Using model-based methods to analyze NGS data

Steve Qin
Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University
“... deep things in science are not found because they are useful; they are found because it was possible to find them”

-- Robert Oppenheimer
Next generation sequencing technologies
Illumina sequencing technology
Sequence data

```
AAAAATCTCTTCTGAAACCATTTCAGAAAATGC
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AACAGACCTAAATCGCTATTGCAATATCTTT
AACCCAGCGACCTGCGACTCCCTGGACGTGTAC
ATGTAGGGTTGTACGTAGAACTCTATTAT
ATTGCCAGAAAGTACTGAGCTATCATGAGAT
ATCCCGATCCCCTGTTAACAGAGTCATTGTA
ACCACCACAATGAATATCAAACCACCTCTC
ATGGGGGAAATATTGCAATTATGTAAAGGTAA
ATGTCTAAAGATCCACTTTTTAAAACTATATT
ATATAAAGTCTCTCTCTCTACTCTTCTCCTC
AGGGAACTACTCCACCCTCTGAGCTCTCCTAG
AAAAAGATATATATATATATATATCTATAATA
AGTCCGACCTCGACCTGTCTCTGCTCTGAGA
ATTTGGTGAGTAAATTTAAGAGAGTGAGTACAT
GGTCGTGTGTGTCAATGGCCGTCTCTTCTCCT
ATTGAAAAGAGTCCTCTCTAGAAATGTTAAATA
AGGGACTGAAGCTGCTGGGCCCATGTTTTAG
AGAAAAATATAAAAATCTTTGAAGAAGAAGAAG
AAGGGGATTAGGGGTTTCTGCGGCGCAATTTT
AGAACCCTCCATAAACCTGAGAGTACTATATG
AATAAGTCGGTCTAGAGATCCAGGAAACTCTT
ATTGGGTTTGGCTATCCACCCCGTTACAA
CGGGGATAAGTGTTGTTTCGAAAGAAGATATAA
```
Applications of NGS

So many –seq, so little time
ALEXA-Seq, Apopto-Seq, AutoMeDip-Seq, Bind-n-Seq, Bisulfite-Seq, ChIP-Seq, CLIP-Seq, CNV-Seq, DGE-Seq, DNA-Seq, DNase-Seq, F-Seq, FRT-Seq, HITS-CLIP, indel-Seq, MBD-Seq, MeDIP-Seq, MethylCap-Seq, microRNA-Seq, mRNA-Seq, NA-Seq, NSR-Seq, PAS-Seq, Peak-Seq, ReChIP-Seq, RIP-Seq, RNA-Seq, rSW-Seq, SAGE-Seq, Sono-Seq, Tn-Seq...
So many –seq, so little time
ALEXA-Seq, Apopto-Seq, AutoMeDip-Seq, Bind-n-Seq, Bisulfite-Seq, ChIP-Seq, CLIP-Seq, CNV-Seq, DGE-Seq, DNA-Seq, DNase-Seq, F-Seq, FRT-Seq, HITS-CLIP, indel-Seq, MBD-Seq, MeDIP-Seq, MethylCap-Seq, microRNA-Seq, mRNA-Seq, NA-Seq, NSR-Seq, PAS-Seq, Peak-Seq, ReChIP-Seq, RIP-Seq, RNA-Seq, rSW-Seq, SAGE-Seq, Sono-Seq, Tn-Seq...
Chromatin Immunoprecipitation
ChIP-chip and ChIP-Seq technologies

- Cross-link and shear
- IP
- Purify, amplify, and label
- Hybridization
- genomics DNA

Ren et al. 1999; Iyer et al. 2000
ChIP sequencing

High-Resolution Profiling of Histone Methyllations in the Human Genome

Artem Barski,1,5 Suresh Cuddapah,1,5 Kairong Cui,1,5 Tao-Young Roh,1,5 Dustin E. Schones,1,5 Zhibin Wang,1,5 Gang Wei,1,5 Iouri Chepelev,2 and Koji Zhao1,5

1Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA
2Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles, Los Angeles, CA 90095, USA
3These authors contributed equally to this work and are listed alphabetically.
4Correspondence: zhaok@nibib.nih.gov
DOI 10.1016/j.cell.2007.05.009

Genome-Wide Mapping of in Vivo Protein-DNA Interactions

David S. Johnson,2,4 Ali Mortazavi,2, Richard M. Myers,2, Barbara Wold2,4

www.sciencemag.org SCIENCE VOL 316 8 JUNE 2007

Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing

Gordon Robertson1, Martin Hirst1, Matthew Bainbridge1, Misha Bilenky1, Yongjun Zhao1, Thomas Zeng1, Ghaia Euskirchen2, Bridget Bernier1, Richard Varhol1, Allen Denley1, Nina Thiesen1, Obi I. Griffith1, Ann He1, Marco Marra1, Michael Snyder2 & Steven Jones1

1British Columbia Cancer Agency Genome Sciences Center, 675 West 10th Avenue, Vancouver, British Columbia V5Z 4S6, Canada. 2Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA. Correspondence should be addressed to S.J. (sjones@bgsc.ca).
RECEIVED 11 MAY; ACCEPTED 5 JUNE; PUBLISHED ONLINE 11 JUNE 2007; DOI:10.1038/nmeth1068

ChIP-Seq papers

Graph showing the number of ChIP-Seq papers published from 2007 to 2014.
Outline

• Hidden Markov model for peak detection
• Hierarchical Hidden Markov model for combining ChIP-seq and ChIP-chip data
• Hybrid Monte Carlo strategy for Motif finding
Align reads to genome, get summary statistics, estimate model parameters.

Get read coverage for each bin on all chromosomes.

Build HMM to infer whether a bin belongs to peak or background.

Post-processing on identified peaks.
Motif enrichment results for NRSF and STAT1 data
HPeak performance

Laajala et al. *BMC Bioinformatics*, 2009
GP and ZIP distribution

• Do not require mean equal to variance which is useful to model over-dispersion and under-dispersion.

\[
P(Y = y | \lambda, \phi) = \left( \frac{\lambda}{1 + \phi \lambda} \right)^y \frac{(1 + \phi \lambda)^{y-1}}{y!} \exp \left\{ -\frac{\lambda(1 + \phi \lambda)}{1 + \phi \lambda} \right\}
\]

\[
E(Y) = \lambda
\]

\[
Var(Y) = \lambda (1 + \phi \lambda)^2
\]

• Zero-inflated Poisson distribution

\[
f(Y | \pi, \mu) = \begin{cases} (1 - \pi) + \pi e^{-\mu} & \text{if } x = 0 \\ \frac{\pi e^{-\mu} \mu^x}{x!} & \text{if } x = 0 \end{cases}
\]
Comparison between ChIP-seq and ChIP-chip
Outline

• Hidden Markov model for peak detection

• Hierarchical Hidden Markov model for combining ChIP-seq and ChIP-chip data

• Hybrid Monte Carlo strategy for Motif finding
Joint analysis of ChIP-chip and ChIP-seq

Hierarchical HMM
Simulated data results
Real Data Analyses

- **NRSF**

<table>
<thead>
<tr>
<th>Method</th>
<th>#Match(^a) (#Permute(^b))</th>
<th>#Peaks</th>
<th>Coverage(Kb)</th>
<th>OR(^c)</th>
<th>$\chi^2$</th>
<th>Match Rate(^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHMM</td>
<td>46 (11)</td>
<td>424</td>
<td>179.2</td>
<td>4.56</td>
<td>21.74</td>
<td>0.19</td>
</tr>
<tr>
<td>Union</td>
<td>67 (24)</td>
<td>860</td>
<td>293.0</td>
<td>2.94</td>
<td>20.47</td>
<td>0.15</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>25 (4)</td>
<td>61</td>
<td>26.5</td>
<td>9.89</td>
<td>18.09</td>
<td>0.79</td>
</tr>
<tr>
<td>ChIP-chip</td>
<td>52 (17)</td>
<td>830</td>
<td>272.9</td>
<td>3.20</td>
<td>17.48</td>
<td>0.13</td>
</tr>
<tr>
<td>Intersect</td>
<td>10 (1)</td>
<td>25</td>
<td>6.6</td>
<td>16.00</td>
<td>7.46</td>
<td>1.36</td>
</tr>
</tbody>
</table>

