Lecture 8 - Epigenomics
read mapping – peak calling – multivariate HMMs
<table>
<thead>
<tr>
<th>Project</th>
<th>Psets</th>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
<th>Lec</th>
<th>Topic</th>
<th>Read*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PS1</td>
<td>1</td>
<td>Thu, Sep 7</td>
<td>Introduction</td>
<td>L1</td>
<td>Intro: Biology, Algorithms, Machine Learning, Course Overview</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Sep 8</td>
<td>Recitation 1: Biology and Probability Review</td>
<td>R1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PS2</td>
<td>2</td>
<td>Tue, Sep 12</td>
<td>Alignment I: Dynamic Programming, Global and local alignment</td>
<td>L2</td>
<td></td>
<td>2,3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Sep 14</td>
<td>Recitation 2: Deriving Parameters of Alignment, Multiple Alignment</td>
<td>L3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Sep 21</td>
<td>Recitation 3: Affinity Propagation Clustering and Random Forest Classification</td>
<td>R2</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>PS4</td>
<td>4</td>
<td>Tue, Sep 26</td>
<td>Module II: Gene Expression and Epigenomics</td>
<td>L5</td>
<td>Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Sep 29</td>
<td>Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute</td>
<td>R3</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>PS5</td>
<td>5</td>
<td>Thu, Oct 3</td>
<td>Foundations II: Regulatory Genomics and Networks</td>
<td>L6</td>
<td>Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian</td>
<td>15,16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Oct 4</td>
<td>Recitation 5: Gapped Motif Discovery, DNAShape, PBMs, Selex</td>
<td>R4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PS6</td>
<td>6</td>
<td>Tue, Oct 10</td>
<td>Module III: Regulatory Genomics and Networks</td>
<td>L7</td>
<td>Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Exp</td>
<td>14,15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Oct 12</td>
<td>Recitation 6: Networks review, Recommendation systems, EHR, PhEWS</td>
<td>R5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PS7</td>
<td>7</td>
<td>Thu, Oct 17</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td>L8</td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td>20,21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Oct 19</td>
<td>Disease Association Mapping, GWAS, organismal phenotypes</td>
<td>L9</td>
<td>Deep Learning. Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder</td>
<td>20,7</td>
</tr>
<tr>
<td></td>
<td>PS8</td>
<td>8</td>
<td>Tue, Oct 24</td>
<td>Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507</td>
<td>L10</td>
<td>No Classes - Columbus Day Holiday</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Oct 26</td>
<td>Recitation 7: Linkage Disequilibrium, haplotype phasing, genotype imputation</td>
<td>R6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PS9</td>
<td>9</td>
<td>Thu, Nov 2</td>
<td>Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq</td>
<td>L11</td>
<td>Panel Discussion: reconciling critiques, strategies for improvement, feedback to author 32D-507</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 3</td>
<td>Recitation 8: Rare Variants, ExAC</td>
<td>R7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PS10</td>
<td>10</td>
<td>Thu, Nov 7</td>
<td>Comparative genomics and evolutionary signatures</td>
<td>L12</td>
<td>No Recitation, Veterans Day</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 10</td>
<td>Genome Scale Evolution, Genome Duplication</td>
<td>L13</td>
<td></td>
<td>4,5,7</td>
</tr>
<tr>
<td></td>
<td>PS11</td>
<td>11</td>
<td>Tue, Nov 14</td>
<td>No more posts (work on your final project)</td>
<td>L14</td>
<td>Quiz</td>
<td>No lecture, thanksgiving break - Thu Nov 26, 2015</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Nov 21</td>
<td>In Class Quiz (the only quiz - the class has no final exam) - covers L1-L20,R1-R9</td>
<td>L15</td>
<td>No recitation, thanksgiving break</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PS12</td>
<td>12</td>
<td>Thu, Nov 28</td>
<td>Quiz</td>
<td>L16</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 24</td>
<td></td>
<td>L17</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PS13</td>
<td>13</td>
<td>Tue, Dec 1</td>
<td>Module V: Comparative Genomics and Evolution</td>
<td>L18</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tue, Dec 5</td>
<td>Recitation 9: Phylogenetic distance metrics, Coalescent Process</td>
<td>R8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PS14</td>
<td>14</td>
<td>Thu, Dec 7</td>
<td>Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference</td>
<td>L19</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Dec 8</td>
<td>Recitation 10: Project Feedback, results, interpretation, directions</td>
<td>R9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PS15</td>
<td>15</td>
<td>Tue, Dec 12</td>
<td>Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface</td>
<td>L20</td>
<td></td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tue, Dec 15</td>
<td>Recitation 11: Presentation Tips - Intro, outline, Slides, Presentation skills</td>
<td>R10</td>
<td></td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tue, Dec 18</td>
<td>Genome Engineering with CRISPR/Cas9 and related technologies</td>
<td>L21</td>
<td></td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tue, Dec 22</td>
<td>Final Presentations - Part I (11am), 32-G8 reading room</td>
<td>L22</td>
<td></td>
<td>38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tue, Dec 25</td>
<td>Final Presentations - Part I (1pm), 32-141</td>
<td>L23</td>
<td></td>
<td>39</td>
</tr>
</tbody>
</table>

* readings refer to chapters in compiled 2016 scribe notes, available in the materials folder on Stellar
** recitation topics will be adjusted to respond to lecture and student needs
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
One Genome – Many Cell Types

ACCGTTACGACGGTCA
GGTACTGATACCCCAA
ACCGTTGACCGCATTTA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTAGCTACTGGTTTAG
CAATTTACCCTTACAAC
GTTTACAGGGTTACCGT
TGGGATTTGAAAAAAAG
TTTGAGTTGGTTTTTTC
ACGGTAGAACGTACCGT
TACCAGTA

Image Source: wikipedia
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
Three types of epigenetic modifications

**EPIGENETIC MECHANISMS**
are affected by these factors and processes:
- **Development** (in utero, childhood)
- Environmental chemicals
- Drugs/Pharmaceuticals
- Aging
- Diet

**HEALTH ENDPOINTS**
- Cancer
- Autoimmune disease
- Mental disorders
- Diabetes

**DNA methylation**
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.

**Histone modification**
The binding of epigenetic factors to histone “tails” alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

**Histones** are proteins around which DNA can wind for compaction and gene regulation.

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
Combinations of marks encode epigenomic state

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

Enhancers
- H3K4me1
- H3K27ac
- DNase

Promoters
- H3K4me3
- H3K9ac
- DNase

Transcribed
- H3K36me3
- H3K79me2
- H4K20me1

Repressed
- H3K9me3
- H3K27me3
- DNAmethyl
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)

Diverse epigenomic assays:
1. Histone modifications
   - H3K4me3, H3K4me1
   - H3K36me3
   - H3K27me3, H3K9me3
   - H3K27/9ac, +20 more
2. Open chromatin:
   - DNase
3. DNA methylation:
   - WGBS, RRBS, MRE/MeDIP
4. Gene expression
   - RNA-seq, Exon Arrays
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

6. Epigenome imputation by exploiting chromatin mark correlations
ChIP-seq review
(Chromatin immunoprecipitation followed by sequencing)
ChIP-chip and ChIP-Seq technology overview

Modification-specific antibodies $\rightarrow$ 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

ChromHMM: multi-variate hidden Markov model

Chromatin state track summary

WashU Epigenome Browser
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

...ATAGTCTTCTGCATATGTGCTTTTCTTGCACAGGACGGTAAATACACTCAACCTTTTTGTAT...
How would you do it:

• 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 $O(m)$
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?

Reference genome (> 3 gigabases) 
<table>
<thead>
<tr>
<th>Chr1</th>
<th>Chr2</th>
<th>Chr3</th>
<th>Chr4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Short read 
ACTCCCCGTACTCTAAAT

Extract seeds
Multi-seed hashing

Position N
Position 2
CTGC CGTA AACT AATG

Position 1
ACTG **** AAAC ****
**** CCCT **** TAAT
**** **** AAAC TAAT
**** **** **** ****
**** CCCT AAAC ****

Index seed pairs
Seed index (tens of gigabytes)

ACTG **** AAAC ****
**** CCCT **** TAAT
**** **** AAAC TAAT
**** **** **** ****
**** CCCT AAAC ****

Six seed pairs per read/fragment

Look up each pair of seeds in index

ACTC CCCT ACTC TAAT
1 2 3 4 5 6

Hits identify positions in genome where spaced seed pair is found

Confirm hits by checking “****” positions

Conventional hashing

Burrows-Wheeler Transform

Reference genome (> 3 gigabases)

Short read
ACTCCCCGTACTCTAAAT

Concatenate into single string

Look up ‘suffixes’ of read

Hits identify positions in genome where read is found

Convert each hit back to genome location

Report alignment to user
Burrows-Wheeler Transform (BWT)


- **Transform:** \(^{\text{BANANA}}@ \) INTO: \(\text{BNN}^{\text{AA}}@\)A

**function** BWT *(string s)*

- create a table, rows are all possible rotations of \(s\)
- sort rows alphabetically

**return** (last column of the table)

**Reversible**

**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

<table>
<thead>
<tr>
<th>Add 1</th>
<th>Sort 1</th>
<th>Add 2</th>
<th>Sort 2</th>
<th>Add 3</th>
<th>Sort 3</th>
<th>Add 4</th>
<th>Sort 4</th>
<th>Add 5</th>
<th>Sort 5</th>
<th>Add 6</th>
<th>Sort 6</th>
<th>Add 7</th>
<th>Sort 7</th>
<th>Add 8</th>
<th>Sort 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>A</td>
<td>BA</td>
<td>AN</td>
<td>AN</td>
<td>ANA</td>
<td>BANA</td>
<td>ANAN</td>
<td>BANAN</td>
<td>ANANA</td>
<td>BANANA</td>
<td>ANANA</td>
<td>BANANA</td>
<td>ANANA</td>
<td>BANANA</td>
<td>ANANA</td>
</tr>
</tbody>
</table>

Last 1st col pairs 2nd col triples 3rd col 4mers 4th col 5mers 5th col 6-mers 6th col 7-mers 7th col 8-mers Full matrix
Searching for an Exact Match

P is the input substring
C[c] – is how many characters occur before c lexicographically in the genome
Occ(c,k) is the number of occurrence of the character c before index k in the far right column

<table>
<thead>
<tr>
<th>OLIS</th>
<th>OLIS</th>
<th>OLIS</th>
<th>OLIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. $MANOLISKELLIS$</td>
<td>1. $MANOLISKELLIS$</td>
<td>1. $MANOLISKELLIS$</td>
<td>1. $MANOLISKELLIS$</td>
</tr>
<tr>
<td>2. ANOLISKELLIS$M</td>
<td>2. ANOLISKELLIS$M</td>
<td>2. ANOLISKELLIS$M</td>
<td>2. ANOLISKELLIS$M</td>
</tr>
<tr>
<td>3. ELLIS$MANOLISK</td>
<td>3. ELLIS$MANOLISK</td>
<td>3. ELLIS$MANOLISK</td>
<td>3. ELLIS$MANOLISK</td>
</tr>
<tr>
<td>4. IS$MANOLISKELL</td>
<td>4. IS$MANOLISKELL</td>
<td>4. IS$MANOLISKELL</td>
<td>4. IS$MANOLISKELL</td>
</tr>
<tr>
<td>5. ISKELLIS$MANOL</td>
<td>5. ISKELLIS$MANOL</td>
<td>5. ISKELLIS$MANOL</td>
<td>5. ISKELLIS$MANOL</td>
</tr>
<tr>
<td>6. LIS$MANOLISKEL</td>
<td>6. LIS$MANOLISKEL</td>
<td>6. LIS$MANOLISKEL</td>
<td>6. LIS$MANOLISKEL</td>
</tr>
<tr>
<td>7. LISKELLIS$MANO</td>
<td>7. LISKELLIS$MANO</td>
<td>7. LISKELLIS$MANO</td>
<td>7. LISKELLIS$MANO</td>
</tr>
<tr>
<td>8. LLIS$MANOLISKE</td>
<td>8. LLIS$MANOLISKE</td>
<td>8. LLIS$MANOLISKE</td>
<td>8. LLIS$MANOLISKE</td>
</tr>
<tr>
<td>10. MANOLISKELLIS$</td>
<td>10. MANOLISKELLIS$</td>
<td>10. MANOLISKELLIS$</td>
<td>10. MANOLISKELLIS$</td>
</tr>
<tr>
<td>11. NOLISKELLIS$MA</td>
<td>11. NOLISKELLIS$MA</td>
<td>11. NOLISKELLIS$MA</td>
<td>11. NOLISKELLIS$MA</td>
</tr>
<tr>
<td>12. OLISKELLIS$MAN</td>
<td>12. OLISKELLIS$MAN</td>
<td>12. OLISKELLIS$MAN</td>
<td>12. OLISKELLIS$MAN</td>
</tr>
<tr>
<td>14. SKELLIS$MANOLI</td>
<td>14. SKELLIS$MANOLI</td>
<td>14. SKELLIS$MANOLI</td>
<td>14. SKELLIS$MANOLI</td>
</tr>
</tbody>
</table>

Hashing vs. Burrows Wheeler Transform

Today: How does the BW transform actually work?

- Extract seeds
- Multi-seed hashing
- Index seed pairs
- Seed index (tens of gigabytes)
- Look up each pair of seeds in index
- Hits identify positions in genome where spaced seed pair is found
- Confirm hits by checking "****" positions
- Report alignment to user
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 coords
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
Quality control metrics

ChIP vs. Input DNA
Read quality
Mappability
Library complexity
ENCODE uniform processing pipeline

Signal Generation (read extension and mappability correction)

Mapped reads

Uniform Peak Calling Pipeline

Segmentation

ChromHMM/Segway

Self Organising Maps

Motif Discovery

Co-association analysis

Signal Aggregation over elements

IDR Processing, Quality control and Blacklist Filtering

Good reproducibility

Poor reproducibility

Rep2

Rep1
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
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
How many distinct uniquely mapping read? 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
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
Cross-correlation analysis

Exploiting forward and reverse reads
Fragment-length peak
Phantom read-length peak
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
**Strand cross-correlation (CC) analysis**

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

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
Cross-correlation at read vs. fragment length

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

\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*}
Where does \textit{read} cross-correlation come from?

- Input dataset (no ChIP) shows ‘phantom’ peak at \textit{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
- \textit{Fragment}-length peak should always dominate the read-length peak
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

\[ \text{RSC should be } > 1 \]
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 $f$ 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
Peak calling: identify discrete intervals

<table>
<thead>
<tr>
<th>Program</th>
<th>Reference Version</th>
<th>Graphical user interface?</th>
<th>Window-based scan</th>
<th>Tag clustering</th>
<th>Gaussian kernel density</th>
<th>Strand-specific scoring</th>
<th>Peak height of fold enrichment (FE)</th>
<th>Background subtraction</th>
<th>Compensates for genomic duplications of deletions</th>
<th>False Discovery Rate</th>
<th>Compare to normalized control data (FE)</th>
<th>Compare to normalized control data fitted with control data</th>
<th>Statistical model or test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CisGenome</td>
<td>28 1.1</td>
<td>X*</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>conditional binomial model</td>
<td></td>
</tr>
<tr>
<td>Minimal ChipSeq Peak Finder</td>
<td>16 2.0.1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>chromosome scale Poisson dist.</td>
<td></td>
</tr>
<tr>
<td>E-RANGE</td>
<td>27 3.1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>chromosome scale Poisson dist.</td>
<td></td>
</tr>
<tr>
<td>MACS</td>
<td>13 1.3.5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>local Poisson dist.</td>
<td></td>
</tr>
<tr>
<td>QuEST</td>
<td>14 2.3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X**</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>chromosome scale Poisson dist.</td>
<td></td>
</tr>
<tr>
<td>HPeak</td>
<td>29 1.1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Hidden Markov Model</td>
<td></td>
</tr>
<tr>
<td>Sole-Search</td>
<td>23 1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>One sample t-test</td>
<td></td>
</tr>
<tr>
<td>PeakSeq</td>
<td>21 1.01</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>conditional binomial model</td>
<td></td>
</tr>
<tr>
<td>SISSRS</td>
<td>32 1.4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>conditional binomial model</td>
<td></td>
</tr>
<tr>
<td>spp package (wtd &amp; mtc)</td>
<td>31 1.7</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>conditional binomial model</td>
<td></td>
</tr>
</tbody>
</table>

X* = Windows-only GUI or cross-platform command line interface
X** = optional if sufficient data is available to split control data
X' = method excludes putative duplicated regions, no treatment of deletions

