Discovering gene regulatory control using ChIP-chip and ChIP-seq

“An introduction to gene regulatory control, concepts and methodologies”

Ian Simpson

ian.simpson@ed.ac.uk
The Central Dogma of Molecular Biology
Transcription

Activators
These proteins bind to genes at sites known as enhancers. Activators help determine which genes will be switched on, and they speed the rate of transcription.

Repressors
These proteins bind to selected sets of genes at sites known as silencers. They interfere with the functioning of activators and thus slow transcription.

Coactivators
These “adapter” molecules integrate signals from activators and perhaps repressors and relay the results to basal factors.

TATA box
Core promoter
Coding region

Basal transcription factors
In response to inclusions from activators, these factors position RNA polymerase at the start of the protein-coding region of a gene and send the enzyme on its way.
Gene structure
Example PWM for the human P53 protein

<table>
<thead>
<tr>
<th>CONSENSUS</th>
<th>R</th>
<th>R</th>
<th>R</th>
<th>C</th>
<th>W</th>
<th>W</th>
<th>G</th>
<th>Y</th>
<th>Y</th>
<th>Y</th>
<th>R</th>
<th>R</th>
<th>C</th>
<th>W</th>
<th>W</th>
<th>G</th>
<th>Y</th>
<th>Y</th>
<th>Y</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>p53 target</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>241</td>
</tr>
<tr>
<td>GADD45A</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>MDM2_1</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>MDM2_2</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>BAX</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>MMP2</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>GDF15_1</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>GDF15_2</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>GTSE1</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>GML</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>SCARA3</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>RRM2B</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>PMAIP1</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>TP53INP1</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>TNFRSF10B</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>P53AIP1</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>TP53i3</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>BBC3</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>TNFRSF6</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>IGFBP3</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>SFN</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

| p53_PWM     |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| A           |  7 |  5 | 11 | -177 | 14 |  7 | -177 | -177 | -177 |  4 |  6 |  6 |  9 | -177 | 12 |  7 |  1 |  5 |  3 | -177 |   |
| C           | -177 |  3 |  2 |  20 | -177 |  1 | -177 |  10 |  19 |  10 |  3 | -177 | -177 |  21 | -177 | -177 | -177 |  6 | 10 |  6 |   |
| G           |  11 | 13 |  7 | -177 |  1 | -177 |  21 |  3 |  2 |  1 |  12 |  15 |  11 | -177 | -177 | -177 | -177 |  20 | -177 | -89 |  9 |   |
| T           |  3 | -177 |  1 |  1 |  6 | 13 | -177 |  8 | -177 |  6 | -89 | -177 |  1 | -177 |  9 | 14 | -177 | 10 |  8 |  6 |   |   |
The classic footprinting method

TF binding site screening
- PWM GibbsSampler
- MOODS fast forward

phylogenetic conservation
- PhastCons, UCSCMultiz
- BioProspector

promoter/enhancer detection
- HMM/cis-module
- cluster-buster
- BioTIFFIN
- RedFly
- ModENCODE
Classic phylogenetic footprinting approach
Limitations of the classical approach to finding TFBSs

• The number and quality of binding site sequences is low

• There is no explicit relation between conservation and function
  i.e. sites are often conserved, but conserved sites do not necessarily function

• Assumptions have to be made about where to look and how to score

• Extremely biased information, low number of experiments to determine sites

• Non-physiological conditions used during assessment

• Measurements made only in specific tissue or cells at specific times
  local solutions to the PWM problem, may be wrong for other conditions
Problems with the available data sources

Main source of site specific data remains pattern or PWM (or HMM)

<table>
<thead>
<tr>
<th>Common name</th>
<th>Binary nomenclature</th>
<th>Number of PWMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>Homo sapiens</td>
<td>476</td>
</tr>
<tr>
<td>mouse</td>
<td>Mus musculus</td>
<td>423</td>
</tr>
<tr>
<td>rat</td>
<td>Rattus norvegicus</td>
<td>253</td>
</tr>
<tr>
<td>chick</td>
<td>Gallus gallus</td>
<td>133</td>
</tr>
<tr>
<td>clawed frog</td>
<td>Xenopus laevis</td>
<td>84</td>
</tr>
<tr>
<td>fruit fly</td>
<td>Drosophila melanogaster</td>
<td>68</td>
</tr>
<tr>
<td>thale cress</td>
<td>Arabidopsis thaliana</td>
<td>45</td>
</tr>
<tr>
<td>yeast</td>
<td>Saccharomyces cerevisiae</td>
<td>39</td>
</tr>
<tr>
<td>monkey</td>
<td>Cercopithecus aethiops</td>
<td>29</td>
</tr>
<tr>
<td>gibbon ape</td>
<td>Hylobates lar</td>
<td>24</td>
</tr>
<tr>
<td>cattle</td>
<td>Bos taurus</td>
<td>23</td>
</tr>
<tr>
<td>domestic pig</td>
<td>Sus scrofa</td>
<td>20</td>
</tr>
<tr>
<td>zebra fish</td>
<td>Brachydanio rerio</td>
<td>19</td>
</tr>
</tbody>
</table>

