6.874 Lecture 11
Chromatin modeling

6.874 Spring 2011
Because of circularization, self-ligation PETs have the opposite strandedness of ChIP-Seq reads.
PETs of different strand orientations have the following distance distributions.
Upregulated

H3K4me3

RNA Pol II

ER ChIP

Foa1 ChIP

Downregulated

FOS

JDP2

BATF

Upregulated
Intra-chromosomal results

\[ C_{ij} = \text{Correlation}(B \text{ Row}_i, B \text{ Column}_j) \]
Intra-chromosomal results
Inter-chromosomal results
FOLDED POLYMER

Equilibrium globule

Cross-section view

Fractal globule

Cross-section view
Chromatin and Nucleosome Organization


Nucleaseosome 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
Histone Tail Modifications

Sims III et al., 2003
The ChIA-PET procedure
Classifying PETs as having resulted from self-ligation vs. inter-ligation

- Both ends map within a short distance of each other and have ‘-+’ strand orientation
- Both ends map distally or have a non ‘-+’ strand orientation
The significance of self-ligation event overlap can be estimated by Monte Carlo Simulation

- A conservative estimate of the size of the mappable genome is the number of base pairs covered by at least one self-ligation event
- Repeat N times:
  - randomly generate fragments from the observed fragment length distribution
- The number of times k or more fragments are observed overlapping during the simulation can be used to estimate the significance of observing k or more overlapping fragments under the null hypothesis that there is no binding event
The significance of observing $I$ inter-ligation events between two binding events $A$ and $B$ can be calculated using the Hypergeometric Test.

Let $I_{A,B}$ be the number of inter-ligation events between binding events $A$ and $B$. Let $c_A$ and $c_B$ be the number of ligation event ends associated with $A$ and $B$, respectively. Let $N$ be the total number of ligation events ends. The null hypothesis assumes that each ligation event end has an equal probability of ligating with any other end. Then, under the null hypothesis:

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

$$p = \min\{c_A, c_B\} \sum_{i=I_{A,B}} P(i|N, c_A, c_B)$$
Estimating the False Discovery Rate by random permutation

• By randomly permuting the ends of the inter-ligation events the false discovery rate can be estimated from the number of pairs of binding events with associated p-value under some threshold $T$
Statistical modeling of overlap between event sets

6.874 Spring 2011
Matt Edwards
matted@mit.edu
Enrichment testing

- We want to test significance of overlap between two sets of items
- *Example*: a set of genes implicated in an experiment and a set of genes involved in a particular biological process
- Are the experimentally-identified genes significantly related to the particular biological process?

**Diagram:**

- Query set (via experiments)
- [Possibly] interesting overlap
- Target set (known annotation)
Statistical model

• Model the query set as samples taken without replacement from some larger set (all genes)
• What is the probability that in \( n \) samples we see \( k \) from our target set?
• Depends on how many targets there are (\( m \)) and how many total genes we have (\( N \))
• Hypergeometric distribution:

\[
P(X = k; N, m, n) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}
\]
Significance testing

• Using the hypergeometric distribution, we can calculate $p$-values for overlap sizes for a target set:

$$p = P(X \geq k; N, m, n)$$

• *Caveats*: uncertainty over which genes we could ever recover (value of $N$), multiple hypothesis testing and the complex relationships between multiple target sets
Estimating total events

• We can use this model to infer how many possible items exist \((N)\) given two sample sizes \((m\) and \(n)\) and an overlap \((k)\):

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

• Example: We perform two replicates of an experiment and obtain 1000 hits (events) in each, of which 900 are shared

• Can we estimate the total number of events?
• Using this model, we predict \(\sim1100\) total events
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 model a false positive rate $f$ by assuming that overlapping (shared) events are true positives and that a fraction $f$ of the remaining events are false positives.
- This approximation lets us update $m$ and $n$ and apply the same model:

\[
m' = (1 - f)(m - k) + k \\
n' = (1 - f)(n - k) + k
\]
Example

- 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):