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
  \[ P(count = x) = \frac{\lambda_{\text{local}}^x \exp(-\lambda_{\text{local}})}{x!} \]
- \( \lambda_{\text{local}} = \max(\lambda_{\text{BG}}, [\lambda_{1k}, \lambda_{5k}, \lambda_{10k}]) \) estimated from control data
  - Poisson p-value = \( P(count \geq x) \)
  - 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

(at FDR = 1% cutoff)
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

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

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

\[(x, y) \sim pN(\mu, \mu, \sigma, \sigma, \rho) + (1 - p)N(0,0,1,1,0)\]

- This can be fit via an EM-like algorithm.
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
What if we don’t have good 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

ChromHMM: multi-variate hidden Markov model

WashU Epigenome Browser
Multivariate HMM for Chromatin States

- Observed chromatin marks. Called based on a poisson distribution.
- Most likely Hidden State
- High Probability Chromatin Marks in State
- All probabilities are learned from the data

Ernst and Kellis
Nature Biotech 2010
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 $(p_{k1},...,p_{kM})$ 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 $e_k(\mathbf{x}_i)$

- 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

Ernst and Kellis
Nature Biotech 2010
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 \textit{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

6. Epigenome imputation by exploiting chromatin mark correlations
<table>
<thead>
<tr>
<th>State definitions</th>
<th>State Enrichments</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Chromatin mark frequencies for each chromatin state</td>
<td>b. Genomic and functional enrichments for each state</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State</th>
<th>Percentage of Genes</th>
<th>% of 3' UTR</th>
<th>% of ITSS</th>
<th>% of 3' UTR Exon</th>
<th>% of 5' SPlice Site</th>
<th>% of 5' UTR</th>
<th>% of ITSS</th>
<th>% of 5' UTR Exon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.14</td>
<td>3.0</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>2</td>
<td>0.16</td>
<td>4.2</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>0.14</td>
<td>4.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>0.10</td>
<td>5.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>5</td>
<td>0.11</td>
<td>6.4</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>6</td>
<td>0.11</td>
<td>7.5</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>7</td>
<td>0.08</td>
<td>8.9</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>8</td>
<td>0.07</td>
<td>9.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>9</td>
<td>0.06</td>
<td>9.5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>10</td>
<td>0.08</td>
<td>10.2</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

(see Supplementary Fig 2 for full emission prob matrix)
Functional enrichments enable annotation of 51 distinct states.
Application of ChromHMM to 41 chromatin marks in CD4+ T-cells (Barski'07, Wang'08)

- Promoter states:
  - Promoter upstream high expr; Potential enh looping
  - Promoter upstream med expr; Potential enh looping
  - Promoter upstream low expr; Potential enh looping
  - Repressed promoter
  - TSS low-med expr; most GC rich
  - TSS med expr
  - TSS high expr
  - Transcribed promoter; highest expr, TSS for active genes
  - Transcribed promoter; highest expr, downstream
  - Transcribed promoter; high expr, near TSS
  - Transcribed promoter; high expr, downstream
  - Transcribed 5' proximal; higher expr, open chr, TF binding
  - Transcribed 5' proximal, higher expr
  - Transcribed 5' proximal, high expr, open chr, TF binding
  - Transcribed 5' proximal, high expr
  - Transcribed 5' proximal, med expr; Alu repeats
  - Transcribed less 5' proximal, med expr; open chr
  - Transcribed less 5' proximal, med expr
  - Transcribed less 5' proximal, lower expr; Alu repeats
  - Enhancer in Transcribed region
  - Spliced exons/GC Rich; open chr, TF binding
  - Spliced exons/GC Rich
  - Spliced exons/GC Rich; Alu repeats
  - Transcribed 5' distal; exons
  - Transcribed Further 5' distal; exons
  - Transcribed 5' distal; Alu repeats
  - End of Transcription; exons; high expr
  - ZNF Genes; KAP-1 repressed state

- Transcribed Intergenic:
  - Candidate strong distal enh; higher open chr; higher target expr
  - Candidate strong enh; high open chr; higher target expr
  - Intergenic H2AZ with open chr/TF binding. Candidate distal enhancer
  - Candidate weak distal enhancer
  - Candidate distal enhancer
  - Proximal to active enhancers; Alu repeats
  - Active intergenic regions not enhancer specific
  - Active intergenic further from enhancers; Alu repeats
  - Non-repressive intergenic domains; Alu repeats
  - H2AZ specific state
  - CTCF Island

- Active Intergenic:
  - Unmappable
  - Heterochr; Nuclear Lamina; Most AT rich
  - Heterochr; Nuclear Lamina; ERVL repeats
  - Heterochr; Lower gene depletion
  - Heterochr; ERVL repeats; Lower gene/exon depletion

- Repressed Intergenic:
  - Simple repeats (CA)n, (TG)n
  - L1/LTR Repeats
  - Satellite Repeat
  - Satellite Repeat; moderate mapping bias
  - Satellite Repeat; high mapping bias
  - Satellite Repeat/rRNA; extreme mapping bias

- Repressed Transcribed:
  - Heterochr; Nuclear Lamina, ERVL repeats
  - Heterochr; Lower gene depletion
  - Heterochr; ERVL repeats; Lower gene/exon depletion

- Repressed Promoter:
  - Specific Repression

- Repressed TSS:
  - CTCF Island

- Repressed Enhancer:

- Repressed Spliced exons/GC Rich:

- Repressed Transcribed state:
Functional properties of discovered chromatin states

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 …" MCB 2006.

State 27

Promoter state → gene GO function

Transcription End State

ZNF repressed state recovery

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

TF binding

Motif enrichment

Promoter vs. enhancer regulation

Intergenic state association with level of downstream gene expression

State 30

State 10kb away predictive of expr.
Applications to genome annotation

New protein-coding genes

In promoter (short)/low-expr states

Chromatin signature:
- Promoter / transcribed

Evolutionary signature:
- Not protein-coding

Long intergenic non-coding RNAs/lincRNAs

Known coding

Evolutionary CSF score

Evolutionary signature:
- Not protein-coding

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

Bing Ren, Eddy Rubin
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

6. Epigenome imputation by exploiting chromatin mark correlations
State-conditional mark independence

Do hidden states actually capture dependencies between marks?
Pairwise Expected vs. Observed Mark Co-Occurrence

Each point = one pair of chromatin marks
41x41 pairs plotted
X-axis: $F(\text{mark1}) \times F(\text{mark2})$
Y-axis: $F(\text{mark1} & \text{mark2})$
Diagonal: independence
Off-diag: dependence

 Marks become conditionally independent
 Model captures dependencies

$\mathbf{p}_i$ emission prob for mark $i$

$\mathbf{q}_{i,j}$ freq w/ which marks $i$ and $j$ co-occur

$\mathbf{P}_i$

$\mathbf{p}_j$

$\mathbf{q}_{i,j} = \mathbf{p}_i \times \mathbf{p}_j$

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

- Repetitive states show more dependencies
- Conditional independence does not hold
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
   – 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

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
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 Inferred with all 41 marks

Recovery of states with increasing number of marks

State confusion matrix with 11 ENCODE 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
### ENCODE: Study nine marks in nine human cell lines

<table>
<thead>
<tr>
<th>9 marks</th>
<th>9 human cell types</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me1</td>
<td>HUVEC Umbilical vein endothelial</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>NHEK Keratinocytes</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>GM12878 Lymphoblastoid</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>K562 Myelogenous leukemia</td>
</tr>
<tr>
<td>H3K9ac</td>
<td>HepG2 Liver carcinoma</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>NHLF Normal human lung fibroblast</td>
</tr>
<tr>
<td>H4K20me1</td>
<td>HMEC Mammary epithelial cell</td>
</tr>
<tr>
<td>H3K36me3</td>
<td>HSMM Skeletal muscle myoblasts</td>
</tr>
<tr>
<td>CTCF</td>
<td>H1 Embryonic</td>
</tr>
<tr>
<td>+WCE</td>
<td></td>
</tr>
<tr>
<td>+RNA</td>
<td></td>
</tr>
</tbody>
</table>

**81 Chromatin Mark Tracks**

(2^81 combinations)


---

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

Brad Bernstein ENCODE Chromatin Group

---

**Chromatin Mark Observation Frequency (%)**

### Solution 1: Learn independent models and cluster

<table>
<thead>
<tr>
<th></th>
<th>CTCF</th>
<th>H3K27me3</th>
<th>H3K4me1</th>
<th>H3K4me2</th>
<th>H3K4me3</th>
<th>H3K79ac</th>
<th>WCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.2</td>
<td>72.0</td>
<td>0.2</td>
<td>9.1</td>
<td>47.9</td>
<td>77.8</td>
<td>49.5</td>
</tr>
<tr>
<td>2</td>
<td>11.9</td>
<td>1.5</td>
<td>6.1</td>
<td>5.0</td>
<td>52.7</td>
<td>93.7</td>
<td>55.0</td>
</tr>
<tr>
<td>3</td>
<td>16.4</td>
<td>1.5</td>
<td>2.4</td>
<td>5.5</td>
<td>17.0</td>
<td>92.6</td>
<td>99.0</td>
</tr>
<tr>
<td>4</td>
<td>14.4</td>
<td>0.6</td>
<td>14.5</td>
<td>11.3</td>
<td>96.3</td>
<td>99.3</td>
<td>75.1</td>
</tr>
<tr>
<td>5</td>
<td>5.3</td>
<td>0.2</td>
<td>9.5</td>
<td>2.6</td>
<td>89.1</td>
<td>56.8</td>
<td>5.3</td>
</tr>
<tr>
<td>6</td>
<td>6.7</td>
<td>0.9</td>
<td>1.0</td>
<td>3.2</td>
<td>58.3</td>
<td>74.7</td>
<td>8.4</td>
</tr>
<tr>
<td>7</td>
<td>1.6</td>
<td>0.6</td>
<td>1.6</td>
<td>1.3</td>
<td>56.5</td>
<td>2.7</td>
<td>4.9</td>
</tr>
<tr>
<td>8</td>
<td>91.5</td>
<td>1.8</td>
<td>0.9</td>
<td>2.8</td>
<td>6.3</td>
<td>3.3</td>
<td>0.4</td>
</tr>
<tr>
<td>9</td>
<td>4.6</td>
<td>0.3</td>
<td>43.2</td>
<td>43.1</td>
<td>38.5</td>
<td>36.5</td>
<td>11.5</td>
</tr>
<tr>
<td>10</td>
<td>1.2</td>
<td>0.1</td>
<td>47.2</td>
<td>2.7</td>
<td>0.4</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td>11</td>
<td>0.4</td>
<td>0.1</td>
<td>2.7</td>
<td>1.7</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>12</td>
<td>0.9</td>
<td>26.8</td>
<td>0.0</td>
<td>2.1</td>
<td>0.4</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>13</td>
<td>0.2</td>
<td>0.4</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.4</td>
</tr>
<tr>
<td>14</td>
<td>21.9</td>
<td>27.9</td>
<td>19.1</td>
<td>41.0</td>
<td>5.7</td>
<td>4.8</td>
<td>25.9</td>
</tr>
<tr>
<td>15</td>
<td>85.2</td>
<td>85.1</td>
<td>91.5</td>
<td>88.4</td>
<td>76.2</td>
<td>76.9</td>
<td>90.8</td>
</tr>
</tbody>
</table>

**Promoter**

- Train a k-state model in each cell type independently.
- Cluster models learned independently.
- Merge clusters and re-apply to each cell type.

**Candidate enhancer**

- Using emission probability matrix: most similar definitions.
- Using genome annotation: posterior probability decoding.

**Insulator**

**Transcribed**

**Repressive**

**Repetitive**

---

**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.
Joint learning of states across multiple cell types

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

9 human cell types

<table>
<thead>
<tr>
<th>9 marks</th>
<th>HUVEC</th>
<th>NHEK</th>
<th>GM12878</th>
<th>K562</th>
<th>HepG2</th>
<th>NHLF</th>
<th>HMEC</th>
<th>HSMM</th>
<th>H1</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me1</td>
<td>Umbilical vein endothelial</td>
<td>Keratinocytes</td>
<td>Lymphoblastoid</td>
<td>Myelogenous leukemia</td>
<td>Liver carcinoma</td>
<td>Normal human lung fibroblast</td>
<td>Mammary epithelial cell</td>
<td>Skeletal muscle myoblasts</td>
<td>Embryonic</td>
</tr>
<tr>
<td>H3K4me2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K4me3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K27ac</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K9ac</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K27me3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K27me3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTCF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+WCE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+RNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

81 Chromatin Mark Tracks (2^81 combinations)

Concatenation approach:
- Learned jointly across cell types
- State definitions are common
- State locations are dynamic

Brad Bernstein ENCODE Chromatin Group

Candidate state annotation

<table>
<thead>
<tr>
<th>Candidate state annotation</th>
<th>Coverage</th>
<th>Median</th>
<th>Median</th>
<th>+/-2kb TSS</th>
<th>Conserved</th>
<th>DNase (K562)</th>
<th>C-Myc (K562)</th>
<th>NF-kB (GM12878)</th>
<th>Transcript</th>
<th>Nuclear Lamina (NHLF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Promoter</td>
<td>0.6</td>
<td>0.5</td>
<td>1.3</td>
<td>0.4</td>
<td>1.0</td>
<td>83.3</td>
<td>23.8</td>
<td>1.0</td>
<td>0.2</td>
<td>0.15</td>
</tr>
<tr>
<td>Weak Promoter</td>
<td>0.5</td>
<td>1.2</td>
<td>1.3</td>
<td>0.4</td>
<td>0.6</td>
<td>58</td>
<td>2.8</td>
<td>15.8</td>
<td>0.6</td>
<td>0.30</td>
</tr>
<tr>
<td>Inactive/posied Promoter</td>
<td>0.4</td>
<td>1.0</td>
<td>1.0</td>
<td>0.6</td>
<td>1.2</td>
<td>49</td>
<td>4.3</td>
<td>10.8</td>
<td>0.1</td>
<td>0.68</td>
</tr>
<tr>
<td>Strong enhancer</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>1.2</td>
<td>49</td>
<td>23.7</td>
<td>23.1</td>
<td>1.3</td>
<td>0.05</td>
</tr>
<tr>
<td>Strong enhancer</td>
<td>0.6</td>
<td>1.2</td>
<td>1.3</td>
<td>0.4</td>
<td>1.2</td>
<td>31.8</td>
<td>23.1</td>
<td>1.3</td>
<td>0.05</td>
<td>0.10</td>
</tr>
<tr>
<td>Weak/posied enhancer</td>
<td>0.5</td>
<td>1.3</td>
<td>1.3</td>
<td>0.4</td>
<td>1.2</td>
<td>6.3</td>
<td>15.8</td>
<td>1.4</td>
<td>0.10</td>
<td>0.2</td>
</tr>
<tr>
<td>Weak/posied enhancer</td>
<td>0.4</td>
<td>1.4</td>
<td>1.2</td>
<td>0.4</td>
<td>1.2</td>
<td>6.3</td>
<td>15.8</td>
<td>1.4</td>
<td>0.10</td>
<td>0.2</td>
</tr>
<tr>
<td>Insulator</td>
<td>0.7</td>
<td>0.3</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>4.3</td>
<td>1.1</td>
<td>4.5</td>
<td>0.8</td>
<td>0.02</td>
</tr>
<tr>
<td>Transcriptional transition</td>
<td>0.6</td>
<td>0.3</td>
<td>0.6</td>
<td>0.4</td>
<td>0.3</td>
<td>4.3</td>
<td>1.1</td>
<td>4.5</td>
<td>0.8</td>
<td>0.02</td>
</tr>
<tr>
<td>Transcriptional elongation</td>
<td>0.7</td>
<td>1.3</td>
<td>1.3</td>
<td>0.4</td>
<td>1.2</td>
<td>12.5</td>
<td>1.3</td>
<td>2.6</td>
<td>0.19</td>
<td>0.2</td>
</tr>
<tr>
<td>Weak transcribed</td>
<td>0.5</td>
<td>0.3</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>4.3</td>
<td>1.1</td>
<td>4.5</td>
<td>0.8</td>
<td>0.02</td>
</tr>
<tr>
<td>Polycomb-repressed</td>
<td>0.4</td>
<td>0.3</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>4.3</td>
<td>1.1</td>
<td>4.5</td>
<td>0.8</td>
<td>0.02</td>
</tr>
<tr>
<td>Heterochrom.; low signal</td>
<td>0.1</td>
<td>0.9</td>
<td>0.1</td>
<td>0.0</td>
<td>0.6</td>
<td>7.1</td>
<td>1.0</td>
<td>10.0</td>
<td>0.1</td>
<td>0.7</td>
</tr>
<tr>
<td>Repetitive/CNV</td>
<td>0.6</td>
<td>0.3</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>4.3</td>
<td>1.1</td>
<td>4.5</td>
<td>0.8</td>
<td>0.02</td>
</tr>
<tr>
<td>Repetitive/CNV</td>
<td>0.4</td>
<td>0.3</td>
<td>0.7</td>
<td>0.8</td>
<td>0.8</td>
<td>4.3</td>
<td>1.1</td>
<td>4.5</td>
<td>0.8</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Chromatin Mark Observation Frequency (%), (fold), (kb), (%) Functional enrichments (fold)
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...
Epigenomic 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, H3K36me3
  - 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 $\Rightarrow$ 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, and fetal
Cells/Tissues at extremes of epigenomic variation