TransfacPro2009.1
Replacing classical prediction with direct localisation

What do we need

• Assays that cover the whole genome (aren't biased)
• Applicable to all transcription factors (good coverage)
• Can be measured in lots of different conditions (condition specific, biologically relevant)
• Can be mapped onto precise (and small) genome locations (high resolution)
• Cost effective, accurate and reliable
Chromatin immuno-precipitation (ChIP)

1. Cross-link proteins
2. Shear DNA (sonication)
3. Recover fragments for TF using antibody (typically pulled down on beads)
4. Reverse cross-linking
5. Label fragments
6. Hybridise to chip and detect
How do we get from populations of DNA fragments to positions on chromosomes?

Currently there are two main choices

- **ChIP-chip**: Hybridisation onto a genomic tiling array
- **Chip-seq**: Direct sequencing of the bound (now released) fragments

**ChIP-chip**

Here a manufactured slide is used in which fragments spanning the genome have been synthesised and attached to the slide surface in a geometric Arrangement. We label our TF retrieved fragments, hybridise them to the slide and then read fluorescence from the features.

**ChIP-seq**

Taking advantage of high throughput sequencing technology (so called next-gen) we attempt to sequence all the fragments. This is quantitative.

In both cases we have issues with mapping, signal processing (noise) and significance testing.
Detection method 1 - Genome tiling arrays (ChIP-chip)
Features of genome tiling arrays

- Generally resolution can be as low as ~3kb, Tfs bind to on average 6-8bp
- How do we know which gene to map to? (meta-data)
  - microarray, gene proximity, functional annotation, in-vivo expression
  - comparison to true positive
- Redundancy probes map to more than one location
- Coverage, cannot cover the genome. This introduces bias.
  - even in Drosophila commonly only 50% of genome possible
  - 2 human chromosomes at 35bp resolution → 1 million features
- Can estimate site occupancy frequency
- Cross-hybridisation can be big problem with repetitive DNA (~5% human genome)
- Processed just like a gene expression microarray
  - SAM, limma (modelled error, tight control of FDR)
Detection method 2 – direct sequencing (ChIP-seq)

Illumina/Solexa SBS sequencing system

Ligate adaptors onto DNA fragments

Denature and attach to substrate

Anneal and extend bridge
Detection method 2 – direct sequencing (ChIP-seq)

1. Complete extension
2. Denature ready for next round
3. Repeat to build cluster
Detection method 2 – direct sequencing (ChIP-seq)

Add fluorescent nucleotides and primer

Scan chip for first base

Enzymatically release block and repeat addition of fluorescent base
Detection method 2 – direct sequencing (ChIP-seq)

Read next base

Repeat until complete

Assemble, align and map sequences
Features of high-throughput sequence data

- Very high resolution, typically 25-mers with mid-spacing ~35bp
- Huge datasets, many Gb of sequence, assembly non-trivial
- Complete genome coverage, no assumption, no bias
- Generally superior at identifying bound sites beyond expectation (this is related to a more accurate ability to discriminate signal from noise)
- Sequences are counted to determine the frequency of site occupancy (better than chip, here seq num is proportional to bound sites)
- Sequences are mapped and converted into signal peaks (typical sizes of bound peaks can range from 50bp-1kb)
- Strong correlation between statistical significance of peak and presence of binding motif (might seem obvious!)
Example ChIP-Chip and ChIP-seq data spanning the atonal locus
Real world examples of ChIP-chip and ChIP-seq in use

Science

Genome-Wide Mapping of in Vivo Protein-DNA Interactions
David S. Johnson, et al.
Science 316, 1497 (2007);
DOI: 10.1126/science.1141319

A Temporal Map of Transcription Factor Activity:
Mef2 Directly Regulates Target Genes
at All Stages of Muscle Development

Combinational Binding Leads to Diverse Regulatory
Responses: Lmd Is a Tissue-Specific Modulator of Mef2
Activity

Ian Simpson, Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh
Studying Drosophila musculature development using ChIP-chip
ChIP-chip blocks integrated with gene expression data for Mef2 and Lmd
Validation of enhancers and TF binding sites

---

Dmel  ATGCTTACCACCACTATTCCGAACAGCTGTGAAGCGTTGCCAATTGTTCTTGAGGATTAACCAA
Dsec  ATGCTTACCACCACTATTCCGAACAGCTGTGAAGCGTTGCCAATTGTTCTTGAGGATTAACCAA
Dere  -TGCTAGCCAACACTTGTTCCGAACAGCTGGAACAGCGTTGCCAATTGTTCTCGCCGGCTAACCAC
Dyak  ATGTAAACCAACACTTGTTCCGAACAGCTGAAGCGTTGCCAATTGTTCTCGCCGGCTAACCAC

