Computational Analysis of UHT Sequences
Histone modifications, CAGE, RNA-Seq

Philipp Bucher

Wednesday January 21, 2009
SIB graduate school course
EPFL, Lausanne
ChIP-seq against histone variants: Biological questions

ChIP-seq data for histone variants:
  • Which regions of the genome are enriched in a particular variant
  • Position of individual nucleosomes,
  • Epigenetic genome organization, chromatin motifs

Combined analysis
  • Position of TF binding sites relative to nucleosomes.
  • Position of transcription start sites relative to nucleosomes

Advanced analysis:
  Identification of motifs (chromatin pattern recognition)
ChIP-seq data analysis: Biological questions

Histone modifications:

Hypothetical association with functional domains

H3K4me3  promoters
H2AZ      promoters
H3K4me1   enhancers
H3K36me3  transcribed regions
H3K9me3   gene silencing
Histone modifications: Autocorrelation analysis

Top:
Auto-correlation plot of H3K36me3 in mouse ES cells

Bottom:
Auto-correlation plot of H3K4me3 in mouse ES cells

Observations:
H3K36me3 → long range correlation
H3K4me3 → short range correlation

Count densities in enriched regions:
H3K36me3: 0.005 (1 per nucleosome)
H3K4me3: 0.04 (6 per nucleosome)
Fragment size distribution:

ChIP against H3K4me3:

Top: Nucleosome-treated
Barski et al. (2007): human
CD4+ cell lines

Bottom: sonicated material
Mikkelsen et al. (2007):
mouse ES cells

Conclusion:
Nuclease treatment produces higher resolution data
Looking at histone modifications at a particular genomic region

Based on data: from Barski et al. 2007, Cell 129, 823-837.
ChIP-Seq tags from both strands centered by 70 bp.
WIG file resolution: H3K4me1 50 bp, H3K4me3 10 bp, H2A.Z 25 bp.
Analysis of ChIP-Seq signal around transcription start sites (TSS)

Input to ChIP-Cor analysis:

- Reference feature: ~40’000 annotated TSS from ENSEMBL
- Target features ChIP-Seq (Barski et al. 2007):
  - H3K4me3 (centered)
  - H2AZ (centered)
  - POL II (centered)
  - CTCF (centered)
- Target features sequences motifs (Bucher 1990)
  - GC-box (predicted with weight matrix)
  - CCAAT-box (weight matrix)

Questions:
- Arrangement and histone composition of promoter nucleosomes
- Positional preference of promoter motifs relative to nucleosomes
ChIP-Seq signal around annotated TSS

Histone modifications, CTCF and Pol II signal near human TSS

Scaling: global over-representation
Promoter elements and chromatin signatures around TSS

Scaling: global over-representation, each curved divided by mean before superposition
Epigenetic promoter signatures in human:

Summary of observations:

- Strong nucleosome phasing downstream of the TSS
- Strong enrichment in H3K4me3 and H2AZ (known before)
- H3K4me3 prefers downstream nucleosomes.
- first downstream nucleosome centered at pos. +120 (invariable)
- first upstream nucleosome centered at pos. $< -180$ (variable)
- minimal nucleosome-free region of 150 bp
- nucleosome-free region coincides with location of promoter elements

Interesting difference to yeast promoters:

- first downstream nucleosome centered at +60 (TSS within nucleosome)
- nucleosome-free region about 120 bp
Nucleosome phasing around *in vivo* TFBS

Nucleosome phasing: nucleosomes occur at same positions in different chromosomes.

ChIP-Seq data: Phasing leads to periodic signal distributions of about 180 bp

Are nucleosomes phased in human cells?

   Tentative answer: in certain regions only

What determines nucleosome phasing?

