Histone Marks
Chromatin 3D Structure
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
Chromatin and Nucleosome Organization

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.
2. Learning chromatin states
Can we find latent state to explain observed marks?
Hidden Markov Models

Hidden state $x$ in $[1..m]$
  
  For example, $m$ can be 15

Emitted symbol $y$ can be multi dimensional
  
  For example, histone and accessibility data at genomic loci

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

Transition parameters

\[ P(y_t \mid x_t) \]

\[ P(x_{t+1} \mid x_t) \]
ChromHMM segment based chromatin states
Tissues and cell types profiled in the Roadmap Epigenomics Consortium.
Chromatin state annotations in 127 epigenomes

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

*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

iMN: Lhx3 ChIP-seq

iMN: Pol II ChIA-PET

pMN: Olig2 ChIP-seq

pMN: Pol II ChIA-PET

mES: Sox2 ChIP-seq

mES: Pol II ChIA-PET

Sox2
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

<table>
<thead>
<tr>
<th>Factor</th>
<th>Cells (million)</th>
<th>Reads (million)</th>
<th>Percent of total reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smc1a</td>
<td>100</td>
<td>398</td>
<td></td>
</tr>
<tr>
<td>Pol II</td>
<td>100</td>
<td>255</td>
<td></td>
</tr>
<tr>
<td>Rad21</td>
<td>100</td>
<td>287</td>
<td></td>
</tr>
<tr>
<td>Rad21</td>
<td>100</td>
<td>394</td>
<td></td>
</tr>
<tr>
<td>CTCF</td>
<td>100</td>
<td>682</td>
<td></td>
</tr>
<tr>
<td>CTCF</td>
<td>100</td>
<td>531</td>
<td></td>
</tr>
<tr>
<td>Pol II</td>
<td>100</td>
<td>244</td>
<td></td>
</tr>
</tbody>
</table>

ChIA-PET

<table>
<thead>
<tr>
<th>Factor</th>
<th>Cells (million)</th>
<th>Reads (million)</th>
<th>Percent of total reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smc1a (Hs)</td>
<td>25</td>
<td>377</td>
<td></td>
</tr>
<tr>
<td>Smc1a (Mm)</td>
<td>25</td>
<td>274</td>
<td></td>
</tr>
<tr>
<td>Smc1a (Mm)</td>
<td>10</td>
<td>177</td>
<td></td>
</tr>
<tr>
<td>Smc1a (Mm)</td>
<td>5</td>
<td>247</td>
<td></td>
</tr>
<tr>
<td>Smc1a (Mm)</td>
<td>1</td>
<td>118</td>
<td></td>
</tr>
</tbody>
</table>
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?

\[
\begin{align*}
N \text{ total ends} \\
I_{a,b} \text{ interactions observed} \\
c_a \text{ ends} & \quad c_b \text{ ends}
\end{align*}
\]
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

iMN: Lhx3 ChIP-seq

iMN: Pol II ChIA-PET
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 ($k$):

$$\hat{N} = \arg\max_{N} [P(X = k; N, m, n)]$$

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:

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

$$\approx \text{Normal} \left( X = k; \mu = \frac{mn}{N}, \sigma^2 = \frac{mn}{N} \left( 1 - \frac{m}{N} \right) \right)$$
5b. Discovering interactions: Density-based
Method 2: CID uses density-based clustering to discover chromatin interactions

**Figure 1.** CID uses density-based clustering to discover chromatin interactions. (A) ChIA-PET read pairs. (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 labeled. (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

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.

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

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
Cluster interaction origins

A

POLR2A ChIA-PET reads pairs

CID interaction calls

ChIA-PET2 interaction calls

ChIA-PET2 peak calls

Mango interaction calls

Mango peak calls

POLR2A ChIA-PET ChIP

DNase-seq

CTCF ChIP-seq
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
TargetFinder Training Data

<table>
<thead>
<tr>
<th>Cell Line</th>
<th>Promoters</th>
<th>Enhancers</th>
<th>Interacting Pairs</th>
<th>Non-Interacting Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>K562</td>
<td>8196</td>
<td>82806</td>
<td>1977</td>
<td>39500</td>
</tr>
<tr>
<td>GM12878</td>
<td>8453</td>
<td>100036</td>
<td>2113</td>
<td>42200</td>
</tr>
<tr>
<td>HeLa-S3</td>
<td>7794</td>
<td>103460</td>
<td>1740</td>
<td>34800</td>
</tr>
<tr>
<td>HUVEC</td>
<td>8180</td>
<td>65358</td>
<td>1524</td>
<td>30400</td>
</tr>
<tr>
<td>IMR90</td>
<td>5253</td>
<td>108996</td>
<td>1254</td>
<td>25000</td>
</tr>
<tr>
<td>NHEK</td>
<td>5254</td>
<td>144302</td>
<td>1291</td>
<td>25600</td>
</tr>
</tbody>
</table>
TargetFinder – Ratio of the CTCF and RAD21 ChIP-seq signals occurring within interacting enhancers and non-interacting enhancers

![Graph showing co-occurring signal for CTCF and RAD21 with peak anchor annotations.](image)
TargetFinder – Enrichment of signals at transcription start sites (TSS)

Enhancer–promoter pairs

- Non-interacting
- Interacting

Dark – interacting; Light – non-interacting
Features for enhancers and promoters only (E/P), extended enhancers and promoters (EE/P), and enhancers and promoters plus the windows between them.
Deep learning network for predicting enhancer-promoter interactions

Diagram:
- Promoter Module
  - Sequence Module
  - Chromatin Module
- Enhancer Module
  - Sequence Module
  - Chromatin Module
- Fully-connected layer
- Interaction Network
- Interacting
- Non-interacting
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