---

Ian Simpson, Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh
Validation of enhancer activity for Mef2/Lmd candidate target genes

<table>
<thead>
<tr>
<th>Enhancer binding</th>
<th>Enhancer activity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>blow</strong></td>
<td>![A']</td>
</tr>
<tr>
<td><img src="#" alt="A" /></td>
<td><img src="#" alt="A''" /></td>
</tr>
<tr>
<td><img src="#" alt="B" /></td>
<td>![B'']</td>
</tr>
<tr>
<td><strong>CG5080</strong></td>
<td>![B']</td>
</tr>
<tr>
<td><img src="#" alt="B" /></td>
<td>![B'']</td>
</tr>
<tr>
<td><img src="#" alt="C" /></td>
<td>![C'']</td>
</tr>
<tr>
<td><strong>gol</strong></td>
<td>![C']</td>
</tr>
<tr>
<td><img src="#" alt="C" /></td>
<td>![C'']</td>
</tr>
<tr>
<td><img src="#" alt="D" /></td>
<td>![D'']</td>
</tr>
<tr>
<td><strong>tik</strong></td>
<td>![D']</td>
</tr>
<tr>
<td><img src="#" alt="D" /></td>
<td>![D'']</td>
</tr>
</tbody>
</table>
Temporal binding profiles of over-represented Mef2 bound blocks
Synthesis of the target gene network and known myogensis pathway
Chip-seq analysis of the neuron restrictive silencing factor (NRSF)
ChIP-seq reveals new binding motif flexibility for NRSF

Canonical NRSF PWM logo

Novel NRSF PWM logo
The gene regulatory network downstream of NRSF constructed from ChIP-seq data
Summary for ChIP based target prediction methods

• ChIP-chip and ChIP-seq allow for the first time physical identification of bound regions on the genomic scale

• ChIP-seq presents higher resolution and is replacing ChIP-chip

• Both methods require large data-processing and analysis

• Novel methods have been developed to call bound regions from these data, they are predominantly based on hidden markov models (HMM) and are naturally normally 2-state models (peak, non-peak)

• The resulting regions can be used with classical methods to refine the nature of the regulatory element (PWM Gibbs/HMM profiling, motif detection, conservation)

• Can also be refined by more precise experiments on the ChIP DNA such as targeted PCR

• Revolutionises the analysis of gene regulatory networks by integration with gene expression data
Discovering gene regulatory control using ChIP-chip and ChIP-seq

“Practical analysis of ChIP derived data”

Ian Simpson

ian.simpson@ed.ac.uk
A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences

Wei Li, Clifford A. Meyer and X. Shirley Liu*

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, MA 02115, USA

Received on January 15, 2006; accepted on March 27, 2006
HMMtiling

(1) Initial probabilities: $J/K$ for ChIP-enriched state, $1 - J/K$ for non-enriched state.

(2) Transition probabilities: $J/K$ for transition to a different state, $1 - J/K$ for staying in the same state.

(3) Emission probability distribution of probe $i$ in single dataset: $N(\mu_i + 2\sigma_i, (1.5\sigma_i)^2)$ for ChIP-enriched state, $N(\mu_i, \sigma_i^2)$ for non-enriched state. The parameters are based on the results on the Affymetrix SNP arrays (Lieberfarb et al., 2003).

(4) A probe $i$, with (PM-MM) value $p_i$, is defined as an outlier if its $Z$-value is $>3$ or $<-2.5$. We reassigned the $Z$-value of each outlier probe as $3$ if $Z > 3$ and $-2.5$ if $Z < -2.5$.

(5) If two adjacent probes are farther apart than 500 bp in the genome (usually due to a long repeat sequence between the two probes), in the forward and backward procedure, the enriched and non-enriched state probabilities of the latter probe are reset to the initial probabilities.
Comparison to known p53 binding sites

Tiling Array MDScan profile

Published PWM in Transfac

The canonical consensus
HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data

Zhaohui S Qin¹,²,³, Jianjun Yu³,⁴, Jincheng Shen¹, Christopher A Maher²,³,⁴, Ming Hu¹, Shanker Kalyana-Sundaram³,⁴, Jindan Yu⁵ and Arul M Chinnaian²,³,⁴,⁶,⁷,⁸
HPeak

Hypothetical DNA fragment

Model architecture

Acceptor DNA fragment

Reads

Range of size selection

Read coverage

Observed data
ChIP-seq, from short sequence reads to enriched intervals analysis pipeline using HPeak

- short seq reads (e.g. NCBI-SRA)
- genome sequence

Bowtie

Reads aligned to genome (SAM file)

SAM tools

BAM file

BED tools

BED file

HPeak

Genome Browser (e.g. UCSC, Ensembl)

BigWiggle file

enriched sequence intervals (peaks)
Finding and using resources

Where to find the data

How to visualise genome scale data