- **CTCF**

<table>
<thead>
<tr>
<th>Method</th>
<th>#Match(^a) (#Permute(^b))</th>
<th>#Peaks</th>
<th>Coverage(Mb)</th>
<th>OR(^c)</th>
<th>$\chi^2$</th>
<th>Match Rate(^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHMM</td>
<td>23,772 (4,815)</td>
<td>65,808</td>
<td>30.31</td>
<td>7.16</td>
<td>16,057.36</td>
<td>0.63</td>
</tr>
<tr>
<td>Union</td>
<td>26,788 (6,200)</td>
<td>83,325</td>
<td>40.08</td>
<td>5.89</td>
<td>16,018.71</td>
<td>0.51</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>16,771 (1,836)</td>
<td>25,372</td>
<td>9.33</td>
<td>25.00</td>
<td>18,926.85</td>
<td>1.60</td>
</tr>
<tr>
<td>ChIP-chip</td>
<td>16,599 (5,134)</td>
<td>69,246</td>
<td>33.83</td>
<td>3.94</td>
<td>7,172.77</td>
<td>0.34</td>
</tr>
<tr>
<td>Intersect</td>
<td>6,310 (719)</td>
<td>9,576</td>
<td>3.06</td>
<td>23.80</td>
<td>7,023.18</td>
<td>1.83</td>
</tr>
</tbody>
</table>
Outline

• Hidden Markov model for peak detection
• Hierarchical Hidden Markov model for combining ChIP-seq and ChIP-chip data
• Hybrid Monte Carlo strategy for Motif finding
Example: cyclic receptor protein (CRP)
Example: cyclic receptor protein (CRP)
Transcription factor binding site (TFBS)
Existing *de novo* motif finding algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus</td>
<td>Hertz <em>et al.</em> 1990</td>
</tr>
<tr>
<td>Gibbs Motif Sampler</td>
<td>Lawrence <em>et al.</em> 1993</td>
</tr>
<tr>
<td>MEME</td>
<td>Bailey and Elkan 1994</td>
</tr>
<tr>
<td>AlignACE</td>
<td>Roth <em>et al.</em> 1998</td>
</tr>
<tr>
<td>BioProspector</td>
<td>Liu <em>et al.</em> 2001</td>
</tr>
<tr>
<td>MDScan</td>
<td>Liu <em>et al.</em> 2002</td>
</tr>
<tr>
<td>Mobydick</td>
<td>Bussemaker <em>et al.</em> 2000</td>
</tr>
</tbody>
</table>

... Review                  | Tompa *et al.* 2005       |
Motif identification model

Alignment variable $A = \{a_1, a_2..., a_J\}$
Posterior distributions

- The posterior conditional distribution for alignment variable $A$

$$p(a_j = l | \theta_0, \Theta, R, A_{-j}) \propto \prod_{k=1}^{4} \theta_{0k}^{h_k(R_j)} \prod_{i=1}^{w} \prod_{k=1}^{4} \left( \frac{\theta_{ik}}{\theta_{0k}} \right)^{h_k(r_{j,i+l-1})}$$

DNA sequence data

$R = (R_1, ..., R_J)$

Why *de novo* motif search

- The only option when the TF binding motif pattern is unknown.
- Reassuring to be able to rediscover the known TFBS motif.
- Many “known” motif patterns are biased and inaccurate.
- Multiple co-factors are often required in transcription regulation in eukaryotes.
- Binding specificity for some TFs may change under different conditions.
Challenges faced

• How to handle large number of input sequences?
• How to utilize sequencing depth information?
Features of our new algorithm

• Incorporate sequencing depth information in the statistical model.

