An Overview of the Data & Some of the Key Analyses of the ENCODE & modENCODE Consortia:

Interpreting the Transcriptome in terms of the Regulome

M Gerstein, Yale

See last slide for references & more info.

Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via @markgerstein).
Trying to interpret RNA-seq in terms of Regulation

• Large amount of transcriptome (RNA-seq) data generated on diverse systems & in diverse conditions

• Less but still considerable amount of regulatory network & chromatin structure data available, mostly on canonical human & model organism systems

• One goal is to interpret the RNA-seq in light of frameworks provided by the regulatory data
Comparative ENCODE Functional Genomics Resource
(EncodeProject.org/modENCODE.org)

• Broad sampling of conditions across transcriptomes & regulomes for human, worm & fly
  – embryo & ES cells
  – developmental time course (worm-fly)

• In total: ~3000 datasets (~130B reads)
Also Large Amount of Yeast Functional Genomics Data
An Overview of the Data & Some of the Key Analyses of the ENCODE & modENCODE Consortia:
Interpreting the Transcriptome in terms of the Regulome

• Using Logical Gates to Interpret RNA-seq
  - Preponderance of OR gates in the human network vs yeast
  - Relation to cancer (myc)

• Globally Organizing Regulation into a Hierarchy
  - Construction: local BFS vs global simulated annealing
  - Differences between kinase & TF hierarchy
  - More logical structure at top of hierarchy

• HM Models Relating Gene Expression to Promoter Activity
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla

• Similarly constructed TF Models [if time]
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for ‘prediction'
  - Surprisingly, a few TFs are quite predictive
An Overview of the Data & Some of the Key Analyses of the ENCODE & modENCODE Consortia:

Interpreting the Transcriptome in terms of the Regulome

- **Using Logical Gates to Interpret RNA-seq**
  - Preponderance of OR gates in the human network vs. yeast
  - Relation to cancer (myc)

- **Globally Organizing Regulation into a Hierarchy**
  - Construction: local BFS vs. global simulated annealing
  - Differences between kinase & TF hierarchy
  - More logical structure at top of hierarchy

- **HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla

- **Similarly constructed TF Models [if time]**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive
Loregic: A method to characterize the cooperative logic of regulatory factors


General-purpose tool, R package: github.com/gersteinlab/loregic
A gene can be regulated by multiple gene regulatory factors

Next generation sequencing techniques (e.g., ChIP-seq, CLIP-seq) predict **gene regulatory factors (RFs)** and their target genes
- transcription factors (TFs)
- micro-RNAs

Many genes are regulated by multiple RFs. How RFs coordinate to regulate target gene expression?
- cooperative?
- competitive?
- independent?

Gene regulatory network

<table>
<thead>
<tr>
<th>Regulatory Factor (RF)</th>
<th>Target (T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF 1</td>
<td>Gene 1</td>
</tr>
<tr>
<td>TF 2</td>
<td>Gene 1</td>
</tr>
<tr>
<td>TF 3</td>
<td>Gene 2</td>
</tr>
<tr>
<td>miRNA 1</td>
<td>Gene 1</td>
</tr>
<tr>
<td>miRNA 2</td>
<td>Gene 3</td>
</tr>
<tr>
<td>miRNA 3</td>
<td>Gene 2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

...
Modeling cooperativity between RFs to target gene using logic gates

A regulatory triplet

Input type (RF1, RF2)

<table>
<thead>
<tr>
<th>RF1</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Output

| T | X | X | X | X |

0 – gene off
1 – gene on

after binarizing gene expression data*

Binarized expression

X can be 0 or 1, so there are $2^4 = 16$ possible output combinations, each of which corresponds to a unique 2-input-1-output logic gate

*BoolNet, R package
An example: selection of the best-matched logic gate

Laplace’s rule of succession

\[ s = \frac{\text{# of selected output state for the input type} + 1}{\text{# of input type} + 2} \]

Consistency score:

\[ \frac{6}{7} \times \frac{5}{7} \times \frac{6}{7} \times \frac{5}{7} = 0.37 \]