![Likelihood plots versus $N$ for several true positive rates (TPR)](image)
Approximate closed form solution

- 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:

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

$$\approx \text{Normal} \left( X = k; \mu = \frac{mn}{N}, \sigma^2 = \frac{mn}{N} \left( 1 - \frac{m}{N} \right) \right)$$
### Table 1 | Summary statistics of library PET sequences

<table>
<thead>
<tr>
<th>Library code</th>
<th>Library identity</th>
<th>Total PET</th>
<th>Unique PET</th>
<th>Self-ligation</th>
<th>Intrachromosome inter-ligation</th>
<th>Interchromosome inter-ligation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PET</td>
<td>PET clusters*</td>
<td>PET</td>
</tr>
<tr>
<td><strong>Small-scale testing of the ChIA-PET method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IHM001N</td>
<td>ChIA-PET</td>
<td>715,369</td>
<td>271,648</td>
<td>78,706</td>
<td>2,701</td>
<td>16,677</td>
</tr>
<tr>
<td>IHM001H</td>
<td>ChIA-PET</td>
<td>764,899</td>
<td>293,754</td>
<td>103,740</td>
<td>3,405</td>
<td>17,718</td>
</tr>
<tr>
<td>IHM043</td>
<td>ChIP-PET</td>
<td>1,118,509</td>
<td>745,251</td>
<td>634,993</td>
<td>1,158</td>
<td>7,386</td>
</tr>
<tr>
<td>SHC007</td>
<td>ChIP-PET</td>
<td>361,241</td>
<td>214,668</td>
<td>192,511</td>
<td>489</td>
<td>2,196</td>
</tr>
<tr>
<td>IHM062</td>
<td>ChIA-PET (IgG)</td>
<td>436,248</td>
<td>217,708</td>
<td>40,847</td>
<td>0</td>
<td>11,254</td>
</tr>
<tr>
<td><strong>Analysis of chimaeras</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IHH015M</td>
<td>ChIA-PET</td>
<td>4,246,429</td>
<td>2,049,719</td>
<td>953,384</td>
<td>3,909</td>
<td>129,492</td>
</tr>
<tr>
<td>(AA + BB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IHH015C</td>
<td>ChIA-PET</td>
<td>5,904,476</td>
<td>1,790,714</td>
<td>15,490</td>
<td>35</td>
<td>98,805</td>
</tr>
<tr>
<td>(chimaeras)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Large-scale ChIA-PET analysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IHM001F</td>
<td>ChIA-PET</td>
<td>31,828,194</td>
<td>4,638,633</td>
<td>1,249,081</td>
<td>14,560</td>
<td>234,400</td>
</tr>
<tr>
<td>IHH015F</td>
<td>ChIA-PET</td>
<td>19,590,581</td>
<td>6,125,099</td>
<td>1,841,684</td>
<td>6,665</td>
<td>348,057</td>
</tr>
</tbody>
</table>

CHA-PET data mapped at satellites and structural variation sites were removed.

* Self-ligation PET clusters for identifying binding sites (FDR < 0.01, PET count at least 5).

† Inter-ligation PET clusters for identifying interactions include at least two (small-scale) or three (chimaeras and large-scale analysis) overlapping PETs (FDR < 0.05). Interchromosomal interactions were subjected to manual curation.

‡ One interaction has a genomic span of less than 5 kb, suggesting that it results from extra-long self-ligation PETs, and the other has a genomic span of more than 10 Mb and PET counts of only 2, and so could be non-specific.
<table>
<thead>
<tr>
<th>Tag</th>
<th>Singleton PET</th>
<th>Duplex Interaction</th>
<th>Complex Interaction</th>
</tr>
</thead>
</table>

**The looping model of chromatin interaction**

- **Anchor region**
- **Large loop**
- **Small gene-centric loop**
- **Loop genes (grey)**
- **Anchor genes (blue & green)**
- **Genes outside of interaction (grey)**