Computational Analysis of Ultra-high-throughput sequencing data: ChIP-Seq

Philipp Bucher

Wednesday January 21, 2009
SIB graduate school course
EPFL, Lausanne
Data flow in ChIP-Seq data analysis

Level 1: Image files (hundreds of Gbytes)

Level 2: Base call primary output: fluorescent dye intensities

Level 3: Tag sequences with quality scores

Level 4: Unsorted mapped sequence tags (Eland output)

Level 5: Genomic count distribution file (sga, gff, WIG format)

Peak recognition  Segmentation  Higher level analysis
Mapping tags to the genome

Method used by Eland software (Illumia)

Try to match sequences of fixed length (e.g. 32) exactly to genome
  Accept unique matches
  Discard non-unique matches
  Keep non-matching segments for next round
Repeat sequencing matching allowing for one mismatch
Repeat sequencing matching allowing for two mismatches

Problem with repetitive regions
  Under-representation of tags in highly conserved repeat elements
  However: most repeat elements are not highly conserved
  Longer reads produce more unique matches

Known artifacts:
  Some positions attract many tags (up to 50’000) – reasons unknown (to us).
  PCR-amplification of small populations also produces multiple hits per positions
Remedy: Set maximal values for counts per positions
Higher level analysis with ChipSeq Tools

Design principle:

• Simple tools (easy to understand to non-specialists)
• Fast algorithms (speed of computers will not drastically increase)
• Generic methods if possible (not restricted to ChIP-seq data)
• C-programs for elementary procedures (glued together with Perl)

Web interface:

• Access and analysis of selected public data
• Upload and analysis of private (user-owned) data
• Combined analysis of private and public data
• Interoperability with sequence analysis programs (e.g. motif discovery)
• Link to genome browser (preparation of customized WIG and BED files)
ChIP-seq against TF: Biological questions

ChIP-seq data for specific transcription factor:

- Average length of an immunoprecipitated fragment (protected DNA regions)
- The location of in vivo occupied sites
- The strength of in vivo occupied sites
- The \textit{in vivo} binding specificity (consensus sequence, matrix)
- Contextual features of \textit{in vivo} occupied binding sites
  - Adjacent sequence motifs
  - Chromatin context
  - Sequence conservation scores
- Correlation with functional sites (\textit{e.g.} transcription start sites)
ChIP-Seq Technique and Data Structure

Our representation: SGA (simple genome annotation) format:

<table>
<thead>
<tr>
<th>sequence ID</th>
<th>feature</th>
<th>position</th>
<th>strand</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_000001.9</td>
<td>stim</td>
<td>559139</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>NC_000001.9</td>
<td>stim</td>
<td>559333</td>
<td>+</td>
<td>1</td>
</tr>
<tr>
<td>NC_000001.9</td>
<td>stim</td>
<td>559356</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>NC_000001.9</td>
<td>stim</td>
<td>559765</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>NC_000001.9</td>
<td>stim</td>
<td>559766</td>
<td>+</td>
<td>3</td>
</tr>
<tr>
<td>NC_000001.9</td>
<td>stim</td>
<td>559767</td>
<td>+</td>
<td>1</td>
</tr>
<tr>
<td>NC_000001.9</td>
<td>stim</td>
<td>559768</td>
<td>+</td>
<td>1</td>
</tr>
<tr>
<td>NC_000001.9</td>
<td>stim</td>
<td>559777</td>
<td>+</td>
<td>3</td>
</tr>
<tr>
<td>NC_000001.9</td>
<td>stim</td>
<td>559778</td>
<td>+</td>
<td>2</td>
</tr>
</tbody>
</table>

... optional additional fields

Fields of SGA format:

1. sequence ID
2. feature
3. position
4. strand
5. Counts
Example: STAT1 ChIP-Seq

Input data:

Stimulated: 15.1 million tags mapped to genome
Unstimulated: 12.9 million tags mapped to genome

Mapping software Eland (Illumina)

First control experiment: Distribution of 5’ end 3’ tags around 37 experimentally defined STAT1 sites.

Results: see next slides:

Conclusions:

The average length of immunoprecipitated fragments is approx 140 bp (distance between 5’ and 3’ peaks)

Application ChIP-Cor

Input:
- genomic tag count distributions for two features (reference, target)
- features may be + and − strand tags from same experiments
- applicable to other types of features, e.g. TSS positions

Output:
- a count correlation histogram
- computes # of times tag pairs are found at a particular distance (range) from each other (large number of tag pairs !).
- different normalization options:
  - count density of target feature (tags per bp)
  - global → relative target tag frequency (normalized to 1)

Purpose:
- identification of average fragment size
- reveals length distribution of enriched domains
- provides clues for choosing parameters for peak and partitioning algorithms

Applications:
- Exploratory analysis
- Rapid quality control
Correlation plot: Example 1

Input data:

Ref: STAT1 5’ tags
Target: STAT1 3’ tags

Analysis parameters:

Range: $-400, +600$
Window width: 10
Count cut-off: 3
Y-axis: count-density

Observations:

Peak center: ~140
Peak count density: 0.03
Background: < 0.007
Correlation plot: Example 2

Input data:
Ref: CTCF 5' tags  
Target: CTCF 3' tags

Analysis parameters:
Range: $-400,+400$  
Window width: 5  
Count cut-off: 3  
Y-axis: count-density

Observations:
Peak center: $\sim75$  
Peak count density: 0.06  
Background: $<0.002$
ChIP-cor reveals differences between fragment site distributions

CTCF, Barski et al. 2007
STAT1, Robertson et al. 2007
NF1: Pjanic et al. (in prep.)

Minimal length 140: CTF/NF1 co-precipitated with nucleosomes?
Application ChIP-Center

Purpose:
• To map tag counts to expected center position of a protein-DNA complex

Input:
• Oriented tag counts for a ChIP-Seq features

Output:
• centered, un-oriented tag counts
• WIG files for viewing data in a genome browser environment

Motivation:
• 5’ and 3’ tag positions show relative displacement to each other
• best estimates for protein-binding site center position:
  5’ end position + ½ fragment length
  or 3’ end position − ½ fragment length
• centered tag count distribution more useful for
  1) peak recognition and partitioning
  2) data viewing in genome browser
Auto-correlation of centered ChIP-Seq tags

Input data:

Ref: STAT1 centered-70
Target: STAT1 centered-70

Analysis parameters:

Range: \(-400, +600\)
Window width: 10
Count cut-off: 3
Y-axis: count-density

Observations:

Peak center: \(~0\)
Peak count density: 0.06
Peak volume: \(~5\) counts
Background: < 0.014
Application ChIP-peak

Purpose:
• identification of peaks corresponding to *in vivo* protein-DNA complexes

Input:
• centered tag counts

Output:
• list of peak center positions

Method:
• consider only positions which have at least one tag count.
• for each position, determine cumulative tag count in window of width $w$.
• select as peaks those positions, which
  • have cumulative tag count $\geq$ threshold $t$.
  • are local maximum with range $\pm r$.
• Optional, refinement of peak center (center of gravity within window)

Interface to sequence analysis programs:
• download of sequences around peak center positions
Example ChIP-Peak: Locating *in vivo* STAT1-binding sites

Input data:
- Cell material: Interferon \( \gamma \)-stimulated HeLa S3 cells.
- About 15 million tags in total

Analysis parameters:
- Centering: 70bp, window \( w=200\)bp, exclusion range \( r=200 \) bp, threshold \( t=100 \) counts

Result: 4446 peaks,
- sequence extraction range for downstream analysis: \(-1000, 1000\)

**Downstream sequence analysis:**

Distribution of TTCNNNGAA around STAT1 peak
- Sliding window size 50
- Figure produced with OPROF (SSA server)
Peak extraction: how to choose the parameters

How to choose the peak width:
   Based on auto-correlation plots

How to choose the threshold?
   Based on a statistical model (requires some assumptions)
   By comparison of the results with a control set (see below)
   By measuring the enrichment of consensus binding sites

<table>
<thead>
<tr>
<th>Peak threshold</th>
<th>STAT1 stim</th>
<th>STAT1 unstim</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>219300</td>
<td>123300</td>
</tr>
<tr>
<td>10</td>
<td>51809</td>
<td>10185</td>
</tr>
<tr>
<td>20</td>
<td>13112</td>
<td>560</td>
</tr>
<tr>
<td>50</td>
<td>2602</td>
<td>48</td>
</tr>
<tr>
<td>100</td>
<td>853</td>
<td>18</td>
</tr>
</tbody>
</table>
From ChIP-Seq peaks to TF binding site matrices

Some considerations

*Ab initio* motif discovery or consensus sequence-initiated refinement?
Speed (time complexity) of downstream programs
Assumptions about binding site: fixed length or variable lengths

Some software tools:

**MEME:**
Seed search, EM refinement in one program
Time complexity $O(N^2)$

**Signal Search Analysis**
Seed search, matrix refinement in two separate programs (SList, PatOp)
Time complexity $O(N)$
Ideally suited to find motifs over-represented at a particular location

**HMM training (e.g. MAMOT)**
Needs initial model
Flexible, can handle palindromes with variable spacing between half-sites
Some results:

In vivo and in vitro specificity of CTF/NF1
(Data from Milos Pjanic and Nicolas Mermod, University of Lausanne)

In vivo specificity: ChIP-Seq

Cellular source:
Mouse embryonic fibroblasts
Mapping software: Eland (Illumina)
ChIP protocol:
fixation (formaldehyde) within cells
Sonication
Chromatin-IP
DNA extraction
Input data for sequence analysis:
Total: 9.7 million tags
within repeats: 3.4 million tags
unique regions: 6.3 million tags

In vitro specificity:

HTP-SELEX
Weight matrix from over >5000 sites
Identical in vivo and in vitro binding specificity for CTF/NF1

Training set:
- Synthetic DNA, SAGE/SELEX, 5579 sequences, Length 25 bp

Initial:
- EM training:
  - In vitro specificity:

Training set:
- Mouse genome, ChIP-seq peaks, 1265 sequences, Length 200 bp

EM training:
- In vivo specificity:
CTCF sites: Correlation between matrix score and ChIP-Seq counts

(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
Results: Are all CTCF sites occupied *in vivo*?

No!

<table>
<thead>
<tr>
<th>Classare</th>
<th>Score range</th>
<th>Count range</th>
<th># of sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>strong-high</td>
<td>65-90</td>
<td>≥10</td>
<td>4476</td>
</tr>
<tr>
<td>strong-low</td>
<td>65-90</td>
<td>&lt;10</td>
<td>6256</td>
</tr>
<tr>
<td>weak-high</td>
<td>50-64</td>
<td>≥10</td>
<td>7655</td>
</tr>
<tr>
<td>weak-low</td>
<td>50-64</td>
<td>&lt;10</td>
<td>59537</td>
</tr>
<tr>
<td>strong-zero</td>
<td>65-90</td>
<td>0</td>
<td>1935</td>
</tr>
</tbody>
</table>

Why are some high scoring sites not occupied?