Associating differential ER binding with clinical outcome in breast cancer

RORY STARK
18 JULY 2013
ChIP-seq for functional genomics

Most ChIP-Seq studies to date have focused on **mapping**, not **function** (cf ENCODE)
- Comparisons limited to peak overlaps (co-occupancy)
- Limited quantitative analysis

Most **functional** studies to date have focused on RNA levels
- Well established design/analysis
- Unable to directly distinguish driver/upstream from passenger/downstream changes
- Regulatory schema **inferred** (knockouts, modelling)

Can we use ChIP-Seq to more directly **observe** regulatory events?
Agenda

• Differential ER binding in breast cancer: Overview of results
  - Identification of differentially bound sites
  - Performance of prognostic signature
  - Downstream analysis (differential co-factor motifs)

• Method: Differential binding analysis
  - Occupancy analysis
  - Quantitative analysis
  - Bioconductor package: DiffBind
Differential oestrogen receptor binding is associated with clinical outcome in breast cancer


Affiliations  |  Contributions  |  Corresponding authors

Received 19 May 2011  |  Accepted 23 November 2011  |  Published online 04 January 2012
Functional genomics of breast cancer

• Tumors cluster into subtypes based on gene expression

• 70% of tumors over-express primary prognostic marker ER

• ER+ tumors respond to hormone and/or tamoxifen treatment

• Two secondary prognostic markers: PR and HER2

• Prognostic gene expression signatures readily derivable from expression data
ER binding in breast cancer

- ER is a transcription factor
- ERE (estrogen response element) regulatory complex
  - E2 binds ER
  - ER-E2 complexes dimerize
  - Pioneer factor (e.g. FoxA1) opens chromatin
  - ER-E2 dimers bind to DNA at ERE
  - Other TF factors co-bind at ERE
- Most ER binding is intergenic (enhancers, not promoters)
- Evidence of DNA looping
- Previously, all genomic ER binding data derived from a single cell line (MCF-7)
ER ChIP-seq in clinical samples

- 20 BC tumours
- 18 ER+, 2 ER-
- 15 primary, 3 metastases
- 3 sampled in replicate
- Additional controls: 3 normal breast, 2 normal liver
- Two peak callers MACS/SWEMBL (42 peaksets)
- Good/poor prognosis based on PR/HER2 status
Differentially bound sites separate tumours by prognosis

• **1,791** sites identified as differentially bound between good and poor prognosis
  • **599** enhanced in good prognosis
  • **1,192** enhanced in poor prognosis/metastases
Genes near DB sites form prognostic gene signatures

- Signature composed of genes within 20k bases of DB sites
  - 265 genes in Poor outcome signature
  - 109 genes in Good outcome signature
- Classifier based on up/down regulation in mRNA expression sets
- Validated in 7 publicly available BC expression datasets

Expression data: Loi et al.
Differentially enriched co-factor motifs

<table>
<thead>
<tr>
<th>Poor/Metastatic tumours</th>
<th>Tumour Prognosis</th>
<th>Tamoxifen Resistance</th>
<th>Mitogenic Cocktail</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERE</td>
<td></td>
<td>Pax2</td>
<td>Pax2</td>
</tr>
<tr>
<td>FoxA1</td>
<td></td>
<td>AP-1</td>
<td>NFE2L2</td>
</tr>
<tr>
<td></td>
<td>FoxA1</td>
<td></td>
<td>FoxA1</td>
</tr>
<tr>
<td></td>
<td>ERE</td>
<td>ERE</td>
<td>ERE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Good tumours Responsive</th>
<th>Tumour Prognosis</th>
<th>Tamoxifen Resistance</th>
<th>Mitogenic Cocktail</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERE</td>
<td></td>
<td>GATA</td>
<td>GATA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ERE</td>
<td>ERE</td>
</tr>
</tbody>
</table>
Differential Binding Analysis
Differential binding analysis: Observations

- ChIP-seq is highly variable
  - [Technical]
  - Biological
  - Experimental

- Many samples involved
  - Conditions and treatments (contrasts)
  - Factors, marks, antibodies
  - Replicates required to capture variance

- Peak calling is noisy
  - Profusion of peak callers
  - Highly parametric
  - Callers have low agreement on marginal peaks which form majority
Differential binding analysis: Goals

Be robust to noise
- Noisy experiments
- Noisy peak calling

Determine DB without defining global binding maps for each ChIP

Exploit quantitative **affinity** (read scores) beyond binary **occupancy** (peak calls)

Link differential regulatory events (DB) with differential mRNA levels (DE)
Example: Occupancy (peak) analysis

11 Samples, 145586 sites in matrix:

<table>
<thead>
<tr>
<th>ID</th>
<th>Tissue Factor</th>
<th>Condition</th>
<th>Replicate</th>
<th>Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>JC398</td>
<td>MCF7</td>
<td>ER Responsive</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>JC430</td>
<td>MCF7</td>
<td>ER Responsive</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>JC448</td>
<td>MCF7</td>
<td>ER Responsive</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>JC432</td>
<td>T47D</td>
<td>ER Responsive</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>JC439</td>
<td>T47D</td>
<td>ER Responsive</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>JC431</td>
<td>ZR75</td>
<td>ER Responsive</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>JC438</td>
<td>ZR75</td>
<td>ER Responsive</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>JC403</td>
<td>BT474</td>
<td>ER Resistant</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>JC381</td>
<td>BT474</td>
<td>ER Resistant</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>JC511</td>
<td>TAMR</td>
<td>ER Resistant</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>JC510</td>
<td>TAMR</td>
<td>ER Resistant</td>
<td>2</td>
</tr>
</tbody>
</table>

![Graph showing occupancy analysis](image)

- 104,051 ER binding events
- 6,920 events
Binding affinity matrix

1. **Rows**: decide interval (binding site) “universe”
   - Peak callers -> occupancy/overlaps
     - High-confidence sites (stringent)
     - All potential sites (lenient)
   - Genomic intervals
     - Promoters
     - Windows

2. **Columns**: count and normalize reads for all samples in all intervals
   - Duplicate reads
   - Controls
   - Normalization
Differential binding analysis

1. Determine contrasts
   • Single-factor
   • Multi-factor (GLM/blocking)
     - Matched tumour-normal
     - Common tissue
     - Replicate groups (batch)

2. Run RNA-Seq DE package
   - edgeR, DESeq, etc.
   - Fit negative binomial distribution
   - Exact test
   - Multiple testing correction (B&H FDR)
Differential binding analysis: Occupancy vs. Affinity

All Sites

Responsive

Resistant

60,800

5,456

37,793

Differentially Bound Sites

Responsive

Resistant

5,658

3,279

4,964
R/Bioconductor package -- DiffBind

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>dba</td>
<td>Construct a DBA object</td>
</tr>
<tr>
<td>dba.peaksset</td>
<td>Add a peakset to a DBA object</td>
</tr>
<tr>
<td>dba.overlap</td>
<td>Compute binding site overlaps</td>
</tr>
<tr>
<td>dba.count</td>
<td>Count reads in binding sites</td>
</tr>
<tr>
<td>dba.contrast</td>
<td>Establish contrast(s) for analysis</td>
</tr>
<tr>
<td>dba.analyze</td>
<td>Execute differential binding analysis</td>
</tr>
<tr>
<td>dba.report</td>
<td>Generate report for a contrast analysis</td>
</tr>
<tr>
<td>dba.plotHeatmap</td>
<td>Heatmap plots (correlation/affinity)</td>
</tr>
<tr>
<td>dba.plotPCA</td>
<td>Principal Components Analysis plot</td>
</tr>
<tr>
<td>dba.plotMA</td>
<td>MA/scatter plot</td>
</tr>
<tr>
<td>dba.plotBox</td>
<td>Boxplot</td>
</tr>
<tr>
<td>dba.plotVenn</td>
<td>Venn diagram plot of overlaps</td>
</tr>
</tbody>
</table>

```r
> tamoxifen = dba(sampleSheet="tamoxifen.csv")
> tamoxifen = dba.count(tamoxifen)
> tamoxifen = dba.contrast(tamoxifen, categories=DBA_CONDITION)
> tamoxifen = dba.analyze(tamoxifen)
> tamoxifen.DB = dba.report(tamoxifen)
```
Functional analysis of genome-scale regulatory data

- Focus primarily on differential *expression* limits ability to identify upstream/driver genes

- Direct study of differential *regulation* should result in gene signatures enriched for upstream events

- Categorization of differentially regulated genes helps identify co-regulators

- These analysis techniques can be applied to epigenomic regulatory data
Hands-on workshop Friday @ 1PM

Thanks to:

• Cancer Research UK (CRUK)
• Gordon Brown
• CRI Bioinformatics Core
  - Matthew Eldridge
  - Thomas Carroll
• Jason Carroll and his laboratory
  - Caryn Ross-Innes
  - Vasiliki Therodorou
- **Histone marks**
  - H3K4me3
  - H3K36me3
  - H3K9me2
  - H3K9me3
  - H3K27me3

- **Conditions:**
  - Growing vs. Senescent

- **Treatment:**
  - WT vs. treated

- **Replicates:**
  - 1-3 for each mark/condition/treatment

- **“Peaks”:**
  - Windows around TSSs (-1000, +4000)