Application 1 – transcription factor cooperativity in Yeast cell cycle

Yeast Cell Cycle

<table>
<thead>
<tr>
<th>Triplet ID</th>
<th>RF1</th>
<th>RF2</th>
<th>Common Target Gene (T)</th>
<th>Matched logic gate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>YHR084W</td>
<td>YBR083W</td>
<td>YBR082C</td>
<td>AND</td>
</tr>
<tr>
<td>2</td>
<td>YKL112W</td>
<td>YIL131C</td>
<td>YMR198W</td>
<td>OR</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>39011</td>
<td>YOR113W</td>
<td>YBL103C</td>
<td>YDR042C</td>
<td>XOR</td>
</tr>
</tbody>
</table>

Target gene: 2464
TF: 176
Triplet: 39,011
Time point: 59

## Application 2 – transcription factor cooperativity in Acute Myeloid Leukemia (AML)

<table>
<thead>
<tr>
<th>Target gene</th>
<th>1824</th>
<th>ENCODE Data (K562, ChIP-seq) <a href="http://encodenets.gersteinlab.org/">http://encodenets.gersteinlab.org/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>TF</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>Regulatory triplet</td>
<td>50,865</td>
<td>TCGA Data (AML, level 3, RNA-seq) <a href="https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp">https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp</a></td>
</tr>
<tr>
<td>Patient sample</td>
<td>197</td>
<td></td>
</tr>
</tbody>
</table>

*Wang, et al., PLoS Computational Biology, 2015*
Application 2 – transcription factor cooperativity in Acute Myeloid Leukemia (AML)

Human TF-TF-target

<table>
<thead>
<tr>
<th>RF1</th>
<th>RF2</th>
<th>Common Target Gene (T)</th>
<th>Matched logic gate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATF3</td>
<td>BDP1</td>
<td>YPEL1</td>
<td>AND</td>
</tr>
<tr>
<td>MYC</td>
<td>BCL3</td>
<td>BCR</td>
<td>T=RF1</td>
</tr>
<tr>
<td>ATF3</td>
<td>BRF2</td>
<td>AIF1L</td>
<td>AND</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

Cancer-related TF, MYC universally amplifies target expression

2,153 (RF1=MYC, RF2=other TFs, T=all common targets) triplets

- RF1
- OR(RF1, RF2)
- OR(RF1, NOT RF2)

High expression of MYC is sufficient for high target gene expression

c-Myc is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells

Zuqin Nie,1,6 Gangqing Hu,2,6 Gang Wei,2 Kairong Cui,2 Arito Yamane,3 Wolfgang Resch,3 Ruoning Wang,4 Douglas R. Green,4 Lino Tessarollo,5 Rafael Casellas,3 Keji Zhao,2,* and David Levens1,*

An Overview of the Data & Some of the Key Analyses of the ENCODE & modENCODE Consortia:

Interpreting the Transcriptome in terms of the Regulome

- **Using Logical Gates to Interpret RNA-seq**
  - Preponderance of OR gates in the human network v yeast
  - Relation to cancer (myc)

- **Globally Organizing Regulation into a Hierarchy**
  - Construction: local BFS v global simulated annealing
  - Differences between kinase & TF hierarchy
  - More logical structure at top of hierarchy

- **HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla

- **Similarly constructed TF Models [if time]**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for ‘prediction'
  - Surprisingly, a few TFs are quite predictive
Gene regulatory pathways have logic-circuit behaviors

Hierarchy Height Statistic = (normalized TF Out deg. – In deg.)

Global Hierarchical Structure

TF out-degree = 3

In-degree = 1

TF target

non-TF target

Algorithms for Defining Hierarchical Structure

- Breadth-first Search
- Globally minimize upward edges
- Globally maximize hierarchy “score”
Breadth-first Search (Locally Optimal)

I. Example network with all 4 motifs

II. Finding terminal nodes (Red)

III. Finding mid-level nodes (Green)

IV. Finding top-most nodes (Blue)

[Yu et al., PNAS (2006)]
Using Simulated Annealing to Globally Minimize the Number of Upward Pointing Edges

Probing direction is an optimization problem

relocating nodes
Hierarchical Score Maximization Algorithm

**Definition**

\[ HS = \frac{N_d + N_h}{N_u + N_h} \]

**Simulated annealing**

maximize \( HS \)

---

### HSM algorithm

**Discretized hierarchy network**

**Probabilistic hierarchy network**

**Hierarchy Score Maximization Algorithm**

\[ HS = N_d + N_h + N_u \]

