Statistical analysis of ChIP-chip experiments: data, questions and partial answers

University of British Columbia
Department of Statistics
Raphael Gottardo

Joint with: Shirley Liu, Wei Li, Evan Johnson, Clifford Meyer, Jason S. Carroll and Myles Brown
Outline

• ChIP-chip, what is it?
• Review of the technology
• Statistical goals
• Challenges
• Model based approach
• Preliminary results
• Conclusion/Future work
Microarrays

- Microarrays widely used for gene expression
- Enable to answers many questions:
  - Differential expression
  - Co-expression
- What about gene regulation?
  - Which gene(s) regulates which?
  - Where does this happen? (Regulatory regions)
ChIP-chip

- Chromatin immunoprecipitation (ChIP) on microarrays (chip)
- Enables researchers to identify regions (in a given genome) that are bound by transcription factors
- Recent technology
  - ChIP on cDNA microarrays (Lee et al. 2002)
  - ChIP on tiling arrays (Cawley et al. 2004)
ChIP-chip

Cross link proteins to DNA

Sonication to shear DNA

Add antibody

Antibody

IP enriched DNA

Reverse cross links

Proteins of interest are enriched (Immunoprecipitation)

Sonication

Probes close to a binding site

PCR + fragmentation

Tiling arrays

Label and hybridize

Statistical analysis of ChIP-chip experiments - Raphael Gottardo
ChIP-chip (Control)

Cross link proteins to DNA  Reverse cross links

Sonication to shear DNA

1 kb

PCR + fragmentation

25-100 bps

Tiling arrays

Measure the background intensity

Label and hybridize
Tiling arrays

Unbiased mapping → Tile all non repetitive sequences of the genome

Affymetrix tiling arrays (Krapanov et al. 2002)
High density arrays (~1 million probes)
2/3 of chromosomes 21-22

Current resolution

Probes ordered by locations

25 bps ↔ 35 bps
Details of IP-enriched hybridization

IP-enriched DNA

Probes ordered by locations

Intensity decreases

Random shearing

25 pbs

35 pbs
Analysis of ChIP-chip experiments

• Can be divided into two major steps
  1. Identify bound regions that are in the order of 1000 bps
  2. Sequence analysis of bound regions to identify the actual binding sites and locations

• Here we are concerned with 1

• Note 2 depends on 1!
Challenges

- Huge amount of data
- Small number of replicates (large p small n)
- Noise (Background)
  - Normalization
  - Background adjustment
- Probe dependence (spatial)

(Source NHGRI)
Data

- ER (estrogen receptor) ChIP-chip data (Carroll et al. 2005) on chromosomes 21/22 using Affymetrix tiling arrays (3 arrays)
- For each sample (control and IP enriched) we have three replicates
- 83 verified bound regions (qPCR)
### Data Format

<table>
<thead>
<tr>
<th>Array</th>
<th>Chr</th>
<th>Probe</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>21</td>
<td>1</td>
<td>-1.15</td>
<td>-1.05</td>
<td>-2.52</td>
<td>-2.11</td>
<td>-0.26</td>
<td>-1.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.02</td>
<td>0.58</td>
<td>0.73</td>
<td>0.29</td>
<td>0.62</td>
<td>1.14</td>
</tr>
<tr>
<td>B</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td>-0.83</td>
<td>-0.62</td>
<td>-0.33</td>
<td>-0.66</td>
<td>-0.38</td>
<td>-0.10</td>
</tr>
</tbody>
</table>
Normalization
Normalization

Probe intensities Chip A, control sample
Normalization

We have the sequence information! Does the sequence composition have an effect?

![Box plot showing log(PM) against GC count.](image-url)
Normalization

• Use the sequence information to normalize
  ‣ Build a model that best predicts the observed intensities from the sequence information alone
  ‣ Compute predictions for each array
  ‣ For each probe subtract the predicted value from the observed value

• MATCH (Model-based Analysis of Tiling-arrays for ChIP-chip)
After MATCH Normalization

No jump!
After MATCH Normalization

More symmetric
# Bound regions detection

<table>
<thead>
<tr>
<th>Chr</th>
<th>Probe</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>-1.15</td>
<td>-1.05</td>
<td>-2.52</td>
<td>-2.11</td>
<td>-0.26</td>
<td>-1.36</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.02</td>
<td>0.58</td>
<td>0.73</td>
<td>0.29</td>
<td>0.62</td>
<td>1.14</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>.</td>
<td>.</td>
<td>+</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>.</td>
<td>.</td>
<td>+</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>.</td>
<td>.</td>
<td>+</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>P</td>
<td></td>
<td>-0.83</td>
<td>-0.62</td>
<td>-0.33</td>
<td>-0.66</td>
<td>-0.38</td>
<td>-0.10</td>
</tr>
</tbody>
</table>

Average in the enriched sample > Average in the control sample
Bound regions detection

Probe mean intensity in enriched - mean intensity in control vs. position in chr21

Bound region!  What else?
Posterior probabilities

Likely to be close to a binding sites

Better, but ... still a bit noisy!

Uncertainty

Unlikely
Statistical model

Our posterior probabilities look at individual probes

\[ \Pr(\gamma_p > 0 | \text{data}) \]

Look at several consecutive probes

For each window of size 11 probes:
Probability that at least six probes have positive enrichment effects \( \gamma_p > 0 \)
Posterior probabilities - revisited

Uncertainty
Conclusion/Future work

• Analysis of ChIP-chip data is very challenging
• Huge amount of data and more to come!
  ‣ How to threshold? (binding regions)
  ‣ Do we need a control sample?
Acknowledgements

• Jenny Bryan
• UBC startup fund
• Westgrid (CFI)