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

Misha Bilenky, Wouter Meuleman
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
5. Correlation-based links of enhancer networks

Regulators $\rightarrow$ Enhancers $\rightarrow$ Target genes
Chromatin state annotations across 127 epigenomes

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

Anshul Kundaje
2.3M enhancer regions ⇔ only ~200 activity patterns
Introducing multi-cell activity profiles

Link enhancers to target genes

Gene expression

Chromatin States

Active TF motif enrichment

TF regulator expression

Dip-aligned motif biases

HUVEC
NHEK
GM12878
K562
HepG2
NHLF
HMEC
HSMM
H1

ON
Active enhancer
Motif enrichment
TF On
Motif aligned

OFF
Repressed
Motif depletion
TF Off
Flat profile
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
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
Chromatin dynamics: linking enhancer networks

TFs $\rightarrow$ enhancers $\rightarrow$ target genes
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>
</tr>
</thead>
<tbody>
<tr>
<td>HUVEC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHEK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM12878</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K562</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HepG2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHLF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMEC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSMM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Link TFs to target enhancers
Predict activators vs. repressors

- **ON**: Active enhancer
- **OFF**: Repressed
- **Motif enrichment**
- **Motif depletion**
- **TF On**
- **TF Off**
- **Motif aligned**
- **Flat profile**
Coordinated activity reveals activators/repressors

Enhancer activity

Activity signatures for each TF

- Enhancer networks: Regulator $\rightarrow$ enhancer $\rightarrow$ target gene
Regulatory motifs predicted to drive enhancer modules

- Activator and repressor motifs consistent with tissues

Pouya Kheradpour
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
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

![Epigenomic Imputation Problem](image)

Ernst and Kellis, *Nature Biotech* 2015
ChromImpute: Two classes of features

Other marks in same tissue

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

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

Same mark in other tissues
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
Imputed data is a close match to observed at multiple resolutions.

- 200kb region, 1 tissue per mark
Imputed data is a close match to observed at multiple resolutions

- 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
Imputed data capture tissue specificity/relationships

- Better tissue coherence than observed datasets!

AUC for Correlation Classifying Pairs of Experiments as the Same Group (Excluding Other and ENCODE)
Imputed: Better agreement with TSS and gene annotations

+/-2kb TSS recovery with H3K4me3

Gene recovery with H3K36me3

- Unbiased comparison of observed/imputed data
Observed/imputed discrepancy ➔ Flag low-quality data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Reference AUC 5%</th>
<th>Read Depth</th>
<th>Poison</th>
<th>SPOT</th>
<th>FindPeaks</th>
<th>NSC</th>
<th>BSC</th>
<th>Impute Top 1%</th>
<th>Impute correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>E081</td>
<td>127</td>
<td>68</td>
<td>125</td>
<td>126</td>
<td>126</td>
<td>126</td>
<td>125</td>
<td>127</td>
<td>127</td>
</tr>
<tr>
<td>E113</td>
<td>126</td>
<td>111</td>
<td>86</td>
<td>121</td>
<td>124</td>
<td>115</td>
<td>102</td>
<td>126</td>
<td>121</td>
</tr>
<tr>
<td>E083</td>
<td>125</td>
<td>118</td>
<td>102</td>
<td>120</td>
<td>122</td>
<td>124</td>
<td>127</td>
<td>125</td>
<td>125</td>
</tr>
<tr>
<td>E002</td>
<td>124</td>
<td>127</td>
<td>42</td>
<td>59</td>
<td>78</td>
<td>82</td>
<td>1</td>
<td>124</td>
<td>126</td>
</tr>
<tr>
<td>E004</td>
<td>123</td>
<td>28</td>
<td>122</td>
<td>122</td>
<td>120</td>
<td>116</td>
<td>91</td>
<td>114</td>
<td>106</td>
</tr>
<tr>
<td>E106</td>
<td>122</td>
<td>115</td>
<td>106</td>
<td>106</td>
<td>114</td>
<td>106</td>
<td>96</td>
<td>122</td>
<td>122</td>
</tr>
<tr>
<td>E065</td>
<td>121</td>
<td>89</td>
<td>117</td>
<td>113</td>
<td>108</td>
<td>112</td>
<td>103</td>
<td>113</td>
<td>115</td>
</tr>
<tr>
<td>E109</td>
<td>120</td>
<td>122</td>
<td>103</td>
<td>102</td>
<td>115</td>
<td>109</td>
<td>122</td>
<td>123</td>
<td>123</td>
</tr>
<tr>
<td>E096</td>
<td>119</td>
<td>86</td>
<td>99</td>
<td>111</td>
<td>112</td>
<td>95</td>
<td>34</td>
<td>119</td>
<td>93</td>
</tr>
<tr>
<td>E007</td>
<td>118</td>
<td>28</td>
<td>104</td>
<td>112</td>
<td>110</td>
<td>90</td>
<td>58</td>
<td>116</td>
<td>94</td>
</tr>
<tr>
<td>Median</td>
<td>100</td>
<td>104</td>
<td>113</td>
<td>115</td>
<td>111</td>
<td>111</td>
<td>99</td>
<td>123</td>
<td>122</td>
</tr>
</tbody>
</table>

- Dataset Rank for each QC metric

Observation-Imputed Agreement

- Existing QC metrics can fail for wrong Ab, cross-reactivity, label-swap
Predictive Performances Increases for More Broadly Expressed Genes

Expressed level RPKM >= 0.5
Mark prioritization from imputation performance

<table>
<thead>
<tr>
<th>Mark/Feature Set</th>
<th>H3K27me3</th>
<th>H3K36me3</th>
<th>H3K4me1</th>
<th>H3K4me3</th>
<th>H3K27ac</th>
<th>H3K9ac</th>
<th>DNase</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Tier 1-3 Features</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Same Same Sample Features Only</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Same Mark Features Only</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Core5</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Core + H3K18ac</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Core + H3K27ac</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Core + H3K9ac</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Core + DNase</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Core + H3K27ac (same sample only)</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Core + H3K9ac (same sample only)</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Core + DNase (same sample only)</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Core (same epig only)</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Tier 1 and 2 Marks</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Tier 1 and 2 Marks (same sample only)</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1 1 1 1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Mark/Feature Set</th>
<th>H3K7me3</th>
<th>H3K36me3</th>
<th>H3K4me3</th>
<th>H3K4m1</th>
<th>H3K36ac</th>
<th>DNase</th>
<th>H2A.Z</th>
<th>H2B.Z</th>
<th>H4K20me1</th>
<th>H4K20me2</th>
<th>H4K20me3</th>
<th>H4K20me1me2</th>
<th>H4K20me2me3</th>
<th>H4K20me3me4</th>
<th>H3K4me1</th>
<th>H3K4me3</th>
<th>H3K27ac</th>
<th>H3K9ac</th>
<th>H3K79me2</th>
<th>H3K18ac</th>
<th>H3K18ac + H3K79me2</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Tier 1-3 Features</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Same Sample Features Only</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Same Mark Features Only</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Core5</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Core + H3K18ac</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Core + H3K27ac</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Core + H3K9ac</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Core + DNase</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Core + H3K27ac (same sample only)</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Core + H3K9ac (same sample only)</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Core + DNase (same sample only)</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Core (same epig only)</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Tier 1 and 2 Marks</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
<tr>
<td>Tier 1 and 2 Marks (same sample only)</td>
<td>90.36</td>
<td>91.86</td>
<td>92.24</td>
<td>92.52</td>
<td>93.03</td>
<td>93.67</td>
<td>94.07</td>
<td>94.17</td>
<td>94.26</td>
<td>94.52</td>
<td>94.89</td>
<td>95.06</td>
<td>95.12</td>
<td>95.28</td>
<td>95.41</td>
<td>95.44</td>
<td>95.49</td>
<td>95.58</td>
<td>95.61</td>
<td>95.70</td>
<td>95.71</td>
</tr>
</tbody>
</table>

H3K18ac + H3K79me2 more informative for most mark imputations than core set in a new cell type given an existing roughly uniform coverage compendium
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

Imputed H3K27ac Association (-log_{10} P)

Observed H3K27ac Association (-log_{10} P)
Imputation improves trait-relevant tissue association

Most significant enrichment shown for observed or imputed data.
Imputation improves trait-relevant tissue association
Significant Sample-Study Combinations Additional Marks

H3K4me1
- Imputed
- Observed
- Randomized

H3K4me3

Number of Combinations

H3K9ac
- 62 samples

DNase
- 53 samples

H3K27me3

Number of Combinations

-log_{10} p-val
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
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
GWAS hits in enhancers of relevant cell types
Linking traits to their relevant cell/tissue types

Liver
- Cholesterol total
- Lipid metabolism phenotypes
- Mean platelet volume
- Platelet counts
- Digestive
- Ulcerative colitis
- Inflammatory bowel
- Multiple sclerosis
- Rheumatoid arthritis
- Primary biliary cirrhosis
- Chronic lymphocytic leukemia
- Systemic lupus erythematosus

Brain
- Attention deficit hyperactivity disorder
- Alzheimer's (late onset)
- Red blood cell traits

Heart
- PR heart repolarization interval
- Type 1 diabetes autoantibodies
- Crohn's
- Type 1 diabetes

T cells
- Celiac and Rheumatoid arthritis
- Self-reported allergy

B cells
- Celiac
- Rheumatoid arthritis
- Primary biliary cirrhosis
- Chronic lymphocytic leukemia
- Systemic lupus erythematosus

ES cell
- Fibroblast
- Keratinocyte
- Prefrontal Cortex
- Cingulate Gyrus

Liver
- HDL cholesterol
- LDL cholesterol
- Liver enzyme levels (g-glut. transferase)
- Urate levels

Digestive
- Fetal Intestine
- Sigmoid Colon
- Fetal Bladder
- Rectal Mucosa
- Small Intestine
- Stomach Mucosa
- Duodenum Mucosa
- Colonic Mucosa
- Left Ventricle

Tissue types
- Adipocyte
- Adipose Nuclei
- Fetal Placenta
- Adult Liver
- Lung
- Melanocyte
- Fibroblast
- Keratinocyte
- ES cell
- ES cell

Other terms
- CD3
- CD4
- CD8
- CD14
- CD15
- CD19
- CD34
- Mean corpuscular volume
HaploReg: systematic mining of GWAS variants

• Start with any list of SNPs or select a GWA study
  – Mine ENCODE and Roadmap epigenomics data for hits
  – Hundreds of assays, dozens of cells, conservation, motifs
  – Report significant overlaps and link to info/browser

• Try it out: http://compbio.mit.edu/HaploReg

Ward, Kellis NAR 2011