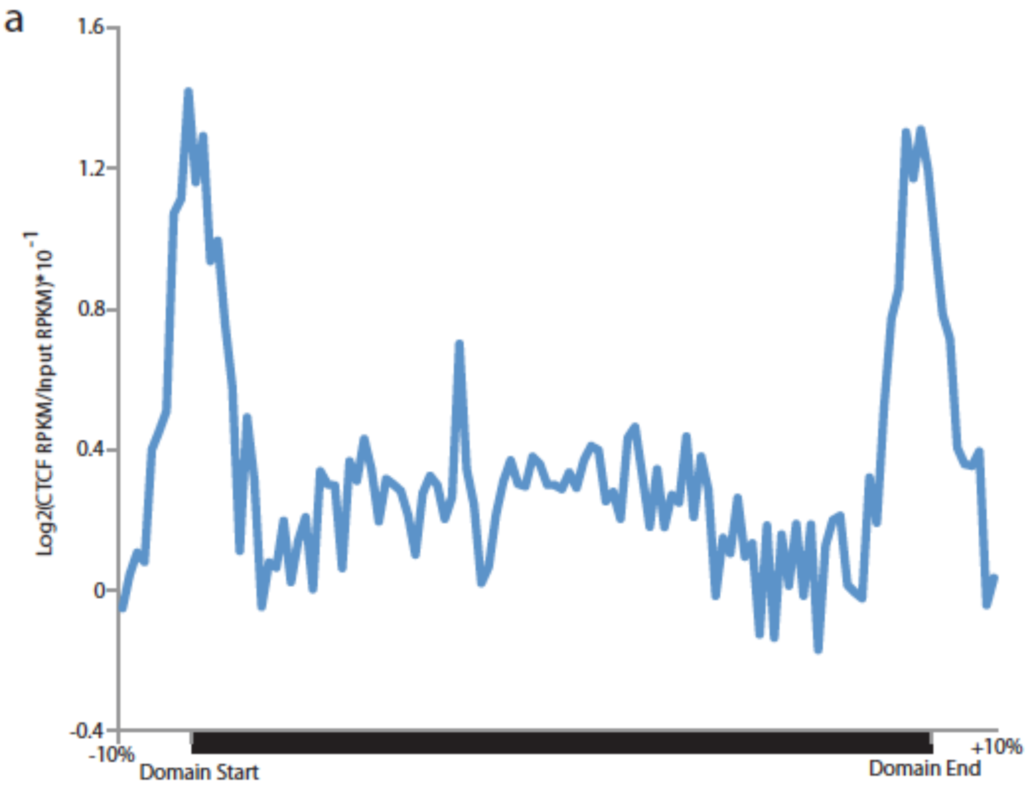


Dixon et al. 2012

CTCF enrichment at topological boundary regions

Average enrichment plot of CTCF over topological domains.

- Each TAD divided into 100 equal size bins (+/- 10 bins from each end of the domain).
- log2 ratio of CTCF RPKM over Input (control) calculated for each bin, shown as an average over TADs.
- CTCF enriched on the edges.

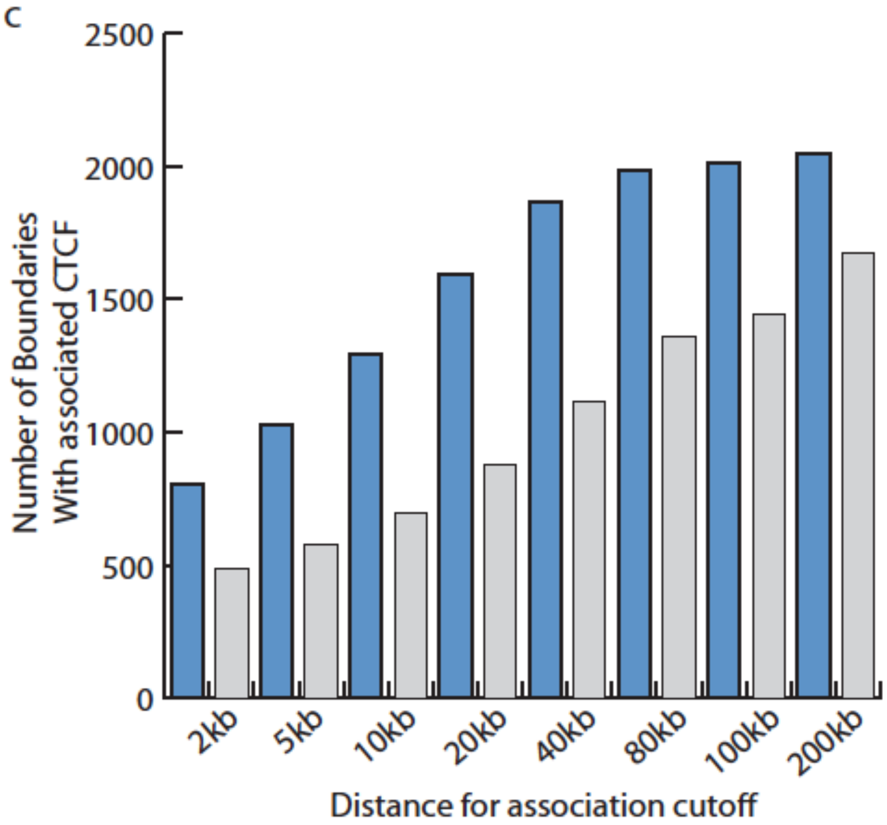


Dixon et al. 2012

CTCF enrichment at topological boundary regions

Number of boundaries with an associated CTCF site for varying window size cut offs.

- Blue: For each distance D, the number of boundaries with a CTCF within +/- D.
- Gray: the number expected at random at the same distance cut-off.

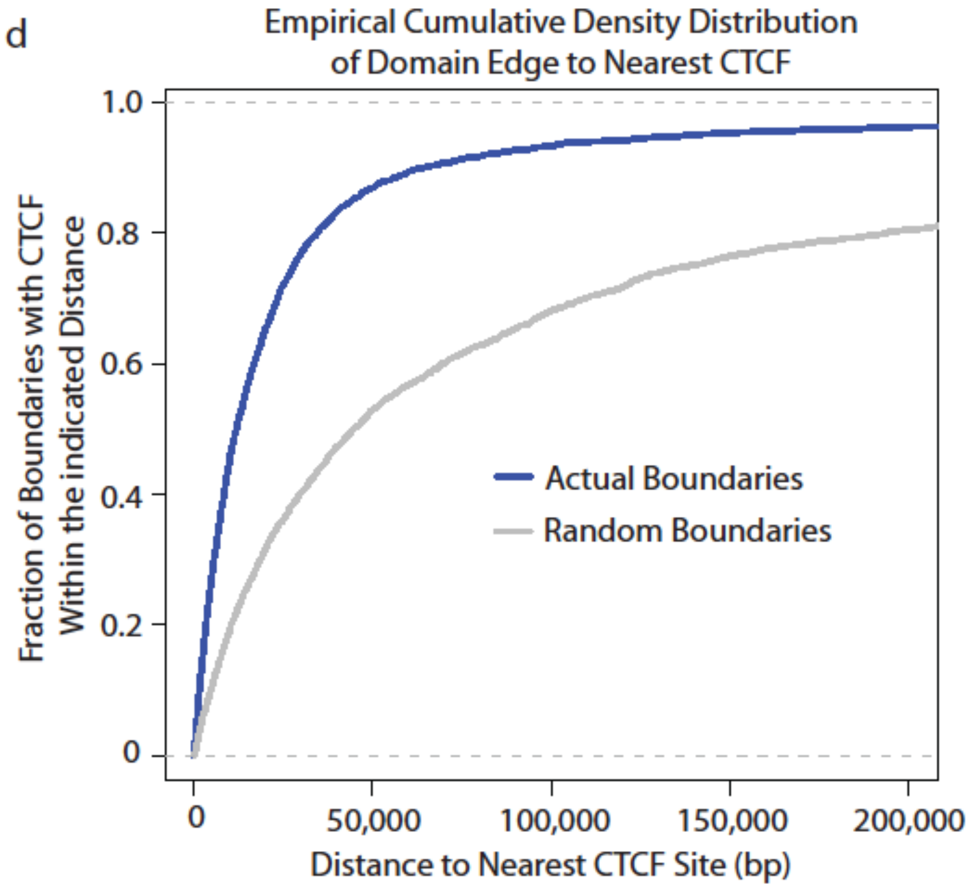


Dixon et al. 2012

CTCF enrichment at topological boundary regions

The empirical cumulative density distribution of the distance between the domain border and the nearest CTCF binding site (in bp).

- Blue: The distance between the actual boundaries and the nearest CTCF site
- Gray: The distance to randomized boundaries



Dixon et al. 2012

Bibliography

- Erez Lieberman-Aiden et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome*. Science 9 October 2009: Vol. 326 no. 5950 pp. 289-293*.

- Neph et al. *An expansive human cis-regulatory lexicon encoded in transcription factor footprints.* Nature 489:83-90, 2012

- Piper et al. *Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data.* Nucleic Acids Research, 2013, Vol. 41, No. 21 e201

- Boyle et al. *F-Seq: a feature density estimator for high-throughput sequence tags.* Bioinformatics Vol. 24 no. 21 2008, pages 2537–2538