   Tentative answer: TF binding sites and space competition

   Nucleosome-intrinsic binding preferences to specific sequences

Expected ChIP-Seq profiles:
Example: CTCF sites
(Chip-Seq data from Barski et al. 2007)

CTCF weight matrix (Sequence logo):
Matrix derived from 3888 highly enriched peak regions
Maximal score of weigh matrix: 90
Threshold score: 50
Human genome contains 77924 sites with score $\geq$ 50

Average # of counts per predicted site
H2AZ nucleosome distribution around predicted CTCF sites

Reference feature matrix-predicted CTCF sites (score ≥ 50)
Target: centered (±65 bp) H2AZ tags

Related paper presentation:

The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome.
Fu et al. - *PLoS Genetics* - 2008
Partitioning algorithms

Purpose:
To partition the genome into signal-rich and signal-poor regions

Applications:
Variable peaks size
Differential chromatin state analysis

Algorithm type 1:
Maximal scoring segment approach
A segment is optimally delineated if
  It does not include a higher scoring segment
  It is not included in a higher scoring segment

Algorithm type 2:
Mixture model/optimal parsing approach
Example: Chip-part (see following slides)
Overview of ChIP-partition

Purpose:
    Segmentation: splitting of genome in signal-rich and signal-poor regions

Input:
    • Centered (or un-centered) tag counts

Output:
    • List of signal-enriched regions (beginning, end)

Principle:
    • Optimization of a partition scoring function by a fast dynamic programming algorithm
    • Two parameters: count density threshold, transition penalty

Scoring function for global results: sum of scores of
    • Transitions (penalty)
    • Scores of signal-rich region: length × (local count-density − threshold)
    • Score of signal-poor region: length × (threshold − local count-density)
Viewing the results of the partitioning program in the genome browser

Custom tracks:
Mikkelsen07: results of ChIP-partition program (BED file)
Chromatin domains are differentially occupied in different cell lines

Mouse Nanog region, H3K4me3 profiles.
ES, ESHyb: embryonic stem cell lines, MEF: embryonic fibroblasts, NP neural progenitors
Using partitioning for differential histone modifications

A) 
ES
ES.Hyb
MEF
NP

B) 
CD1  CD2  CD3  CD4  CD5  CD6  CD7  CD8

C) 
<table>
<thead>
<tr>
<th></th>
<th>CD1</th>
<th>CD2</th>
<th>CD3</th>
<th>CD4</th>
<th>CD5</th>
<th>CD6</th>
<th>CD7</th>
<th>CD8</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES</td>
<td>9</td>
<td>21</td>
<td>46</td>
<td>15</td>
<td>11</td>
<td>19</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>ES.Hyb</td>
<td>13</td>
<td>38</td>
<td>37</td>
<td>26</td>
<td>10</td>
<td>13</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>MEF</td>
<td>3</td>
<td>0</td>
<td>12</td>
<td>7</td>
<td>6</td>
<td>17</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>NP</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>14</td>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>
CAGE - Cap Analysis of Gene expression

Method based on chemical modification of Cap-structure.

Figure shows old version for SAGE-inspired concatemer-sequencing

Cage Data Analysis

Goals:

- Identification of transcription start sites for promoter analysis
- Quantitative measurement of alternative promoter usage
- Detection of new transcripts

Analysis methods

- Mapping of tags to genome (as with ChIP-Seq)
- Segmentation and Peak identification
- Classification of promoters according to initiation site patterns.

Related paper presentation:

A code for transcription initiation in mammalian genomes.
Frith et al. - Genome Research - 2008
Example of CAGE tag distribution around a promoter site

Larger regions (H3K4me1 in blue)

Detailed view below
RNA-Seq

RNA-Seq: Some technical issues

Mapping reads to the genome:
1. First try to map co-linear reads by Eland or similar software
2. Try to map remaining reads to synthetic splice-junction database

Synthetic splice junction database may be based on
- Combinations of donors and acceptors from transcriptome database
- Combination of predicted donors and acceptors within genome interval

Normalized expression measure:
RPKM: reads per kilobase of exon model per million mapped reads

Related paper presentation:

Further Reading:
Methylome analysis

Goal:
Identification of the degree of CpG methylation at regional scale or bp resolution

MeDIP-Seq:
IP against methylated DNA
UHT sequencing
Computational issues: mapping of reads, segmentation

Bisulfate sequencing:
Bisulfate converts non-methylated C into U (sequenced as T)
Computational issues: Mapping of reads to genome not entirely trivial!

Related paper presentation:
A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Down et al. - Nature Biotechnology - 2008