Epigenetic Reduction
Marco Cusumano-Towner, Jason Ma, Diego Munoz, Tomas Rube
Introduction: The "Histone Code"

- The DNA is coiled around nucleosomes, each consisting of four histones. e.g. H2, H3...
- Histones have tails sticking out named e.g. K4, K9, etc.
- Tails can be modified, e.g. 2 Methylation: K4me2 1 Acetylation: K4ac
- Markers are involved in gene regulation, differentiation...
- Associated with regions:
  - H3K4me3 - promoter
  - H3K4me2 - enhancer
  - H3K9ac - active regulatory
Introduction: The "Histone Code"

- Histone modifications are read using ChIP-seq, one marker at a time. Expensive!

- Markers are correlated.

- Measuring some markers may be redundant if they can be inferred from other measurements
Our Problem

1. Find a minimum set of ChIP-Seq experiments needed to recover all histone modifications at a specified accuracy.

2. Quantify how accurately we can reconstruct the biologically relevant "chromatin states" from our reduced set of experiments.
Prior Work

• Ernst and Kellis propose a system of 'chromatin states'
  ◦ Each consists of a set of markers that corresponds to some biological function

• To discover these states, Ernst and Kellis trained a hidden Markov model
  ◦ 8 histone markers yielded 15 distinct states

• Each state fell into several groups:
  ◦ Promoter
  ◦ Transcribed
  ◦ Active intergenic
  ◦ Repressive
  ◦ Repetitive

The Data: 22 markers, 2 cell lines

Acetylated

Methylated

H3K4ac
H2BK12ac
H2BK120ac
H2BK15ac
H2BK20ac
H2AK5ac
H3K18ac
H4K91ac
H4K5ac
H3K56ac
H3K23ac
H3K27ac
H3K9ac
H3K4me1
H3K4me2
H3K4me3
H3K9me3
H3K27me3
H3K36me3
H3K79me1
H3K79me2
H4K20me1
Our Method: Linear Regression

- Treat each position independently, indexed by $a$
- Goal: predict the values of a set of $n$ markers $y^{(a)}_i$, $i = 1...n$, given the values of the other $(22-n)$ markers $x^{(a)}_j$.

- Model: $y_i(x) = \sum_j \beta_{ij} x_j, \quad x_0=1$
- Least squares solution for $\beta$, in matrix notation:
  \[
  \beta = (x^T x)^{-1} x^T y
  \]
- We normalize the data first, so all markers are considered equally.
- Error measured by:
  \[
  \varepsilon_i^2 = \sum_a (y^{(a)}_i - y_i(x^{(a)}))^2
  \]
Great news, everyone: lin-reg works fine!

Example: H3K4ac has $\varepsilon = 0.27$ for IMR90 cell line

Reconstruction errors for markers in IMR90
Reconstructing multiple markers

• Linear regression is used to reconstruct multiple markers
• Problem: find subset of markers with reconstruction errors \( \{\varepsilon_1, \varepsilon_2, \ldots\} \) such that the max error is minimized.
• Large number of possible combinations
• Two strategies for finding optimal combination of markers:
  ○ **Exhaustive search** up to 7 markers (170,544 combinations)
  ○ **Greedy search**: remove markers one at a time, picking marker with least error.
Error as a function of markers removed:

Max($\varepsilon$)

Exhaustive search to level 7 gave same errors
Results: Reconstruct the 22 markers
Results: Consistency between cell types and chromosomes

Linear Regression parameters trained in one cell line are general enough to be predictive for another.
Results: Reconstruct the 8 markers

Optimal markers to reconstruct for marker subsets, for $k=1,...,8$ and across different chromosomes (17,...,22) and cell-lines (IMR90 & H1)
Predicting chromatin states

Using Ernst et al. (2011)'s HMM

Emission matrix

Transition matrix

Images from Ernst et al (2011)
Predicting chromatin states

1. Binarize Marker Data

2. Run HMM
   - Using all markers
   - With each marker removed from HMM
   - With reconstructed marker values

3. Compare results
State confusion matrices

When each marker is removed:

- H3K27ac
- H3K27me3
- H3K36me3
- H3K4me1
- H3K4me2
- H3K4me3
- H3K9ac
- H4K20me1

"H3K4me2 associated with promoters and enhancers"
"H3K4me1 associated with enhancers"
"H3K36me3 and H4K20me1 associated with transcribed regions"
"H3K27me3 associated with polycomb repressed states"

Mostly diagonal

H3K9ac is least essential
Good Reconstruction Helps

H3K27ac binary
reconstructed binary

posterior with all markers

posterior with H3K27ac removed

posterior with H3K27ac reconstructed

KL-div. from all-markers posterior with H3K27ac removed with H3K27ac reconstructed
Bad Reconstruction Hurts

H4K20me1 reconstructed

H4K20me1 binary reconstructed binary

posterior with all markers

posterior with H4K20me1 removed

posterior with H4K20me1 reconstructed

KL-div. from all-markers posterior with H4K20me1 removed with H4K20me1 reconstructed
Conclusions

Findings
• Histone markers are correlated
• Linear regression provides robust results
• A subset of the markers in Ernst et al. can be used to calculate states

Future work
• Binarized data, logistic regression, neural networks
• Position-dependent models?
Thank you!

AND THAT’S MY LAST SLIDE. ANY COMMENTS?
Our goals with respect to prior work

We run our algorithms on both:

1. The full set of 22 markers, in an agnostic manner

2. The set of 8 markers used in Ernst (2011)
   - Prior work showed that eight markers suffice to annotate the genome
   - Can we further reduce the number of markers?
Correlation matrix

Correlations persist when averaging over full chromosome
Results: Greedy Reconstruction error

Errors of removed markers if removed individually

Max(\(\sigma\))

\(Nr\)
Results: Reconstruct the 22 markers

Binary reconstruction error.

Sensitivity
Specificity

Nr
For any given state, you can remove most markers (one-at-a-time)
'Averaged' confusion matrix between states

confusion matrix over all marker-hold-outs (diagonal removed)

active promoter
weak promoter
inactive/poised promoter
strong enhancer
strong enhancer
weak/poised enhancer
transcriptional transition
transcriptional elongation
weak transcribed
polycomb repressed
heterochrom; low signal
repetitive/ICNV
repetitive/ICNV
active promoter
Normalization

\[ \sqrt{(y - y(x))^2 / (y - \bar{y})^2} \]

\[ \sqrt{(y - y(x))^2} \]

BLUE = IMR90, RED = H1

\[(y - \bar{y})^2\]

BLUE = IMR90, RED = H1
Future work

Reconstructing held-out markers:
Our prediction models all assumed independence of positions. Using the sequence context could likely improve reconstruction of held-out markers considerably.

Annotating Chromatin States with reduced subset of markers: