Lecture 8 - Epigenomics
read mapping – peak calling – multivariate HMMs

Goals for today: Computational Epigenomics
1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks
6. Epigenome imputation by exploiting chromatin mark correlations

DNA packaging
• Why packaging
  – DNA is very long
  – Cell is very small
• Compression
  – Chromosome is 50,000 times shorter than extended DNA
• Using the DNA
  – Before a piece of DNA is used for anything, this compact structure must open locally
• Now emerging:
  – Role of accessibility
  – State in chromatin itself
  – Role of 3D interactions

One Genome – Many Cell Types

Three types of epigenetic modifications

Image source: http://nihroadmap.nih.gov/epigenomics/
100s of histone tail modifications

- 100+ different histone modifications
- Histone protein → H3/H4/H2A/H2B
- AA residue → Lysine4(K4)/K36...
- Chemical modification → Met/Pho/Ubi
- Number → Me-Me-Me(me3)
- Shorthand: H3K4me3, H2BK5ac

In addition:
- DNA modifications
- Methyl-C in CpG / Methyl-Adenosine
- Nucleosome positioning
- DNA accessibility
- The constant struggle of gene regulation
- TF/histone/nucleo/GFs/Chrom compete

100s of known modifications, many new still emerging
Systematic mapping using ChIP-, Bisulfite-, DNase-Seq

Epigenomics Roadmap across 100+ tissues/cell types

Diverse tissues and cells:
1. Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood...)
2. Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood...)
3. ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench, trophobl)

Ongoing epigenomic mapping projects

- Mapping multiple modifications
- In multiple cell types
- In multiple individuals
- In multiple species
- In multiple conditions
- With multiple antibodies
- Across the whole genome

ChIP-seq review
(Chromatin immunoprecipitation followed by sequencing)

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data

2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)

3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states

4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence

5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

6. Epigenome imputation by exploiting chromatin mark correlations
ChIP-chip and ChIP-Seq technology overview

Modification-specific antibodies → Chromatin Immuno-Precipitation followed by: ChIP-chip: array hybridization ChIP-Seq: Massively Parallel Next-gen Sequencing

ChIP-Seq Histone Modifications: What the raw data looks like

- Each sequence tag is 30 base pairs long
- Tags are mapped to unique positions in the ~3 billion base reference genome
- Number of reads depends on sequencing depth. Typically on the order of 10 million mapped reads.

Summarize multiple marks into chromatin states

Chromatin state track summary

Mapping millions of short reads to the genome

Traditional Hashing Schemes
Burrows-Wheeler Transform (BWT)

Mapping Reads to the Genome

- Assign reads to best matching location in reference genome
- 10,000,000s of reads, ~30 bases long
- Example: CAGGGCTGATTGAGGACATTCATCACG
- Allow mismatches: sequencing errors, or SNPs
- Algorithmic and memory efficiency is critical

- L2: Sequence alignment: O(m*n)
- L3: Hashing / BLAST: O(m+n)
  - Solution until 2008 (e.g. MAQ, Li et al, GR 2008)
- Other advanced algorithms:
  - Linear-time string matching: O(m+n). L3 addendum
  - Suffix trees and suffix arrays: O(m). L13 addendum
- Challenge: memory requirements
  - Hash table, suffix tree/array require O(m*n) space
- Today: Burrows-Wheeler transformation O(m)
  - Ultrafast/memory efficient. New norm since 2009.
  - Introduced in: Bowtie (Langmead GB 2009).
Second Generation Mappers have Leveraged the Burrows Wheeler Transformation

"...35 times faster than Maq and 300 times faster than SOAP under the same conditions"

Hashing vs. Burrows Wheeler Transform

Today: How does the BW transform actually work?

Burrows-Wheeler Transform (BWT)

• Transform: ^BANANA@ INTO: BNN^AA@A

function BWT (string s)
create a table, rows are all possible rotations of s
sort rows alphabetically
return (last column of the table)

function inverseBWT (string s)
create empty table
repeat length(s) times
insert s as a column of table before first column of the table // first insert creates first column sort rows of the table alphabetically return (row that ends with the 'EOF' character)

Last column only suffices to reconstruct entire matrix, and thus recover original string

Searching for an Exact Match

e.g. Searching for OLIS
In MANOLISKELLIS
For simplicity (here):
- only exact matches
- Show entire matrix
In practice: only pointers

OLIS
1. SMANOLISKELLIS
2. MANOLISKELLIS
3. ANOLISKELLIS
4. NOLISKELLIS
5. OLISKELLIS
6. LISKELLIS
7. KELLIS
8. SSLIS
9. KILIS
10. ISL
11. ILS
12. LS
13. S
14. SKELLIS
15. KELLIS
16. LISK
17. IS
18. SL
19. LS
20. SL
21. LS
22. SL
23. LS
24. SL
25. LS
26. SL
27. LS
28. SL
29. LS
30. SL
31. LS
32. SL
33. LS
34. SL
35. LS
36. SL
37. LS
38. SL
39. LS
40. SL
41. LS
42. SL
43. LS
44. SL
45. LS
46. SL
47. LS
48. SL
49. LS
50. SL
51. LS
52. SL
53. LS
54. SL
55. LS
56. SL
57. LS
58. SL
59. LS
60. SL
61. LS
62. SL
63. LS
64. SL
65. LS
66. SL
67. LS
68. SL
69. LS
70. SL
71. LS
72. SL
73. LS
74. SL
75. LS
76. SL
77. LS
78. SL
79. LS
80. SL
81. LS
82. SL
83. LS
84. SL
85. LS
86. SL
87. LS
88. SL
89. LS
90. SL
91. LS
92. SL
93. LS
94. SL
95. LS
96. SL
97. LS
98. SL
99. LS
100. SL


Key properties of Burrows-Wheeler Transform

• Very little memory usage. Same as input (or less)
  – Don’t represent matrix, or strings, just pointers
  – Encode: Simply sort pointers. Decode: follow pointers

• Original application: string compression (bZip2)
  – Runs of letters compressed into (letter, runlength) pairs

• Bioinformatics applications: substring searching
  – Achieve similar run time as hash tables, suffix trees
  – But: very memory efficient ➔ practical speed gains

• Mapping 100,000s of reads: only transform once
  – Pre-process once; read counts in transformed space.
  – Reverse transform once, map counts to genome coordinates
**Goals for today: Computational Epigenomics**

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
     - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks
6. Epigenome imputation by exploiting chromatin mark correlations

---

**ENCODE uniform processing pipeline**

**QC1: Use of input DNA as control dataset**

- **Challenge:**
  - Even without antibody: Reads are **not** uniformly scattered
- **Sources of bias in input dataset scatter:**
  - Non-uniform fragmentation of the genome
  - Open chromatin fragmented more easily than closed regions
  - Repetitive sequences over-collapsed in the assembled genome.
- **How to control for these biases:**
  - Remove portion of DNA sample before ChIP step
  - Carry out control experiment without an antibody (input DNA)
  - Fragment input DNA, sequence reads, map, use as background

**QC2: Read-level sequencing quality score Q>10**

- **High quality reads**
  - Each column is a color-coded histogram
  - Encodes fraction of all mapped reads that have base score Q (y-axis) at each position (x-axis)
  - Darker blue = higher density
  - Read quality tends to drop towards the ends of reads
  - Low average per base score implies greater probability of mismappings.
  - Typically, reject reads whose average score Q < 10

- **Low quality reads**

**QC3: Fraction of short reads mapped >50%**

- Reads can map to:
  - exactly one location (uniquely mapping)
  - multiple locations (repetitive or multi-mapping)
  - no locations (unmappable)

- **Dealing with multiply-mapping reads:**
  - Conservative approach: do not assign to any location
  - Probabilistic approach: assign fractionally to all locations
  - Sampling approach: pick one location at random, averages across many reads
  - EM approach: map according to density, estimated from unambiguous reads
  - Pair-end approach: use paired end read to resolve ambiguities in repeat reads

- **Absence of reads in a region could be due to:**
  - No assembly coverage in that region (e.g. peri-centromeric region)
  - Too many reads mapping to this location (e.g. repetitive element)
  - No activity observed in this location (e.g. inactive / quiescent / dead regions)

- **Dealing with mappability biases:**
  - ‘Black-listed’ regions, promiscuous across many datasets
  - ‘White-listed’ regions, for which at least some dataset has unique reads
  - Treat unmappable regions as missing data, distinguish from ‘empty’ regions

---

**Quality control metrics**

ChIP vs. Input DNA
Read quality
Mappability
Library complexity
How many distinct uniquely mapping reads? How many duplicates?
If your sample does not contain sufficient DNA and/or you over-sequence, you will simply be repeatedly sequencing PCR duplicates of a restricted pool of distinct DNA fragments. This is known a low-complexity library and is not desirable.

- Histogram of no. of duplicates
- Non-redundant fraction (NRF) = \frac{\text{No. of 'distinct' unique-mapping reads}}{\text{No. of unique-mapping reads}}
- NRF should be > 0.8 when 10M < #reads < 80M unique-mapping reads

Cross-correlation analysis
Exploiting forward and reverse reads
- Fragment-length peak
- Phantom read-length peak

Cross-correlation at read vs. fragment length

1. Calculate forward and reverse strand signals
2. Shift both by specified offset towards each other
3. Calculate correlation of two signals at that shift
4. Correlation peaks at fragment length offset \( f \)

\( f \) is the length at which ChIP DNA is fragmented

Goals for today: Computational Epigenomics
1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinators
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks
6. Epigenome imputation by exploiting chromatin mark correlations

ChIP-seq: exploiting forward and reverse reads
(Chromatin immunoprecipitation followed by sequencing)
Multiple IP fragments are obtained corresponding to each binding event

- Ends of the fragments are sequenced i.e. “Short-reads/tags”
  - Typically ~36 bp, 50 bp, 76 bp or 101 bp

- Single-end (SE) sequencing
  - Randomly sequence one of the ends of each fragment

- Paired-end (PE) sequencing
  - sequence both ends of each fragment

Canonical “stranded mirror distribution of short-reads” after mapping reads to genome
- Heaps of reads on the + strand and – strand separated by a distance ~ fragment length

Cross-correlation (CC) analysis

- Strand shift \( s = \frac{f}{2} \pm \frac{f}{2} \)

- Sign of a good dataset:
  - High absolute cross-correlation at fragment length (NSC)
  - High fragment length CC relative to read length CC (RSC)
• Input dataset (no ChIP) shows ‘phantom’ peak at read length only
• Due to read mappability:
  • If position ‘x’ is uniquely mappable on + strand
  • Then position ‘x+r-1’ is uniquely mappable on – strand
  • Fragment-length peak should always dominate the read-length peak

Where does read cross-correlation come from?

Example of good, medium, bad CC datasets

Goals for today: Computational Epigenomics
1. Introduction to Epigenomics
   – Overview of epigenomics, Diversity of Chromatin modifications
   – Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   – Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   – Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   – A multi-variate HMM for chromatin combinatorics
   – Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   – Selecting the number of states, selecting number of marks
   – Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   – Stacking vs. concatenation approach for joint multi-cell type learning
   – Defining activity profiles for linking enhancer regulatory networks
6. Epigenome imputation by exploiting chromatin mark correlations

Peak Calling

Continuous signal → Intervals

Peak calling: detect regions of enrichment

Goal: Transform read counts into normalized intensity signal

Steps:
1. Estimate fragment-length l using strand cross-correlation analysis
2. Extend each read from 5’ to 3’ direction to fragment length l
3. Sum intensity for each base in ‘extended reads’ from both strands
4. Perform same operation on input-DNA control data (correct for sequencing depth differences)
5. Calculate enrichment ratio value for every position in the genome

Result: Enrichment fold difference for ChIP / control signal

Peak calling: identify discrete intervals

doi:10.1371/journal.pone.0011471
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0011471
Peak calling thresholds

Poisson p-value thresholds
• Read count model: Locally-adjusted Poisson distribution
  \[ P(\text{count} = x) = \frac{\lambda_{\text{local}} \exp(-\lambda_{\text{local}})}{x!} \]
• \[ \lambda_{\text{local}} = \max(\lambda_{\text{BG}}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k}) \] estimated from control data
  • Poisson p-value = \( P(\text{count} \geq x) \)
  • q-value: Multiple hypothesis correction

Peaks: Genomic locations that pass a user-defined p-value (e.g. 1e-5) or q-value (e.g. 0.01) threshold

Empirical False discovery rates
• Swap ChIP and input-DNA tracks
• Recompute p-values
  • At each p-value, eFDR = Number of control peaks / Number of ChIP peaks
• Use an FDR threshold to call peaks

Issues with peak calling thresholds

Cannot set a universal threshold for empirical FDRs and p-values
• Depends on ChIP and input sequencing depth
• Depends on binding ubiquity of factor
• Stronger antibodies get an advantage