• Generalize the product multinomial model to allow inter-dependent positions within the motif.

• Adopt a hybrid Monte Carlo strategy to speed up the traditional Gibbs sampler-based algorithm.
The informative prior

• The prior is symmetric and centered at the peak summit.
• The prior probabilities stem from Student’s $t$-distribution with $df=3$.

$$p(a_j = l) \propto t_3 \left( \int \frac{|l + w/2 - s_j| + u/2}{u} \right)$$
Modeling inter-dependent positions

• Zhou and Liu
  *Bioinformatics* 2005

• Barash *et al.*
  *RECOMB* 2003
Detect intra-dependent position pairs

\[ d_{ij} = \sum_{x=1}^{4} \sum_{y=1}^{4} \left| \hat{\eta}_{xy}(r_i, r_j) - \hat{\eta}_x(r_i)\hat{\eta}_y(r_j) \right| \]
New algorithm

- The posterior conditional distribution of alignment variable $A$ under the new statistical model.
Prioritized hybrid Monte Carlo

- Subject each sequence to either stochastic sampling or greedy search.
- Input sequences are not created equal.
- ChIP-enrichment is indicative of binding affinity.
Implementation

• Hybrid Motif Sampler (HMS).
• Gibbs sampler type iterative procedure.
• Run multiple chains to avoid trapping in local mode.
Performance comparison

- Two established and popular motif discovery tools:
  - MEME (Bailey and Elkan 1994),
    - EM-based motif finding algorithm,
    - widely used.
  - MDscan (Liu et al. 2002),
    - designed to analyze ChIP-chip data,
    - combines word enumeration and probability matrix updating,
    - take into account ChIP-chip ranking,
    - very fast.
## Real data analysis

<table>
<thead>
<tr>
<th>TF</th>
<th>Cell type</th>
<th>Antibody</th>
<th># of peaks</th>
<th>Coverage</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRSF</td>
<td>Jurkat T cell</td>
<td>12C11 Monoclonal</td>
<td>4,982</td>
<td>1.4 MB</td>
<td>Johnson et al. (2007)</td>
</tr>
<tr>
<td>STAT1</td>
<td>HeLa S3 cell</td>
<td>Polyclonal</td>
<td>27,470</td>
<td>8.1 MB</td>
<td>Robertson et al. (2007)</td>
</tr>
<tr>
<td>CTCF</td>
<td>CD4+ T cell</td>
<td>Upstate 07-729</td>
<td>22,159</td>
<td>7.4 MB</td>
<td>Barski et al. (2007)</td>
</tr>
<tr>
<td>ER</td>
<td>MCF7 cell</td>
<td>ER α (HC-20)</td>
<td>10,072</td>
<td>2.5 MB</td>
<td></td>
</tr>
</tbody>
</table>
Performance evaluation

• Cross validation
  – Randomly separate all peaks into two halves: training and testing.
  – Run motif finding algorithms on the training data to predict the motif pattern.
  – Scan testing data using the identified motif pattern and compare to a set of control sequences.

• Testing
  – Using Chi-square test statistics to quantify motif enrichment.
  – Estimate FDR and plot FDR versus Chi-square test statistics.
Compare ER motif patterns

- V$ER01*
- V$ER02*
- V$ER03*
- MEME
- HMS
Positions show inter-dependency inside the ER motif
Compare ER motif enrichment

Carroll et al. Nature Genetics 2006
Compare NRSF motif enrichment

Johnson et al. Science 2007
Compare CTCF motif enrichment

Kim et al. Cell 2007
Compare STAT1 motif enrichment

Euskirchen et al. Genome Res
Computation time
Summary

• ChIP-Seq data offers abundant information and provides much improved opportunity for studying protein-DNA interaction.