[Cheng et al. Genome Biol. ('15)]
Apply HSM to a toy example

| A1 | A2 | B1 | B2 | B3 | C1 | C2 | C3 | D1 | D2 | D3 | D4 | D5 | E1 | E2 | E3 | E4 | E5 |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 0  | 0  | 0  | 0  | 0  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  |
| 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 0  | 0  |
| 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1  |

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.43</td>
<td>0.57</td>
<td>0.49</td>
<td>0.51</td>
<td>0.57</td>
<td>0.49</td>
<td>0.51</td>
<td>0.57</td>
</tr>
</tbody>
</table>

**ILI=2**
- HS=2.1
- CHS=2.4
- PHS=2.1

**ILI=3**
- HS=5.5
- CHS=5.5
- PHS=5.5

**ILI=4**
- HS=8.7
- CHS=5.4
- PHS=8.7

**ILI=5**
- HS=∞
- CHS=∞
- PHS=∞

**ILI=6**
- HS=∞
- CHS=∞
- PHS=16.1

**ILI=7**
- HS=∞
- CHS=∞
- PHS=10.4

**ILI=8**
- HS=∞
- CHS=∞
- PHS=9.4

[Cheng et al. GenomeBiol. (in press, '15)]
Example of Path Through Regulatory Network

Expression of MOT3 is activated by heme and oxygen. Mot3 in turn activates the expression of NOT5 and GCN4, mid-level hubs. GCN4 activates two specific bottom-level TFs, Put3 and Uga3, which trigger the expression of enzymes in proline and nitrogen utilization.
Biological Insights from Hierarchy in Yeast TF Regulatory Network

[ Cheng et al. GenomeBiol. (in press, '15) ]
Hierarchical organization of human transcriptional regulatory network
Hierarchical organization of human transcriptional regulatory network

The features are significantly different across levels.

<table>
<thead>
<tr>
<th>TOP</th>
<th>MIDDLE</th>
<th>BOTTOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{\text{TOP}}$</td>
<td>$f_{\text{MID}}$</td>
<td>$f_{\text{BOT}}$</td>
</tr>
</tbody>
</table>

- **TOP**
  - No. PPI TF partners: 27
  - No. miRNA regulators: 16
  - No. ncRNA targets: 630
  - No. miRNA targets: 35
  - No. distal targets: 92
  - Amount of rewiring: 0.59
  - Expression level ($\times 10^3$): 2.6
  - Binding-expression correlation: 0.62
  - ns-SNP density ($\times 10^3$): 1.0
  - Allelicity: 0.20
  - Betweenness full network ($\times 10^3$): 5.5
  - Betweenness TF network: 49

- **MIDDLE**
  - No. PPI TF partners: 18
  - No. miRNA regulators: 24
  - No. ncRNA targets: 593
  - No. miRNA targets: 33
  - No. distal targets: 116
  - Amount of rewiring: 0.73
  - Expression level ($\times 10^3$): 10.6
  - Binding-expression correlation: 0.63
  - ns-SNP density ($\times 10^3$): 3.1
  - Allelicity: 0.13
  - Betweenness full network ($\times 10^3$): 14.0
  - Betweenness TF network: 115

- **BOTTOM**
  - No. PPI TF partners: 11
  - No. miRNA regulators: 10
  - No. ncRNA targets: 321
  - No. miRNA targets: 17
  - No. distal targets: 54
  - Amount of rewiring: 0.80
  - Expression level ($\times 10^3$): 3.0
  - Binding-expression correlation: 0.56
  - ns-SNP density ($\times 10^3$): 3.8
  - Allelicity: 0.18
  - Betweenness full network ($\times 10^3$): 13.5
  - Betweenness TF network: 107

[Nature 489: 91]
Hierarchical organization of human transcriptional regulatory network

The feature are significantly different across levels

More influential TFs at the top

[ Nature 489: 91 ]
Logical cooperativity across hierarchical layers in gene regulatory network

The regulations of middle and top TFs more likely follow logical operations than the bottom TFs.