FDRs quite unstable
• Small changes in threshold => massive changes in peak numbers

Difficult to compare results across peak callers with a fixed threshold
• Different methods to compute eFDR or q-values

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   – Overview of epigenomics, Diversity of Chromatin modifications
   – Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   – Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   – Quality Control, Cross-correlation, Peak calling (IDR (similar to FDR))
3. Discovery and characterization of chromatin states
   – A multi-variate HMM for chromatin combinatorics
   – Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   – Selecting the number of states, selecting number of marks
   – Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   – Stacking vs. concatenation approach for joint multi-cell type learning
   – Defining activity profiles for linking enhancer regulatory networks
6. Epigenome imputation by exploiting chromatin mark correlations

Selecting meaningful peaks using reproducibility

Use peak ranks in replicate experiments
IDR: Irreproducible Discovery Rate

IDR Cutoff

How to combine two replicates

• Challenge:
  – Replicates show small differences in peak heights
  – Many peaks in common, but many are unique
• Problem with simple solutions:
  – Union: too lenient, keeps garbage from both
  – Intersection: too stringent, throws away good peaks
  – Sum: does not exploit independence of two datasets

IDR idea: Exploit peak rank similarity in replicates

• Key idea: True peaks will be highly ranked in both replicates
  • Keep going down rank list, until ranks are no longer correlated
  • This cutoff could be different for the two replicates
  • The actual peaks included may differ between replicates
  • Adaptively learn optimal peak calling threshold
  • FDR threshold of 10% \( \Rightarrow \) 10% of peaks are false (widely used)
  • IDR threshold of 10% \( \Rightarrow \) 10% of peaks are not reproducible
The IDR model: A two component mixture model

- Looking only at ranks means that the marginals are uniform, so all the information is encoded in the joint distribution.
- Model the joint distribution of ranks as though it came from a two component Gaussian mixture model:
  \[(x, y) \sim pN(\mu, \sigma) + (1 - p)N(0, 1, 1, 0)\]
- This can be fit via an EM-like algorithm.

IDR leads to higher consistence between peak callers

<table>
<thead>
<tr>
<th>IDR</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td># peaks called by SPP</td>
<td># peaks called by MACS</td>
</tr>
</tbody>
</table>

- Compare number of peaks found by two different peak callers
- IDR thresholds are far more robust and comparable than FDR
- FDR only relies on enrichment over input, IDR exploits replicates

IDR pipeline uses replicates when they are available
IDR pipeline also evaluates each replicate individually
- Pooling strategy to generate pseudo-replicates
  - Can pin-point ‘bad’ replicates that may lead to low reproducibility
  - Can estimate IDR thresholds when replicates are not available

Only one good replicate: Pseudo-replicates

IDR pipeline can be used to rescue datasets with only one good replicate (using pseudo-replicates)
IDR pipeline can also be used to call optimal thresholds on a dataset with a single replicate (e.g. when there isn’t enough material to perform multiple reps)

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks
6. Epigenome imputation by exploiting chromatin mark correlations

Chromatin signatures for genome annotation

- Challenges
  - Dozens of marks
  - Complex combinatorics
  - Diversity and dynamics
- Histone code hypothesis
  - Distinct function for distinct combinations of marks?
  - Both additive and combinatorial effects
- How do we find biologically relevant ones?
  - Unsupervised approach
  - Probabilistic model
  - Explicit combinatorics
Summarize multiple marks into chromatin states

Chromatin state track summary

Design Choice

- How to model the emission distribution
  - Model the signal directly
  - Locally binarize the data
- For M input marks each state $k$ has a vector of $(p_{k1}, \ldots, p_{km})$ of parameters for independent Bernoulli random variables which determine the emission probability for an observed combination of marks

Data Binarization

- Leads to biologically interpretable models that can be robustly learned
- Let $c_{ij}$ be the number of reads for mark $i$ mapping to bin $j$. $\lambda_i$ be the average number of reads mapping to a bin for modification $i$. The input for feature $i$ becomes ‘1’ if $P(X>c_{ij})<10^{-4}$

where $X$ is a Poisson random variable with mean $\lambda_i$

Emission Parameter Matrix $e_k(x_i)$

- Multi-variate HMM emits vector of values, not just one value
- Can emit real values (SegWay) or binary presence/absence values (ChromHMM)
- Use to learn mark combinations

Transition matrix $a_{kl}$

- Learns spatial relationships between neighboring states
- Reveals distinct sub-groups of states
- Reveals transitions between different groups

Example Chromatin State Annotation

- Use Baum Welch to learn hidden states and their annotations
- Learned states correspond to known functional elements
- De novo discovery of major types of chromatin
Model complexity matches that of genome

- Handful of repressed states capture vast majority of genome
- Only 1% of genome split in 14 promoter states
- Modeling power well distributed where needed

Apply genome wide to classify chromatin states **de novo**

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Chromatin state characterization: Functional/positional enrichment
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks
6. Epigenome imputation by exploiting chromatin mark correlations

Functional enrichments enable annotation of 51 distinct states

Application of ChromHMM to 41 chromatin marks in CD4+ T-cells (Barski’07, Wang’08)
ZNF repressed state recovery

**State 28: 112-fold ZNF enrich**

*The achievement of the repressed state by wild-type KAP1 involves decreased recruitment of RNA polymerase II, reduced levels of histone H3 K9 acetylation and H3K4 methylation, an increase in histone occupancy, enrichment of trimethyl histone H3K9, H3K36, and histone H4K20.*

**Functional properties of discovered chromatin states**

<table>
<thead>
<tr>
<th>GO Category</th>
<th>State 2</th>
<th>State 3</th>
<th>State 4</th>
<th>State 5</th>
<th>State 6</th>
<th>State 7</th>
<th>State 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Cycle</td>
<td>2.10</td>
<td>0.57</td>
<td>1.61</td>
<td>1.45</td>
<td>1.15</td>
<td>1.51</td>
<td></td>
</tr>
<tr>
<td>Embryonic Development</td>
<td>1.24</td>
<td>2.82</td>
<td>1.07</td>
<td>0.85</td>
<td>0.54</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Chromatin Response</td>
<td>1.20</td>
<td>0.48</td>
<td>2.20</td>
<td>1.64</td>
<td>0.85</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>DNA Damage Response</td>
<td>1.20</td>
<td>0.35</td>
<td>1.55</td>
<td>2.13</td>
<td>1.97</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>RNA Processing</td>
<td>0.49</td>
<td>0.26</td>
<td>1.31</td>
<td>1.91</td>
<td>2.64</td>
<td>2.45</td>
<td></td>
</tr>
<tr>
<td>T-cell Activation</td>
<td>0.77</td>
<td>0.88</td>
<td>1.27</td>
<td>0.70</td>
<td>0.79</td>
<td>4.72</td>
<td></td>
</tr>
</tbody>
</table>

**Transcription End State**

**ZNF repressed state recovery**

**Promoter vs. enhancer regulation**

- TF binding
- Motif enrichment

**Distinct types of repression**

- Chrom bands / HDAC resp
- Repeat family / composition

**Applications to genome annotation**

- New protein-coding genes
- Long intergenic non-coding RNAs / lincRNAs

**Pairwise Expected vs. Observed Mark Co-Occurrence**

- Each point = one pair of chromatin marks
- 41x41 pairs plotted
- X-axis: F(mark 1) * F(mark 2)
- Y-axis: F(mark 1 & mark 2)
- Diagonal: independence
- Off-diag: dependence

**Goals for today: Computational Epigenomics**

1. **Introduction to Epigenomics**
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChiP-Seq, data generation projects, raw data

2. **Primary data processing: Read mapping, Peak calling**
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)

3. **Discovery and characterization of chromatin states**
   - A multi-variate HMM for chromatin combinators
   - Promoter, transcribed, intergenic, repressed, repetitive states

4. **Model complexity: selecting the number of states/marks**
   - Capturing dependencies. State-conditional mark independence
   - Selecting the number of states, selecting number of marks

5. **Learning chromatin states jointly across multiple cell types**
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

6. **Epigenome imputation by exploiting chromatin mark correlations**

**Multi-variate HMM emits entire vector of marks at a time**

*Model assumes mark independence "conditional" upon state*

In fact, it specifically seeks to "capture" these dependencies.
Test conditional independence for each state

Non-independence reveals cases of model violation

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Capturing dependencies. State-conditional mark independence
   - Selecting the number of states, selecting number of marks
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks
6. Epigenome imputation by exploiting chromatin mark correlations

Comparison of BIC Score vs. Number of States for Random and Nested Initialization

Recovery of 79-state model in random vs. nested initialization

Nested initialization approach:
- First pass: learn models of increasing complexity
- Second pass: form nested set of emission parameter initializations by greedily removing states from best BIC model found

Nested models criteria:
- Maximize sum of correlation of emission vectors with nested model
- Models learned in parallel

As more states are added, dependencies captured
- With only 5 states in HMM, not enough power to distinguish different properties
  - Dependencies remain
- As model complexity increases, states learned become more precise
  - Dependencies captured
Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks
6. Epigenome imputation by exploiting chromatin mark correlations

Solution 1: Learn independent models and cluster

Basic approach:
   a) Train a k-state model in each cell type independently
   b) Cluster models learned independently
   c) Merge clusters and re-apply to each cell type

How to cluster
   a) Using emission probability matrix: most similar definitions
   b) Using genome annotation: posterior probability decoding

Solution 2: Stacking

- Learns each combination of activity as a separate state
- Ex: ES-specific enhancers: enhancer marks in ES, no marks in other cell types

Solution 3: Concatenation

- Requires that profiled marks are the same (or treat as missing data)
- Ensures common state definitions across cell types
Joint learning with different subsets of marks (Solution 3)

Option (a) Treat missing tracks as missing data
- EM framework allows for unspecified data points
- As long as pairwise relationship observed in some cell type

Option (b) Chromatin mark imputation
- Explicitly predict max-likelihood chromatin track for missing data
- Less powerful if ultimate goal is chromatin state learning

Chromatin states dynamics across nine cell types
- Single annotation track for each cell type
- Summarize cell-type activity at a glance
- Can study 9-cell activity pattern across

Epigenetic mapping across 100+ tissues/cell types

Diverse tissues and cells
- Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood...)
- Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood...)
- ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench...)

Diverse epigenomic assays
- Histone modifications: H3K4me3, H3K4me1, H3K27me3, H3K9me3, H3K27/9ac
- >20 more
- Open chromatin: DNA accessibility
- DNA methylation: WGBS, RRBS, MRE/MeDIP
- Gene expression: RNA-seq, Exon Arrays

Chromatin state annotations across 127 epigenomes

Reveal epigenomic variability: enh/prom/tx/repr/het

Anshul Kundaje

States show distinct mCpG, DNase, Tx, Ac profiles

TssA vs. TssBiv: diff. activity, both open, both unmethylated!
Enh vs. ReprPC: diff. activity, both intermediate DNase/Methyl
Tx: Methylated, closed, actively transcribed
⇒ Distinct modes of repression: H3K27me3 vs. DNAme vs. Het
Chromosomal ‘domains’ from chromatin state usage

• State usage → gene density, lamina, cytogenetic bands
• Quies/ZNF/het | gene rich/poor, each active/repressed

H3K4me1 phylogeny reveals common biology

• Grouping of ES, immune, brain, muscle, heart, smooth muscle, fetal

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   – Overview of epigenomics, Diversity of Chromatin modifications
   – Antibodies, ChIP-Seq, data generation projects, raw data

2. Primary data processing: Read mapping, Peak calling
   – Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   – Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)

3. Discovery and characterization of chromatin states
   – A multi-variate HMM for chromatin combinatorics
   – Promoter, transcribed, intergenic, repressed, repetitive states

4. Model complexity: selecting the number of states/marks
   – Selecting the number of states, selecting number of marks
   – Capturing dependencies and state-conditional mark independence

5. Learning chromatin states jointly across multiple cell types
   – Stacking vs. concatenation approach for joint multi-cell type learning
     – Defining activity profiles for linking enhancer regulatory networks

6. Epigenome imputation by exploiting chromatin mark correlations

Cells/Tissues at extremes of epigenomic variation

• ES/Immune/IMR90 most extreme
• ES: ↑Biv, ↓Enh/Tx/TssFlnk/PCwk
• Immune: ↓TssA, ↓TxWk
• IMR90: ↑ReprPC, ↓Quies

Chromatin state annotations across 127 epigenomes

• Reveal epigenomic variability: enh/prom/tx/repr/het
Activity-based linking of enhancers to target genes

Finding correct target of enhancer in divergently transcribed genes

Compare correlations between gene expression levels and enhancer associated histone modification signals

Visualizing 10,000s predicted enhancer-gene links

- Overlapping regulatory units, both few and many
- Both upstream and downstream elements linked
- Enhancers correlate with sequence constraint

Introducing multi-cell activity profiles

Link TFs to target enhancers

Predict activators vs. repressors

Chromatin dynamics: linking enhancer networks

TFs → enhancers → target genes

Gene expression

Chromatin States

Active TF motif enrichment

TF regulator expression

Dip-aligned motif biases

Link enhancers to target genes

- ON: Active enhancer
- OFF: Repressed
- Motif enrichment
- Motif depletion
- TF On
- TF Off
- Motif aligned
- Flat profile
**Goals for today: Computational Epigenomics**

1. **Introduction to Epigenomics**
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChiP-Seq, data generation projects, raw data

2. **Primary data processing: Read mapping, Peak calling**
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)

3. **Discovery and characterization of chromatin states**
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states

4. **Model complexity: selecting the number of states/marks**
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence

5. **Learning chromatin states jointly across multiple cell types**
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

6. **Epigenome imputation by exploiting chromatin mark correlations**

---

**Epigenomic Imputation Problem**

**Problem:** Predict mark, cell type data genome-wide assuming no data for the dataset we are trying to predict
- Complete big (mark, tissue) data matrix
- Combines potentially hundreds of datasets to generate more robust and higher quality versions of observed data sets

**Features for a mark**
- At target position and every 25bp left and right until 500bp.
- At 500bp and every 500bp left and right until 10000 bp.

**Features:**
- Average target mark signal at target position in K-nearest epigenomes for K=1,...,10
- Separate set of features for distance defined based on each mark in target epigenome and local and global distance

---

**ChromImpute: Two classes of features**

**Features for a mark**
- At target position and every 25bp left and right until 500bp.
- At 500bp and every 500bp left and right until 10000 bp.

---

**Causal motifs supported by dips & enhancer assays**

- Dip evidence of TF binding (nucleosome displacement)
- Enhancer activity halved by single-motif disruption

**Motifs bound by TF, contribute to enhancers**
ChromImpute: Training and Prediction strategy

- Assume no training data for target mark in target epigenome
- Separate regression tree(s) for each epigenome where mark is available
- Restrict features to common marks between target and informant tissue
- Apply each regression tree to target epigenome and average predictions

Imputed data is a close match to observed at multiple resolutions

- 2Mb region, 1 tissue per mark
- 200kb region, 1 tissue per mark
- 10kb region, at 25bp bins

Observed/Imputed Data at 2000 Random Positions

- Captures cell type specificity
- Dynamic changes across marks

ChromImpute Outperforms Two Stringent Baselines

- Signal Average – average of mark across all other epigenomes
- Best Case Single Epigenome – upper bound on performance when selecting one epigenome
ChromImpute Outperforms Baselines on Vast Majority of Individual Data Sets

- Better tissue coherence than observed datasets!

Imputed data capture tissue specificity/relationships

- Unbiased comparison of observed/imputed data

Imputed: Better agreement with TSS and gene annotations

+/−2kb TSS recovery with H3K4me3

Gene recovery with H3K36me3

- Existing QC metrics can fail for wrong Ab, cross-reactivity, label-swap

Predictive Performances Increases for More Broadly Expressed Genes

- Mark prioritization from imputation performance

Evaluation of performance for subset of marks/features relative to prediction with all features on deep epigenomes
Mark prioritization from imputation performance

Imputed signal data shows stronger H3K27ac-GWAS associations

Method:
• H3K27ac association for GWAS catalog (Hindorff et al, 2009)
• GWAS-Tissue association vs. all GWAS SNPs (Mann-Whitney test)
• Restrict to 98 common samples (1MB pruned)

Results: Imputed H3K27ac shows higher association than observed
• More significant P-value for most-significant tissue in each trait
• Higher total number of significant tissues across all tissues and traits

Imputation improves trait-relevant tissue association

Most significant enrichment shown for observed or imputed data
25 chromatin states from 12 marks imputed in 127 cells

Chromatin states based on ChromHMM (Ernst and Kellis, 2012)
Observed model based on 5-core marks

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data

2. Primary data processing: Read mapping, Peak calling
   - Read mapping; Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)

3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states

4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence

5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

6. Epigenome imputation by exploiting chromatin mark correlations