• There are many biological and technical factors that affect the ChIP-Seq data we observe, careful modeling is critical in order to process ChIP-Seq data efficiently and thoroughly.

• New sequencing data are different from microarray, ChIP-chip data. Methods developed there do not work well for analyzing sequencing data, new models and algorithms need to be developed.
Apply to cancer genomics
Reference


Statistical model to infer chromosomal structures from Hi-C data
Chromosome folding

How can a two meter long polymer fit into a nucleus of ten micrometer ($10^{-5}$ m) diameter?

http://en.wikipedia.org/wiki/Chromosome
Chromosome Conformation Capture (3C)
Dekker et al. Science 2002

Fine scale: (0-kb)

Naumova and Dekker J of Cell Science 2010
3C-on-chip/Circular 3C (4C)

5C

Intermediate: (0-Mb)

Fine scale: (0-kb)

Naumova and Dekker *J of Cell Science* 2010
Whole genome

Intermediate: (0-Mb)

Fine scale: (0-kb)

Naumova and Dekker *J of Cell Science* 2010
Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

Erez Lieberman-Aiden,1,2,3,4* Nynke L. van Berkum,5* Louise Williams,1 Maxim Imakaev,2 Tobias Ragoczy,6,7 Agnes Telling,6,7 Ido Amit,1 Bryan R. Lajoie,3 Peter J. Sabo,6 Michael O. Dorschner,6 Richard Sandstrom,8 Bradley Bernstein,2,9 M. A. Bender,10 Mark Groudine,6,7 Andreas Gnirke,1 John Stamatoyannopoulos,8 Leonid A. Mirny,2,11 Eric S. Lander,1,12,13† Job Dekker5†

We describe Hi-C, a method that probes the three-dimensional architecture of whole genomes by

We created a Hi-C library from a karyotypically normal human lymphoblastoid cell line (GM06990) and sequenced it on two lanes of an Illumina Genome Analyzer (Illumina, San Diego, CA), generating 8.4 million read pairs that could be uniquely aligned to the human genome reference sequence; of these, 6.7 million corresponded to long-range contacts between segments >20 kb apart.

We constructed a genome-wide contact matrix $M$ by dividing the genome into 1-Mb regions (“loci”) and defining the matrix entry $m_{ij}$ to be the number of ligation products between locus $i$ and locus $j$ (/10). This matrix reflects an ensemble
Hi-C: one cell

Cross-link DNA

Restriction enzyme cut

HindIII

5’-A AGCT T-3’

3’-T TCGA A-5’

Ligation and shear

Paired-end sequencing
We created a Hi-C library from a karyotypically normal human lymphoblastoid cell line (GM06990) and sequenced it on two lanes of an Illumina Genome Analyzer (Illumina, San Diego, CA), generating 8.4 million read pairs that could be uniquely aligned to the human genome reference sequence; of these, 6.7 million corresponded to long-range contacts between segments >20 kb apart.

We constructed a genome-wide contact matrix $M$ by dividing the genome into 1-Mb regions (“loci”) and defining the matrix entry $m_{ij}$ to be the number of ligation products between locus $i$ and locus $j$ (10). This matrix reflects an ensemble
Hi-C Data Representation

3D chromosomal structure

Pair-wise spatial distance matrix

\[
\begin{array}{cccccc}
  d(1,1), & d(1,2), & \ldots & d(1,N) \\
  d(2,1), & d(2,2), & \ldots & d(2,N) \\
  \vdots & \vdots & \ddots & \vdots \\
  d(N,1), & d(N,2), & \ldots & d(N,N) \\
\end{array}
\]

High resolution
High throughput
Challenges

• Quality control and pre-processing of the reads,
• Any bias in the data? and if so, how to normalize?
• Whether it is possible, and if so, how, to infer the 3-dimensionnal chromosomal structure based on the Hi-C data?
Hi-C Data Preprocess

Restriction enzyme cutting site

Self-ligation reads

Dangling reads

PCR amplification reads

Random break

Random break

Valid reads

Downstream analysis
Methods for Hi-C Bias Reduction

• Normalization (equal ‘visibility’, no assumption on biases)
  - Iterative correction and eigenvector decomposition (ICE) (Imakaev, et al, 2012)
  - Sequential component normalization (SCN) (Cournac, et al, 2012)

• Correction (posit a statistical model on biases)
  - Yaffe & Tanay’s method (Yaffe & Tanay, 2011)
    Fragment level (4KB, $10^{12}$), 420 parameters
  - **HiCNorm (Hu et al, 2012)**
    Any resolution level
    1MB, $10^6$, 3 parameters
3D structure prediction

- Challenges:
  - Sequencing uncertainties
  - Biases: enzyme, GC content, mappability
What does the number mean?

• The Hi-C experiment is conducted on millions of cells,
• A captured pair-end read is from one cell,
• A number in the matrix (loci $i$ and $j$) indicates the frequency of capture (link $i$ and $j$) in the cell population,
• Do those numbers say anything about 3D distance?
Motivation and the key assumption

- Number of paired-end reads spanning the two loci is inversely proportional to the 3D spatial distance between them (obtained from fluorescence in situ hybridization (FISH)).

[Graph showing the relationship between average distance in microns and the logarithm of the number of reads.]

Lieberman-Aiden et al, 2009
Existing methods

  ➢ Biophysical properties of chromatin fiber.
  ➢ No consideration of systematic biases.
  ➢ No statistical inference.

• Statistical method: MCMC5C (Rousseau et al, 2011)
  ➢ Normal model for count data.
  ➢ No consideration of systematic biases.
ACGTAAGCTAGATACATTGATACATCGATAGCGTAGTGGTAGGAGGGATCATG
ACGTAAGCTAGATACATTGATACATCGATAGCGTAGTGGTAGGAGGGATCATG
Model
Beads-on-string
Beads-on-string

\[ P_i = (x_i, y_i, z_i) \]

\[ P_{i-1} = (x_{i-1}, y_{i-1}, z_{i-1}) \]

\[ P_{i+1} = (x_{i+1}, y_{i+1}, z_{i+1}) \]
Beads-on-string

\[ p_i = (x_i, y_i, z_i) \]

\[ p_{i-1} = (x_{i-1}, y_{i-1}, z_{i-1}) \]

\[ p_{i+1} = (x_{i+1}, y_{i+1}, z_{i+1}) \]
Bayesian statistical model

- $u_{ij}$: number of reads between loci $i$ and $j$.
- $d_{ij}$: 3D Euclidian distance between loci $i$ and $j$.
- $enz_i$: number of enzyme cut site in locus $i$.
- $gcc_i$: mean GC content in locus $i$.
- $map_i$: mean mappability score in locus $i$.

$$u_{ij} \sim \text{Poisson}(\theta_{ij})$$

$$\log(\theta_{ij}) = \beta_0 + \beta_1 \log(d_{ij}) + \beta_{enz} \log(enz_i \cdot enz_j)$$
$$+ \beta_{gcc} \log(gcc_i \cdot gcc_j) + \beta_{map} \log(map_i \cdot map_j)$$
Bayesian Statistical Model

- Likelihood: \( \begin{pmatrix} N \\ 2 \end{pmatrix} \) data points, \( 3N + 5 \) parameters

\[
L(u_{ij}, 1 \leq i < j \leq N | x_i, y_i, z_i, 1 \leq i \leq N, \beta_0, \beta_1, \beta_e, \beta_g, \beta_m) = \prod_{1 \leq i < j \leq N} \frac{e^{-\theta_{ij} \theta_{ij}^u}}{u_{ij}!}
\]

\[
\log(\theta_{ij}) = \beta_0 + \beta_1 \log\left( \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \right) \\
+ \beta_e \log(e_i e_j) + \beta_g \log(g_i g_j) + \beta_m \log(m_i m_j)
\]
Statistical Inference

- **Algorithm:** Bayesian 3D constructor for Hi-C data (BACH)

  - Initialization 1: use Poisson regression to obtain the initial values of model parameters.
  
  - Initialization 2: use **sequential important sampling** to get the initial 3D chromosomal structure.
  
  - Refinement: use Gibbs sampler with **hybrid Monte Carlo** to refine the initial values for parameters.
SIS in BACH: Outline

• Goal: use sequential importance sampling to **sequentially** put $N$ loci into 3D space, i.e. sample from:

\[ \pi(x_i, y_i, z_i, 1 \leq i \leq N | u_{ij}, 1 \leq i < j \leq N) \]

• Bridging distributions:

\[ \pi_t(x_i, y_i, z_i, 1 \leq i \leq t | u_{ij}, 1 \leq i < j \leq t) \]

• Proposal distributions (given the previous $t-1$ loci, put the $t$ th locus into 3D space):

\[ g_t(x_t, y_t, z_t | x_i, y_i, z_i, 1 \leq i \leq t - 1, u_{ij}, 1 \leq i < j \leq t) \]
Simulation study

• Use random walk to simulate a 3D structure with 33 loci (red lines). Simulate Hi-C contact map from the posited model.

• Predicted 3D structure (blue lines) aligns well with true 3D structure (RMSD = 0.0091).
## Human Hi-C data

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Restriction enzyme</th>
<th># of reads (million)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM06990</td>
<td>HindIII</td>
<td>4.1</td>
</tr>
<tr>
<td>GM06990</td>
<td>HindIII</td>
<td>4.4</td>
</tr>
<tr>
<td>GM06990</td>
<td>HindIII</td>
<td>4.9</td>
</tr>
<tr>
<td>GM06990</td>
<td>HindIII</td>
<td>5.4</td>
</tr>
<tr>
<td>GM06990</td>
<td>NcoI</td>
<td>8.8</td>
</tr>
<tr>
<td>GM06990</td>
<td>NcoI</td>
<td>10.1</td>
</tr>
<tr>
<td>K562</td>
<td>HindIII</td>
<td>12.1</td>
</tr>
<tr>
<td>K562</td>
<td>HindIII</td>
<td>9.7</td>
</tr>
</tbody>
</table>
Real Hi-C data from Lieberman-Aiden et al. 2009

d(L2, L4) = 1.4042, d(L2, L3) = 1.9755, significant
mESC: Hind3 vs. Nco1
Whole Chromosome 3D Model

• Two compartments
  - Compartment A: gene rich, active transcription
  - Compartment B: gene poor, inactive transcription

• Same compartment: strong chromatin interactions, spatially close

• Different compartments: weak chromatin interaction, spatially isolated

Two compartment model
Whole Chromosome Model

Naumova and Dekker, 2010
Other Features (Chromosome 2)

- Compartment
- Gene density
- Gene expression
- Chromatin accessibility

- RNA polymerase II
- DNA replication time
- H3K36me3
- H3K27me3
- H3K4me3
- H3K9me3
- H3K20me3
- Lamina interaction
Conclusions

• BACH--Reconstruct chromosome 3D structures
• Remove systematic biases
• Consistent with FISH data
• Elongation of chromatin is highly associated with genetic/epigenetic features.
• Separation of compartments of A and B can be visualized.
More questions to be answered

• Is there a consensus? Or a dominant 3D chromosomal structure?
  – Completely random?
  – Mixture of distinct structures?

• Rigorous inference
  – Variance of the structure

• Computation
References


Acknowledgements

Ming Hu
Ke Deng
Jun S. Liu
Li Li
Chunhui Hou
Victor Corces
Jesse Dixon
Siddarth Selvarajan
Bing Ren
Thank You

Questions: zhaohui.qin@emory.edu