Putting the regulatory hierarchy in perspective: Kinase network is more hierarchical than the TF reg. network

<table>
<thead>
<tr>
<th></th>
<th>CHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worm neural</td>
<td>2.3</td>
</tr>
<tr>
<td>Political blogs</td>
<td>3.1</td>
</tr>
<tr>
<td>Yeast TF</td>
<td>3.8</td>
</tr>
<tr>
<td>Human TF</td>
<td>5.6</td>
</tr>
<tr>
<td>P2P file sharing</td>
<td>5.8</td>
</tr>
<tr>
<td>Foodweb</td>
<td>6.4</td>
</tr>
<tr>
<td>Human Kinase</td>
<td>13.3</td>
</tr>
<tr>
<td>Yeast Kinase</td>
<td>13.9</td>
</tr>
</tbody>
</table>

[Cheng et al. Genome Biol. (in press, '15)]
An Overview of the Data & Some of the Key Analyses of the ENCODE & modENCODE Consortia: Interpreting the Transcriptome in terms of the Regulome

- **Using Logical Gates to Interpret RNA-seq**
  - Preponderance of OR gates in the human network vs yeast
  - Relation to cancer (myc)

- **Globally Organizing Regulation into a Hierarchy**
  - Construction: local BFS vs global simulated annealing
  - Differences between kinase & TF hierarchy
  - More logical structure at top of hierarchy

- **HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla

- **Similarly constructed TF Models [if time]**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for ‘prediction'
  - Surprisingly, a few TFs are quite predictive
Focus on Promoters

Key Questions
- How do we define the active regions of promoter?
- For an active promoter, how do we relate it bound TFs, its epigenetic marks & its chromatin state to the level of transcription?
- Are these definitions & relationships conserved between very different species?

Relating Genomic Inputs to Outputs

Cell Type 1

Cell Type 2
\[ Y = aX_1 + bX_2 + c \]
Inputs v Outputs:
Upstream Binding/Modification v Expression

PCC: Pol II, 0.33; H3K4me3, 0.28

[Nature 512:445 ('14); doi: 10.1038/nature13424]
Histone Modification (HM) model

TSS

Gene k

Chromatin features: Histone modifications

Predictors

RNA-Seq data

Prediction target: Gene expression level

HM1, 2, 3, ....

Bin 1

Bin 2

Bin 160

Bin 40-1 (TSS-4kb to TSS)

Bin 41-80 (TSS to TSS+4kb)

Bin 120-81 (TTS-4kb to TTS)

Bin 121-160 (TTS to TTS+4kb)
His. mods around TSS & TTS are clearly related to level of gene expression, in a position-dependent fashion

Early work in '09/'10

*Science* 330:6012
[here]

Also:

- Ouyang, Zhou, Wong ('09) *PNAS*;
- Karlic et al. & Vingron ('10) *PNAS*
Integrate all histone modifications to predict gene expression levels

Classify H/L genes (SVM)

Predict expression values

Magnitude of Prediction from a “bin” around the TSS

R=0.75

R=0.60

* = LOG10 RPKM
Human ENCODE Results
Comparison of Models for Gene Expression, Building a Universal Model

-2Kb  TSS  +2Kb

Histone Modifications (HM)

- H3K4me2
- H3K4me3
- H3K27me3
- H3K36me3
- H3K27ac
- H3K4me1
- H4K20me1

Scaled Correlation with Expression

Universal, Human, Worm, Fly

Human, Worm & Fly

Universal Model is Built Simultaneously on Data from all 3 Organisms & Predicts on all 3 with a Single Set of Parameters

Model

Universal Model

[Nature 512:445 ('14); doi: 10.1038/nature13424]
Performance of Universal, cross-organism Model

- works almost as well as species specific models
- works for both mRNAs and ncRNAs
An Overview of the Data & Some of the Key Analyses of the ENCODE & modENCODE Consortia:
Interpreting the Transcriptome in terms of the Regulome

• Using Logical Gates to Interpret RNA-seq
  – Preponderance of OR gates in the human network v yeast
  – Relation to cancer (myc)

• Globally Organizing Regulation into a Hierarchy
  – Construction: local BFS v global simulated annealing
  – Differences between kinase & TF hierarchy
  – More logical structure at top of hierarchy

• HM Models Relating Gene Expression to Promoter Activity
  – Works for ncRNAs as well as genes
  – Universal cross-species model uses same set of parameters across diverse phyla

• Similarly constructed TF Models [if time]
  – Variable importance of regions around genes for HMs & TFs
  – TF & HM signals are redundant for ‘prediction'
  – Surprisingly, a few TFs are quite predictive
Doing a Model with TFs:
Positive and negative regulators from correlating TF signal at TSS with gene expression

[Cheng et al. ('11) PLOS CB]
**Predictor v2:**  
2-levels, now with TFs

[Cheng et al. NAR ('11)]
Human Results

Pearson’s $r=0.81$; RMSE=2.57
Classification: AUC = 0.89
Rrgression: $r = 0.62$; RMSE = 3.06
Models Illuminates Different Regions of Influence for TFs vs HMs

- Datasets
  - ChIP-Seq for 12 TFs (Chen et al. 2008)
  - ChIP-Seq for 7 HMs (Meissner et al.'08; Mikkelsen et al. '07)
  - RNA-Seq (Cloonan et al. 2008)

A TF+HM model that combine TF and HM features does NOT improve accuracy!

TF model accuracy only needs a small number of TFs for high accuracy (>90%)
An Overview of the Data & Some of the Key Analyses of the ENCODE & modENCODE Consortia:

Interpreting the Transcriptome in terms of the Regulome

- **Using Logical Gates to Interpret RNA-seq**
  - Preponderance of OR gates in the human network vs yeast
  - Relation to cancer (myc)

- **Globally Organizing Regulation into a Hierarchy**
  - Construction: local BFS vs global simulated annealing
  - Differences between kinase & TF hierarchy
  - More logical structure at top of hierarchy

- **HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla

- **Similarly constructed TF Models [if time]**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive
An Overview of the Data & Some of the Key Analyses of the ENCODE & modENCODE Consortia:
Interpreting the Transcriptome in terms of the Regulome

• Using Logical Gates to Interpret RNA-seq
  – Preponderance of OR gates in the human network vs yeast
  – Relation to cancer (myc)

• Globally Organizing Regulation into a Hierarchy
  – Construction: local BFS vs global simulated annealing
  – Differences between kinase & TF hierarchy
  – More logical structure at top of hierarchy

• HM Models Relating Gene Expression to Promoter Activity
  – Works for ncRNAs as well as genes
  – Universal cross-species model uses same set of parameters across diverse phyla

• Similarly constructed TF Models [if time]
  – Variable importance of regions around genes for HMs & TFs
  – TF & HM signals are redundant for ‘prediction'
  – Surprisingly, a few TFs are quite predictive
modENCODE/ENCODE Transcriptome subgroup


Hiring Postdocs. See gersteinlab.org/jobs!
Acknowledgements

(11 Main Projects, ~50 labs, >700 substantial contributors + NHGRI)

Networks/Elements (~60 participants):


Hiring Postdocs. See gersteinlab.org/jobs!
• Hierarchy Construction –
  Chao Cheng, Erik Andrews, Koon-Kiu Yan, Matthew Ung, Daifeng Wang
  [papers.gersteinlab.org/papers/hinet coming soon]

• Loregic –
  Daifeng Wang, Koon-Kiu Yan, Cristina Sisu, Chao Cheng, Joel Rozowsky, William Meyerson
  [papers.gersteinlab.org/papers/loregic coming soon]

TF-v-expr:

worm-HM:

ENCODE:
  Chao Cheng, Roger Alexander, Renqiang Min, Kevin Y. Yip, Jing Leng, Joel Rozowsky, Koon-kiu Yan, Xianjun Dong, Sarah Djebali, Yijun Ruan, Carrie A Davis, Piero Carninci, Timo Lassman, Thomas R. Gingeras, Roderic Guigó Serra, Ewan Birney, Zhiping Weng, Michael Snyder

Acknowledgements

Hiring Postdocs. See gersteinlab.org/jobs!
Default Theme

- Default Outline Level 1
  - Level 2
Hierarchical organization of human transcriptional regulatory network

The features are significantly different across levels.

More connected TFs at the top

[Nature 489: 91]
Hierarchical organization of human transcriptional regulatory network

The features are significantly different across levels

[Nature 489: 91]
Info about content in this slide pack

• PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2015 (and beyond). Please read statement at http://www.gersteinlab.org/misc/permissions.html. Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to appropriate place on gersteinlab.org).

• Paper references in the talk were mostly from Papers.GersteinLab.org.

• PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info. In particular, many of the images have particular EXIF tags, such as kwpotppt, that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt