Lecture 14
Epigenomics

Goals for today

- Introduction to Epigenomics
  - Overview of epigenomics, Diversity of Chromatin modifications,
  - Antibodies, ChIP-Seq, data generation projects, raw data
- Primary data processing: Read mapping, Peak calling
  - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
  - Peak calling: One mark at a time, Binarization, cutoff selection
- Discovery and characterization of chromatin states
  - A multi-variate HMM for chromatin combinatorics
  - Promoter, transcribed, intergenic, repressed, repetitive states
- Model complexity: selecting the number of states/marks
  - Selecting the number of states, selecting number of marks
  - Capturing dependencies and state-conditional mark independence
- Dynamics of chromatin marks and states across cell lines
  - ENCODE data, whole-gene modeling
  - Motifs associated with chromatin state changes
- Disease epigenomics: interpreting disease-associated variants

One Genome – Many Cell Types

Diverse epigenetic modifications
Diversity of epigenetic 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 methylation primarily at CpG
- Nucleosome positioning
- DNA accessibility
- The constant struggle of gene regulation
- TF/histone/nucleo/GFs/Chrom compete

Ongoing epigenomic mapping projects

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

ChIP-chip and ChIP-Seq technology

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

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

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

Goals for today

- Introduction to Epigenomics
  - Overview of epigenomics, Diversity of Chromatin modifications,
  - Antibodies, ChIP-Seq, data generation projects, raw data
- Primary data processing: Read mapping, Peak calling
  - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
  - Peak calling: One mark at a time, Binarization, cutoff selection
- Discovery and characterization of chromatin states
  - A multi-variate HMM for chromatin combinatorics
  - Promoter, transcribed, intergenic, repressed, repetitive states
- Model complexity: selecting the number of states/marks
  - Selecting the number of states, selecting number of marks
  - Capturing dependencies and state-conditional mark independence
- Dynamics of chromatin marks and states across cell lines
  - ENCODE data, whole-gene modeling
  - Motifs associated with chromatin state changes
- Disease epigenomics: interpreting disease-associated variants

Read mapping
Mapping Reads to the Genome

- Have tens of millions of reads ~30 bases long, for example
  CAGGCTGATAGGACATTCACAG
  want to assign the read to the best matching location in the reference genome if a
  mismatch tolerance is satisfied
  - Mismatches can be due to sequencing errors or SNPs
- Algorithmic efficiency is a critical issue!
- Problem encountered in all high-throughput sequencing applications

Second Generation Mappers have Leveraged the
Burrows Wheeler Transformation

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

The power of the 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
  - Runs of letters compressed into (letter,run-length) 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 to the genome
  - Pre-process once; map all reads in transformed space.
  - Reverse transform once; counts at each genome coord

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: \( O(m) \), L3 addendum
- Challenge: memory requirements
  - Hash table, suffix tree/array require \( O(m*n) \) space
- Today: Burrows-Wheeler transformation \( O(m) \)
  - Ultrafast/memory efficient. New norm since 2009.
  - Introduced in: Bowtie (Langmead GB 2009).

Burrows-Wheeler Transform (BWT)

- Transform: *BANANA@ INTO: BNN*AA@A

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

  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
    - if first insert creates first column
      - sort rows of the table alphabetically
  return (row that ends with the ‘EOF’ character)

Searching for an Exact Match

e.g. Searching for OLIS

  In MANOLISKELLIS

  For simplicity (here):
  - only exact matches
  - Show entire matrix
  - in practice: only pointers

<table>
<thead>
<tr>
<th>OLIS</th>
<th>OLIS</th>
<th>OLIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. SWALLISLELLIS</td>
<td>1. SWALLISLELLIS</td>
<td>1. SWALLISLELLIS</td>
</tr>
<tr>
<td>2. WOLISLELLIS</td>
<td>2. WOLISLELLIS</td>
<td>2. WOLISLELLIS</td>
</tr>
<tr>
<td>3. LISWOLISLEL</td>
<td>3. LISWOLISLEL</td>
<td>3. LISWOLISLEL</td>
</tr>
<tr>
<td>4. LISWOLISLEL</td>
<td>4. LISWOLISLEL</td>
<td>4. LISWOLISLEL</td>
</tr>
<tr>
<td>5. LISWOLISLEL</td>
<td>5. LISWOLISLEL</td>
<td>5. LISWOLISLEL</td>
</tr>
<tr>
<td>6. LISWOLISLEL</td>
<td>6. LISWOLISLEL</td>
<td>6. LISWOLISLEL</td>
</tr>
<tr>
<td>7. LISWOLISLEL</td>
<td>7. LISWOLISLEL</td>
<td>7. LISWOLISLEL</td>
</tr>
<tr>
<td>8. LISWOLISLEL</td>
<td>8. LISWOLISLEL</td>
<td>8. LISWOLISLEL</td>
</tr>
<tr>
<td>9. LISWOLISLEL</td>
<td>9. LISWOLISLEL</td>
<td>9. LISWOLISLEL</td>
</tr>
<tr>
<td>10. LISWOLISLEL</td>
<td>10. LISWOLISLEL</td>
<td>10. LISWOLISLEL</td>
</tr>
<tr>
<td>11. LISWOLISLEL</td>
<td>11. LISWOLISLEL</td>
<td>11. LISWOLISLEL</td>
</tr>
<tr>
<td>12. LISWOLISLEL</td>
<td>12. LISWOLISLEL</td>
<td>12. LISWOLISLEL</td>
</tr>
<tr>
<td>13. LISWOLISLEL</td>
<td>13. LISWOLISLEL</td>
<td>13. LISWOLISLEL</td>
</tr>
<tr>
<td>14. LISWOLISLEL</td>
<td>14. LISWOLISLEL</td>
<td>14. LISWOLISLEL</td>
</tr>
<tr>
<td>15. LISWOLISLEL</td>
<td>15. LISWOLISLEL</td>
<td>15. LISWOLISLEL</td>
</tr>
<tr>
<td>16. LISWOLISLEL</td>
<td>16. LISWOLISLEL</td>
<td>16. LISWOLISLEL</td>
</tr>
</tbody>
</table>

P is the input substring
C1:C2...CJ is how many characters occur before c geographically in the genome
C(x,c) is the number of occurrence of the character c before index x in the bar right column

© 2009. 8

Example by Jean Ginis.
**Goals for today**

- **Introduction to Epigenomics**
  - Overview of epigenomics, Diversity of Chromatin modifications,
  - Antibodies, ChIP-Seq, data generation projects, raw data
- **Primary data processing: Read mapping, Peak calling**
  - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
  - Peak calling: One mark at a time, Binarization, cutoff selection
- **Discovery and characterization of chromatin states**
  - A multi-variate HMM for chromatin combinatorics
  - Promoter, transcribed, intergenic, repressed, repetitive states
- **Model complexity: selecting the number of states/marks**
  - Selecting the number of states, selecting number of marks
  - Capturing dependencies and state-conditional mark independence
- **Dynamics of chromatin marks and states across cell lines**
  - ENCODE data, whole-gene modeling
  - Motifs associated with chromatin state changes
- **Disease epigenomics: interpreting disease-associated variants**

---

**Peak/interval calling (ChIP-seq/chip)**

- After reads are aligned signal tracks can be computed
- One can then be interested in identifying the coordinates of peaks or enriched domains of a modification
- For histone modifications methods based on scan statistics and univariate HMMs have been used
- To some extent an art, ambiguity as to what is desired especially with broader domains

**Data Binarization**

- Leads to biologically interpretable models that can be robustly learned
- Let \( c_i \) 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 \( \frac{1}{\lambda_i} \) if
  \[
P(X>c_i)<10^{-4}\]
  where \( X \) is a Poisson random variable with mean \( \lambda_i \)
Aggregate analyses: ‘supervised learning’

Analysis design choices

- **Anchoring**: ‘Anchored’ vs. non-anchored
  - Aggregate signal on known feature type
  - Cluster based on signal tracks at these features (e.g. cluster all promoters based on chromatin marks)
- **Learning type**: Supervised vs. unsupervised
  - Derive features predictive of different types of elements. E.g. enhancers/promoters
  - Integrative signal analysis and segmenting genome
- **Resolution**: megabase, kb, nucleosome, nucleotide

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

Goals for today

- **Introduction to Epigenomics**
  - Overview of epigenomics, Diversity of Chromatin modifications,
  - Antibodies, ChIP-Seq, data generation projects, raw data
- **Primary data processing: Read mapping, Peak calling**
  - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
  - Peak calling: One mark at a time, Binarization, cutoff selection
- **Discovery and characterization of chromatin states**
  - A multi-variate HMM for chromatin combinatorics
  - Promoter, transcribed, intergenic, repressed, repetitive states
- **Model complexity: selecting the number of states/marks**
  - Selecting the number of states, selecting number of marks
  - Capturing dependencies and state-conditional mark independence
- **Dynamics of chromatin marks and states across cell lines**
  - ENCODE data, whole-gene modeling
  - Motifs associated with chromatin state changes
- **Disease epigenomics: interpreting disease-associated variants**

Example Chromatin State Annotation

Chromatin states: ‘unsupervised learning’
Cartoon Illustration of ChromHMM

Emission Parameters

Transition matrix

State Enrichments

Transition matrix

State Enrichments

(1) Promoter Associated States: Positional and functional properties

(2) Actively Transcribed States: Diverse marks, expression/position biases
(2) Actively Transcribed States: Recovery of highly specific Kap1 combinations

(3) Active intergenic states: Distinct TF/motif enrichments

(4 & 5) Intergenic and Large Scale Repressed States

(6) Functional enrichments enable annotation of 51 distinct states

Apply genome wide to find novel genes, enhancers, insulators

Discovery power for promoters, transcripts

*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 and H3K36, and histone H4K20...

- Enhancer state predictive of expression level
- Different intergenic states, different dist. activity
- Enhancer states indeed distant from promoters

- Distinct enrichments with lamina-associated regions. Constitutive vs. facultative heterochromatin
- Distinct response to HDAC inhibitors: State 44 acetylated suggesting active acetylation turnover
- Distinct sequence signatures: State 46 CAn/TGn/CATGn low-complexity repeat.
- Distinct enrichments for distinct classes of repeat elements, distinct epigenetic marks
- Importance of jointly observing entire vector of marks → repetitive would overwhelm other’s signal

Apply genome wide to find novel genes, enhancers, insulators

Discovery power for promoters, transcripts

- Significantly outperforms single-marks
- Similar power to supervised learning approach
- CAGE experiments give possible upper bound
Shedding light on GWAS disease SNPs
Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction

- rs12619285 in Chr2 intergenic region 40kb 3' of IKZF2 (lymphocyte devel)
- Strongest disease association with numerous inflamations (Gudbjartsson09)
- Strong hit for State33, while surrounding region unenriched (37 and 41-43)

Goals for today
- Introduction to Epigenomics
  - Overview of epigenomics, Diversity of Chromatin modifications, Antibodies, ChIP-Seq, data generation projects, raw data
- Primary data processing: Read mapping, Peak calling
  - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
  - Peak calling: One mark at a time, Binarization, cutoff selection
- Discovery and characterization of chromatin states
  - A multi-variate HMM for chromatin combinatorics
  - Promoter, transcribed, intergenic, repressed, repetitive states
- Model complexity: selecting the number of states/marks
  - Selecting the number of states, selecting number of marks
  - Capturing dependencies and state-conditional mark independence
- Dynamics of chromatin marks and states across cell lines
  - ENCODE data, whole-gene modeling
  - Motifs associated with chromatin state changes
- Disease epigenomics: interpreting disease-associated variants

Model complexity: How many chromatin states?

Model Selection
- Genome so large standard model selection criteria prefer models with more states than most useful for biological interpretation
- Build model with more states than model selected to verify all states of interest in larger model adequately captured in smaller models
- Selected model also depends on parameter search initialization strategy

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

A 79 State Model with the Best BIC score Found
Best of three Random Initialized Models Correlation with 79 State Model – Randomly initialized models can miss biologically significant states without large number of states

Correlations with 79 State Model using Nested Initialization Strategy

Pairwise Expected vs. Observed Mark Co-Occurrence

Model Initialization

- First pass randomly initialize three models for every number of states 2-80
- Second pass form a nested set of initializations of emission parameters by greedily removing states from best BIC model found
- Criteria for nested models is to maximize sum of correlation of emission vectors from randomly initialized models have with most correlated vector from nested model
- Models learned in parallel
Assessing predictive value for subset of marks

State Inferred with subset of marks

Greedy ordering of marks

State confusion matrix with 11 ENCODE marks

Recovery of states with increasing number of marks

Goals for today

- Introduction to Epigenomics
  - Overview of epigenomics, Diversity of Chromatin modifications,
  - Antibodies, ChIP-Seq, data generation projects, raw data
- Primary data processing: Read mapping, Peak calling
  - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
  - Peak calling: One mark at a time, Binarization, cutoff selection
- Discovery and characterization of chromatin states
  - A multi-variate HMM for chromatin combinatorics
  - Promoter, transcribed, intergenic, repressed, repetitive states
- Model complexity: selecting the number of states/marks
  - Selecting the number of states, selecting number of marks
  - Capturing dependencies and state-conditional mark independence
- Dynamics of chromatin marks and states across cell lines
  - ENCODE data, whole-genome modeling
  - Motifs associated with chromatin state changes
- Disease epigenomics: interpreting disease-associated variants

Chromatin state dynamics: enhancer networks

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

Chromatin dynamics: linking enhancer networks

TFs \(\rightarrow\) enhancers \(\rightarrow\) target genes
Introducing multi-cell activity profiles

- Gene expression
- Chromatin States
- Active TF motif enrichment
- TF regulator expression
- Dip-aligned motif biases

Link enhancers to target genes

- ON: Active enhancer
- OFF: Repressed
- Motif enrichment
- Motif depletion

Introducing multi-cell activity profiles

- NHLF
- HMEC
- HSMM

Distal enhancer hard to integrate in regulatory models

- Linked to target genes based on coordinated activity

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

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

Enhancer-gene links supported by eQTL-gene links

- eQTL study
- Individuals
- Indiv. 1: 0.5
- Indiv. 2: 3.5
- Indiv. 3: 1.6
- Indiv. 4: 1.1
- Indiv. 5: 1.1
- Indiv. 6: 0.9
- Indiv. 7: 0.8
- Indiv. 8: 1.0
- Indiv. 9: 1.1

Expression level of gene

- Sequence variant at distal position

Validation rationale:
- Expression Quantitative Trait Loci (eQTLs) provide independent SNP-to-gene links
- Do they agree with activity-based links?

Example: Lymphoblastoid (GM) cells study
- Expression/genotype across 60 individuals
  - 120 eQTLs are eligible for enhancer-gene linking based on our datasets
  - 51 actually linked (43%) using predictions

Chromatin dynamics: linking enhancer networks

- TFs → enhancers → target genes

- Independent validation of links
- Relevance to disease datasets
Introducing multi-cell activity profiles

- Gene expression
- Chromatin States
- Active TF motif enrichment
- TF regulator expression
- Dip-aligned motif biases

Link TFs to target enhancers
Predict activators vs. repressors

Coordinated activity reveals activators/repressors

- Enhancer networks: Regulator $\rightarrow$ enhancer $\rightarrow$ target gene

Causal motifs supported by dips & enhancer assays

- Enhancer activity in HepG2
- TF motif
- Dip motif
- Predicted causal HNF motifs (that also showed dips) in HepG2 enhancers

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

Motifs bound by TF, contribute to enhancers

Goals for today

- Introduction to Epigenomics
  - Overview of epigenomics, Diversity of Chromatin modifications,
  - Antibodies, ChIP-Seq, data generation projects, raw data
- Primary data processing: Read mapping, Peak calling
  - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
  - Peak calling: One mark at a time, Binarization, cutoff selection
- Discovery and characterization of chromatin states
  - A multi-variate HMM for chromatin combinatorics
  - Promoter, transcribed, intergenic, repressed, repetitive states
- Model complexity: selecting the number of states/marks
  - Selecting the number of states, selecting number of marks
  - Capturing dependencies and state-conditional mark independence
- Dynamics of chromatin marks and states across cell lines
  - ENCODE data, whole-gene modeling
  - Motifs associated with chromatin state changes
- Disease epigenomics: interpreting disease-associated variants

Disease epigenomics: interpreting GWAS variants

Revisiting disease-associated variants

- Disease-associated SNPs enriched for enhancers in relevant cell types
- E.g. lupus SNP in GM enhancer disrupts Ets1 predicted activator
Mechanistic predictions for top disease-associated SNPs

Lupus erythematosus in GM lymphoblastoid
Erythrocyte phenotypes in K562 leukemia cells

- Disrupt activator Ets-1 motif
  - Loss of GM-specific activation
  - Loss of enhancer function
  - Loss of HLA-DRB1 expression

- Creation of repressor Gfi1 motif
  - Gain K562-specific repression
  - Loss of enhancer function
  - Loss of CCDC162 expression

HaploReg: systematic ENCODE mining of top variants
(compbio.mit.edu/HaploReg)

- To use: Paste in list of SNPs or select a GWA study
  - Mine publically available ENCODE data for significant hits
  - Hundreds of assays, dozens of cells, conservation, motifs
  - Report significant overlaps and link to info/browser

Goals for today

- Introduction to Epigenomics
  - Overview of epigenomics, Diversity of Chromatin modifications,
  - Antibodies, ChiP-Seq, data generation projects, raw data
- Primary data processing: Read mapping, Peak calling
  - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
  - Peak calling: One mark at a time, Binarization, cutoff selection
- Discovery and characterization of chromatin states
  - A multi-variate HMM for chromatin combinatorics
    - Promoter, transcribed, intergenic, repressed, repetitive states
- Model complexity: selecting the number of states/marks
  - Selecting the number of states, selecting number of marks
  - Capturing dependencies and state-conditional mark independence
- Dynamics of chromatin marks and states across cell lines
  - ENCODE data, whole-gene modeling
  - Motifs associated with chromatin state changes
- Disease epigenomics: interpreting disease-associated variants