Lecture 10 - Epigenomics
read mapping – peak calling – multivariate HMMs

Andreas Pfenning

Module III: Epigenomics and gene regulation

• Computational Foundations
  – L10: Multi-variate HMMs, IDR, peak calling
  – L11: Gibbs Sampling: between EM and Viterbi training
  – L12: Spectral algorithms, matrix operations, linear algebra

• Biological frontiers:
  – L10: Epigenomics, chromatin states, differentiation
  – L11: Regulatory motif discovery, TF binding
  – L12: Gene networks, regulatory genomics

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
     (Future: Chromatin states to interpret disease-associated variants)

One Genome – Many Cell Types

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
Diverse epigenetic modifications

- Histone modifications provide information about what types of proteins are bound to the DNA and what the function of the region is.
- Where are enhancers, promoters, transcribed regions, and repressed regions and in what different cell types?

Diversity of epigenetic modifications

- 100+ different histone modifications
  - Histone → 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/Chrom compete

Epigenetic Mechanisms

- Histone modifications
  - H3K4me1, H3K4me3, H3K9ac, H3K36me3, H3K27me3, H3K9me3, H3K27ac
  - DNase

DNA wrapped around histone proteins

Epigenomics Roadmap across 100+ tissues/cell types

Diverse epigenomic assays:
1. Histone modifications
   - H3K4me3, H3K4me1
   - H3K36me3
   - H3K27me3, H3K27ac
   - H3K27/9ac, +20 more
2. Open chromatin:
   - DNase
3. DNA methylation:
   - WGBS, RRBS, MRE/MeDIP
4. Gene expression
   - RNA-seq, Exon Arrays

Diverse chromatin signatures encode epigenomic state

Enhancers
- H3K4me3, H3K27ac, DNase

Promoters
- H3K4me3, H3K9ac, DNase

Transcribed
- H3K36me3, H3K79me2, H4K20me1

Repressed
- H3K9me3, H3K18ac

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

- First wave published
- Lots more in pipeline
- Time for analysis!
**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

(Future: Chromatin states to interpret disease-associated variants)

---

**ChIP-seq review**

(Chromatin immunoprecipitation followed by sequencing)

**Mapping Reads to the Genome**

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

---

**Reads**

- 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.
Deep sampling of 9 reference epigenomes (e.g. IMR90)

Chromatin state+RNA+DNAse+28 histone marks+WGBS+Hi-C

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
   - (Future: Chromatin states to interpret disease-associated variants)

Quality control metrics

ChIP vs. Input DNA
Read quality
Mappability
Library complexity

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
- Read quality histograms
  - 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

ENCODE uniform processing pipeline

High quality reads
- average base score per position

Low quality reads
- location in read

• 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
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

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
   (Future: Chromatin states to interpret disease-associated variants)

QC4: Library complexity: non-redundant fraction

How many distinct uniquely mapping read? How many duplicates?
If your sample does not contain sufficient DNA and/or you over-sequencing, 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

Strand cross-correlation (CC) analysis

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 \( f \)

\[ s = \frac{f}{2} + \frac{f}{2} \]

\( f \) is the length at which ChIP DNA is fragmented
Cross-correlation at read vs. fragment length

\[
\begin{align*}
\text{Normalized strand } CC(NSC) &= \frac{CC_f}{\min(CC)} \\
\text{Relative strand } CC(RSC) &= \frac{(CC_f - \min(CC))}{(CC_r - \min(CC))}
\end{align*}
\]

• Sign of a good dataset:
  – High absolute cross-correlation at fragment length (NSC)
  – High fragment length CC relative to read length CC (RSC)

Example of good, medium, bad CC datasets

For highly enriched datasets, fragment length cross-correlation peak should be able to beat read-length phantom peak

RSC should be > 1

Where does read cross-correlation come from?

• 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+1’ is uniquely mappable on – strand
• Fragment-length peak should always dominate the read-length peak

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
(Future: Chromatin states to interpret disease-associated variants)

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.

Peak Calling

Continuous signal → Intervals
Peak calling: detect regions of enrichment

Goal: Transform read counts into normalized intensity signal

Steps:
1. Estimate fragment-length f using strand cross-correlation analysis
2. Extend each read from 5’ to 3’ direction to fragment length f
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

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

Peak calling: identify discrete intervals

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
(Future: Chromatin states to interpret disease-associated variants)

Selecting meaningful peaks using reproducibility

Use peak ranks in replicate experiments
IDR: Irreproducible Discovery Rate

http://anshul.kundaje.net/projects/idr

A. Kundaje, Q. Li, B. Brown, J. Rozowsky, S. Wilder, M. Gerstein, I. Dunham, E. Birney, P. Bick
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

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:
  \[
  (x, y) \sim pN(\mu, \sigma, \rho) + (1 - p)N(0, 1, 1, 0)
  \]
- This can be fit via an EM-like algorithm.

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

IDR leads to higher consistence between peak callers

IDR = Irreproducible Discovery Rate FDR = False Discovery Rate

- 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

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
     (Future: Chromatin states to interpret disease-associated variants)

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

Multivariate HMM for Chromatin States

- 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
  \[ P(X>c_i) < 10^{-4} \]
  where $X$ is a Poisson random variable with mean $\lambda_i$

Emission Parameter Matrix $e_k(\vec{x}_i)$

- 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

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

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

Apply genome wide to classify chromatin states de novo

Now what? Interpret these states biologically

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
   (Future: Chromatin states to interpret disease-associated variants)
Functional enrichments enable annotation of 51 distinct states

Application of ChromHMM to 41 chromatin marks in CD4+ T-cells (Barski'07, Wang'08)

Promoter upstream high avg, Potential enhancer
Promoter upstream low avg, Potential enhancer
Promoter upstream high prom, Potential enhancer
Promoter upstream low prom, Potential enhancer
Promoter upstream high avg, Potential enhancer
Promoter upstream low avg, Potential enhancer
Promoter upstream high prom, Potential enhancer
Promoter upstream low prom, Potential enhancer

Promoter vs. enhancer regulation
State 10kb away predictive of expr.

Distinct types of repression
- Chrom bands / HDAC resp
- Repeat-family / composition

Enrichment in Transcribed region
Spliced exons/5' UTR, open 3' UTR, TF binding
Spliced exons/UTR, 3' flanking, TF binding
Spliced exons/UTR, 5' flanking, TF binding
Spliced exons/UTR, 5' flanking, 3' flanking

TF binding Motif enrichment

Promoter state → gene GO function

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

Known coding
Evolutionary CBF score
Evolutionary signature: not protein-coding

Assign candidate functions to intergenic SNPs from genome-wide association studies

Discovery power for promoters, transcripts

- Significantly outperforms single-marks
- Similar power to supervised learning approach
- CAGE experiments give possible upper bound

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
   (Future: Chromatin states to interpret disease-associated variants)
State-conditional mark independence

Do hidden states actually capture dependencies between marks?

Pairwise Expected vs. Observed Mark Co-Occurrence

Test conditional independence for each state

Non-independence reveals cases of model violation

As more states are added, dependencies captured

Goals for today: Computational Epigenomics
**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

(Future: Chromatin states to interpret disease-associated variants)

**ENCODE: Study nine marks in nine human cell lines**

<table>
<thead>
<tr>
<th>9 marks</th>
<th>9 human cell types</th>
<th>81 Chromatin Mark Tracks (2^9 combinations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me1</td>
<td>HUVEC</td>
<td>Umbilical vein endothelial</td>
</tr>
<tr>
<td>H3K4me2</td>
<td></td>
<td>Keratinocytes</td>
</tr>
<tr>
<td>H3K4me3</td>
<td></td>
<td>Lymphoblastoid</td>
</tr>
<tr>
<td>H3K27ac</td>
<td></td>
<td>Myelogenous leukemia</td>
</tr>
<tr>
<td>H3K9ac</td>
<td></td>
<td>HepG2</td>
</tr>
<tr>
<td>H3K27me3</td>
<td></td>
<td>Liver carcinoma</td>
</tr>
<tr>
<td>CTCF</td>
<td></td>
<td>NHLF</td>
</tr>
<tr>
<td>+WCE</td>
<td></td>
<td>Normal human lung fibroblast</td>
</tr>
<tr>
<td>+RNA</td>
<td></td>
<td>HMEC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mammary epithelial cell</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HSSM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Skeletal muscle myoblasts</td>
</tr>
<tr>
<td>H1</td>
<td></td>
<td>Embryonic</td>
</tr>
</tbody>
</table>

Brad Bernstein ENCODE Chromatin Group

**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

**Functional recovery with increasing numbers of states**

- **Red:** Maximum fold functional enrichment for corresponding biological category
- **Blue:** Percent of that functional category that overlaps regions annotated to this state
- **Top plot:** Correlation of emission parameter vector for that state to closest state

**Chromatin state recovery with increasing numbers of marks**

- **Which states are well-recovered?** Increasing numbers of marks (greedy)
- **Precisely what mistakes are made?** (for a given subset of 11 ENCODE marks)

**State inferred with subset of marks**

**State confusion matrix with 11 ENCODE marks**

**Random initialization** (states appear & disappear)

**Selected 51-state model**

**Nested initialization** (states consistently recovered)

**Transcription End State**

**Simple Repeat state**

**Zinc Finger state**

**Step 1:** Learn a larger model that captures ‘all’ relevant states

**Step 2:** Prune down model greedily eliminating least informative states

**Step 3:** Select arbitrary cutoff based on biological interpretation

**Result:** A 51-state model that captures most biology in least complexity

- Standard model selection criteria fail due to genome complexity: more states always preferred
- Instead: Start w/complex model, keep informative states, prune redundant states. Pick cutoff

**ENCODE: Study nine marks in nine human cell lines**

Brad Bernstein ENCODE Chromatin Group

**How to learn single set of chromatin states?**

Ernst et al. Nature 2011
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 of states across multiple cell types

Epigenome imputation by exploiting mark correlations

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

Epigenomics Roadmap across 110 tissues/cell types

- ChIP-Seq: 8-20 histone marks
- Methyl: WGBS, RRBS, MeDip, MRE
- Accessibility: DNase, Footprints
- RNA: mRNA, smRNA, Exons

Integration: chromatin states, regulatory regions, hi-res, high-coverage

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

Chromatin state annotations across 127 epigenomes

- Jason Ernst

Joint learning with different subsets of marks (Solution 3)
**H3K4me1 phylogeny reveals common biology**

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

**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

**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
   - (Future: Chromatin states to interpret disease-associated variants)

**Introducing multi-cell activity profiles**

<table>
<thead>
<tr>
<th>Gene expression</th>
<th>Chromatin States</th>
<th>Active TF motif enrichment</th>
<th>TF regulator expression</th>
<th>Dip-aligned motif biases</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUVEC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHEK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM12878</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K562</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HepG2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHLF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMEC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSMM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Link enhancers to target genes**

**5. Correlation-based links of enhancer networks**

Regulators \(\rightarrow\) Enhancers \(\rightarrow\) Target genes
Activity-based linking of enhancers to target genes

Finding correct target of enhancer in divergently transcribed genes

- Compute correlations between gene expression levels and enhancer associated histone modification signals

Deep sampling of 7 reference epigenomes (e.g. IMR90)

Chromatin dynamics: linking enhancer networks

Introducing multi-cell activity profiles

- Link TFs to target enhancers
- Predict activators vs. repressors

Coordinated activity reveals activators/repressors

- Enhancer networks: Regulator → enhancer → target gene

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
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, transcriptional, 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

(Future: Chromatin states to interpret disease-associated variants)

Application: Chromatin state changes in mouse model of Alzheimer’s

Andreas Pfennig with Manolis, Li-Huei Tsai, Elizabeth Gjoneska, and Anshul Kundaje

Application: Chromatin state changes in mouse model of Alzheimer’s

Mouse AD decreasing enhancers map to human fetal brain enhancers
Mouse AD increasing enhancers map to human immune enhancers

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

Interpreting disease-association signals

Interpret variants using reference states
- Chromatin states: Enhancers, promoters, motifs
- Enrichment in individual loci, across 1000s of SNPs in T1D

Epigenome changes in disease
- Molecular phenotypic changes in patients vs. controls
- Small variation in brain methylomes, mostly genotype-driven
- 1000s of brain-specific enhancers increase methylation in Alzheimer’s

Application: Chromatin state changes in mouse model of Alzheimer’s

Mouse AD decreasing enhancers map to human fetal brain enhancers
Mouse AD increasing enhancers map to human immune enhancers

Animal model (CK-p25)
Late (6 weeks)
Control (CK)
Late (6 weeks)

Higher-activity enhancer in mouse AD model
Lower-activity enhancer in mouse AD model

Andreas Pfennig with Manolis, Li-Huei Tsai, Elizabeth Gjoneska, and Anshul Kundaje