Comparative functional genomics: A VISION for interpretation and validation of function from epigenomic profiles

PSU Weekly Genomics Seminar
February 18, 2015
Ross C. Hardison, Cheryl A. Keller, Nergiz Dogan, Marta Byrska-Bishop, Belinda Giardine, Robert Harris, Stella Chou, Mitch Weiss, Gerd Blobel, Yu Zhang, ...Olgert Denas, Yong Cheng, Michael Snyder, James Taylor
Collaborations

Stella Chou (CHoP)

Mitch Weiss (CHoP; St Jude’s)

Gerd Blobel (CHoP)

David Bodine (NHGRI)

James Taylor (JHU)

Yu Zhang (PSU)

Qunhua Li (PSU)

Bob Paulson (PSU)

Frank Pugh (PSU)

Tom Loughran (UVa)

Mary Poss (PSU)

Rick Myers, Barbara Wold et al. ENCODE Project Consortium

Mouse ENCODE Project Consortium

Yong Cheng; Snyder

Hardison Lab (PSU): ChIP, RNA, DNase sample preps, library construction, Illumina sequencing, mapping, analysis

2/26/2015
3 things

• 1. Summaries from recent papers on hematopoietic lineage choice
• 2. Comparative functional genomics
• 3. VISION = Validated Integrative Systematic InterpretatiON of ....
  – Epigenomic signatures for gene regulation during differentiation
  – Any multidimensional comprehensive datasets relevant to a process
General model for lineage choice

• Lineage choice, i.e. commitment of a previously multipotential progenitor cell to a lineage, occurs with – or even via – establishment of permissive and repressive chromatin states

• These chromatin states are relatively stable within a lineage – even when expression changes dramatically
  – Notable exceptions are intensively studied loci that led to early observations linking changes in chromatin structure to changes in expression

• Induction and repression within a lineage are largely a result of changes in patterns of TF binding on the stage of the permissive chromatin
Models for roles of chromatin states in regulation

Integral to regulatory mechanism

Establish a permissive state

Commitment

Maturation
CHROMATIN STATES ARE ESTABLISHED BEFORE OR DURING COMMITMENT
Chromatin state maps summarize histone modifications across genes

White dot gives RNA level, log2 FPKM

CH12 cells

Weisheng Wu in Mouse ENCODE Consortium (2014) Nature
Chromatin state maps differ *between* lineages but are largely constant at progressive stages within a lineage.

Little change in chromatin state maps for genes differentially expressed within a lineage.

Changes in chromatin state maps for genes differentially expressed between lineages.

Weisheng Wu in Mouse ENCODE Consortium (2014) Nature
Chromatin states distinguish different cell types

Distance in chromatin state between cell types

Normalize mouse and human separately
General **model** for lineage choice

- Lineage choice, i.e. commitment of a previously multipotential progenitor cell, occurs with – or even *via* – establishment of permissive and repressive chromatin states.
- These chromatin states are relatively stable within a lineage – even when expression changes dramatically.
  - Notable exceptions are intensively studied loci that led to early observations linking changes in chromatin structure to changes in expression.
- Induction and repression within a lineage are largely a result of changes in patterns of TF binding on the stage of the permissive chromatin.
CHANGES IN TF OCCUPANCY DRIVE DIFFERENTIAL REGULATION

Maxim Pimkin, Chris Morrissey, Tejas Mishra, Deepti Jain, Weisheng Wu.
Pimkin et al (2014) Genome Research
Wu et al (2014) Genome Research
Most GATA1 and TAL1 binding sites are distinctive to ERYs vs MEGs

Max Pimkin, Tejas Mishra, Chris Morrissey, et al. (2014) Genome Research
Major shifts in TAL1 occupancy during hematopoiesis

Wu et al. (2014) Genome Research

A. Heatmap showing expression levels of TAL1 in hematopoietic cell lineages. The heatmap indicates the expression levels of TAL1 across different hematopoietic cell types, with colors ranging from blue to red, indicating varying levels of expression. The cell types are represented along the x-axis, and the y-axis shows the different stages of hematopoietic differentiation.

Ter119- erythroid progenitors, primary
G1E cell line: Gata1- model for erythroid progenitors
G1E-ER4+E2 cell line: Gata1 restored model for differentiating erythroblasts
Ter119+ differentiating erythroblasts, primary

HPC7, hematopoietic precursor cell line
commitment
multi-lineage myeloid progenitors
Neutrophils

BFU-e
CFU-e
Pro
Basophilic
Polychromatic
Orthochromatic enucleation
Reticulocyte
Erythrocyte

Erythroblasts
Megakaryocytes

HPC Epro G1E ER4 Ebl Meg

16 (205)
15 (159)
14 (299)
13 (503)
12 (382)
11 (841)
10 (605)
9 (911)
8 (1648)
7 (2184)
6 (1588)
5 (1564)
4 (1164)
3 (3567)
2 (396)
1 (2579)
TF binding around MEG-specific genes

Occupancy by TFs before commitment to MEG: lineage priming.

Wu et al. (2014) Genome Research.
TF binding around ERY-specific gene

Occupancy by TFs *after* commitment to ERY: induction.

Wu et al. (2014) Genome Research.
Distinct cohorts of TFs regulating genes during differentiation

Wu et al. (2014) Genome Research.
N-TERMINAL DOMAIN OF GATA1 IS NEEDED FOR ERYTHROID BUT NOT MYELOID DIFFERENTIATION
N-terminus of GATA1 is required for erythropoiesis

Distribution of lineage-committed cells in EB suspension cultures on differentiation days 12-14

Byrska-Bishop et al. .. Chou (2015) JCI
Expression of erythroid genes is repressed in GATA1s vs GATA1fl

Less binding of GATA1s at erythroid targets

Byrska-Bishop et al. .. Chou (2015) JCI
TF OCCUPANCY IS A GOOD PREDICTOR OF ERYTHROID ENHANCERS: TAL1
Many (56%) TAL1 OSs act as enhancers

Transfections into K562 cells
65% of TAL1 OSs are active enhancers when tested in transgenic mice.

Dogan et al. (2015) Epigenetics & Chromatin: submitted
VISTA Enhancer Browser (Visel et al. 2007)
TF +/- Histone Mods are strong predictors of CRMs

Dogan et al. (2015) Epigenetics & Chromatin: submitted
Comparative functional genomics

Comparative epigenomics

Experimental tests
Power in interpretation of comparisons

Comparative genomics

- purifying selection
- adaptive evolution
- lineage specificity

Motifs for GATA factor binding preserved across mammals

Comparative epigenomics

GATA1 factor occupancy in erythroblasts preserved across mammals
4 categories of functional evolution revealed by comparative epigenomics

In 2nd species: Function conserved

Function active in different tissue

Tissue 2

Tissue 1

Sequence conserved

Not present (Lineage specific)

FunctCons FunctActive SeqCons LineageSpec


Denas, Sandstrom, Cheng, Beal, Herrero, Hardison, Taylor (2015) BMC Genomics; bioRxiv
ARE ALL THOSE TFOS REALLY FUNCTIONAL?
Functional vs opportunistic

• *Functional*: Binding of the TF leads to or participates in regulating levels of expression

• *Opportunistic*: Binding of TF has no effect
  – TFs like to bind DNA, especially to their preferred motifs.
  – TFs *could* bind any DNA sequence with a motif instance that is in accessible chromatin.
  – Is the analogy more to a parking lot or a junk yard?
**Distinguishing functional from opportunistic**

<table>
<thead>
<tr>
<th>Type of TFos</th>
<th>Expect for enhancer assay</th>
<th>Quantitative modeling to predict regulation</th>
<th>Conservation of occupancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional</td>
<td>Significant fraction active</td>
<td>Need many TFos for good model</td>
<td>Significant fraction conserved</td>
</tr>
<tr>
<td>Opportunistic</td>
<td>Active at rate similar to DHS</td>
<td>Many TFos expendable in model</td>
<td>Not conserved</td>
</tr>
</tbody>
</table>
4 categories of functional evolution revealed by comparative epigenomics

In 2nd species: Function conserved  
Function active in different tissue  
Sequence conserved  
Not present (Lineage specific)

Tissue 1  
Tissue 2

FunctCons  
FunctActive  
SeqCons  
LineageSpec


Denas, Sandstrom, Cheng, Beal, Herrero, Hardison, Taylor (2015) BMC Genomics
Lineage-specific candidate enhancers are frequently active

Predict candidate regulatory modules from patterns of histone modifications in mouse ES cells

Select those with no orthologous sequence in human

Test for promoter or enhancer activity by transfection of reporter genes

Feng Yue, Bing Ren, The Mouse ENCODE Consortium (2014) Nature
4 categories of functional evolution revealed by comparative epigenomics

In 2nd species:

- **Function conserved**
  - Tissue 1
  - Tissue 2

- **Function active in different tissue**
  - Tissue 1
  - Tissue 2

- **Sequence conserved**

- **Not present (Lineage specific)**


Denas, Sandstrom, Cheng, Beal, Herrero, Hardison, Taylor (2015) BMC Genomics; bioRxiv
Conserved and divergent occupancy of orthologous DNA


Fraction of TFos with occupancy conserved in 2\textsuperscript{nd} species
Conservation of occupancy by GATA1 is a highly accurate predictor of enhancement

Ascertain TF occupancy in erythroid cells (mouse and human), see enhancement in fetal heart, vasculature, midbrain, ...

Multiple functions, more intense selection

- The *FuncCons* TFos in comparative epigenomics are analogous to ultraconserved sequences in comparative genomics
- Model: **Pleiotropic** functions are subject to stronger constraint
  - Active in multiple tissues
  - Multiple TFs binding
- Leads to preservation of occupancy despite tendency of regulatory regions to “turn over”
FunctCons GATA1 TFos: Activity in multiple tissues at E11.5
For many TFs, conservation of occupancy across species predicts enhancement

<table>
<thead>
<tr>
<th>TF</th>
<th>Nmbr</th>
<th>VISTA segs</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAL1</td>
<td>1,370</td>
<td>59</td>
</tr>
<tr>
<td>EP300</td>
<td>1,612</td>
<td>56</td>
</tr>
<tr>
<td>GATA1</td>
<td>622</td>
<td>41</td>
</tr>
<tr>
<td>ETS1</td>
<td>810</td>
<td>29</td>
</tr>
<tr>
<td>MAX</td>
<td>7,049</td>
<td>44</td>
</tr>
<tr>
<td>MYC</td>
<td>4,321</td>
<td>44</td>
</tr>
<tr>
<td>ELF1</td>
<td>5,379</td>
<td>27</td>
</tr>
<tr>
<td>MXI1</td>
<td>2,862</td>
<td>24</td>
</tr>
<tr>
<td>MAZ</td>
<td>2,631</td>
<td>20</td>
</tr>
<tr>
<td>USF1</td>
<td>2,305</td>
<td>19</td>
</tr>
<tr>
<td>USF2</td>
<td>663</td>
<td>3</td>
</tr>
<tr>
<td>JUND</td>
<td>316</td>
<td>11</td>
</tr>
<tr>
<td>UBTF</td>
<td>1,985</td>
<td>10</td>
</tr>
<tr>
<td>MEF2A</td>
<td>140</td>
<td>9</td>
</tr>
<tr>
<td>CHD2</td>
<td>1,602</td>
<td>6</td>
</tr>
<tr>
<td>MEF2A</td>
<td>140</td>
<td>9</td>
</tr>
<tr>
<td>CHD2</td>
<td>1,602</td>
<td>6</td>
</tr>
<tr>
<td>RCOR1</td>
<td>800</td>
<td>31</td>
</tr>
<tr>
<td>BHLHE40</td>
<td>1,033</td>
<td>19</td>
</tr>
<tr>
<td>SIN3A</td>
<td>3,978</td>
<td>16</td>
</tr>
<tr>
<td>NRF1</td>
<td>2,352</td>
<td>2</td>
</tr>
<tr>
<td>MAFK</td>
<td>183</td>
<td>6</td>
</tr>
<tr>
<td>CHD1</td>
<td>354</td>
<td>1</td>
</tr>
<tr>
<td>E2F4</td>
<td>287</td>
<td>0</td>
</tr>
<tr>
<td>RDBP</td>
<td>225</td>
<td>0</td>
</tr>
<tr>
<td>TBP</td>
<td>5,548</td>
<td>26</td>
</tr>
<tr>
<td>POL2</td>
<td>10,227</td>
<td>35</td>
</tr>
<tr>
<td>CTCF</td>
<td>15,623</td>
<td>57</td>
</tr>
<tr>
<td>RAD21</td>
<td>11,446</td>
<td>44</td>
</tr>
<tr>
<td>SMC3</td>
<td>9,229</td>
<td>39</td>
</tr>
</tbody>
</table>

Vista positives: 15% heart
32% forebrain

N. Dogan
Y. Cheng
M. Snyder
L. Pennacchio
A. Visel
R. Hardison

2/26/2015
Insights into cis-regulatory modules from comparative functional genomics

- For prediction of regulatory modules, use epigenetic features, especially TFos
- Then use evolutionary patterns
  - Comparative genomics: to discern history
  - Comparative epigenomics: refined functional predictions
- Conservation of TF occupancy (FunctCons): Pleiotropic functions, core functions
- Repurposed (exapted) TFos (FunctActive): Preferred DNA for turnover of binding site motifs?
- SeqCons: ...
- Lineage specific TF occupancy: Adaptive functions
Hematopoiesis and datasets

A. Diagram showing the process of hematopoiesis, starting from embryonic stem cells (ES) and ending with various types of blood cells, including megakaryocytes, erythrocytes, platelets, and others.

B. A table detailing the datasets for different cell types, including RNA, Access, DNAme, H3K, and TFs.
ATAC-seq and RNA-seq

Mouse mm9 chr7:110,941,001-111,028,000 (87,000 bp)

ATAC-seq

HSC
CMP
MEP
CFU-E
ERY
CFU-M
MEG

RNA-seq minus strand

Hbb-b2
Hbb-b1
Hbb-bh
Hbb-y

chr1:166,030,001-166,084,000 (54,000 bp)

ATAC-seq

Selp

RNA-seq plus strand

ACH

F5
VISION: Validated Integrative Systematic Interpretation of epigenomic data
Bayesian Network Analysis: Discovers regulatory patterns

A. Prototype Network

B. Regulatory rules recovered from querying network

<table>
<thead>
<tr>
<th></th>
<th>G1E</th>
<th>G1E-ER4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATA2</td>
<td>✓</td>
<td>0.4</td>
</tr>
<tr>
<td>TAL1</td>
<td>x</td>
<td>0.6</td>
</tr>
<tr>
<td>GATA1</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TAL1</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Joint P</td>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td>P(Ind)</td>
<td></td>
<td>0.06</td>
</tr>
<tr>
<td>P(Rep)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(Nonresp)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C. Proposed network incorporating chromatin state (~G1E,~ER4) and TF motif instances (~TF)
Illustration of IDEAS. Regulatory elements are identified by (1) as mixture components. Long intervals of differential regulatory regions (unique regulatory backgrounds) are identified by (2) via iHMMs, and genome positions are classified using information from all samples via iHMM in (3).

Yu Zhang
IDEAS outperforms several other enhancer finders

Precision-Recall comparison of enhancers predicted by IDEAS, ChromHMM, Segway and EnhancerFinder evaluated against four enhancer datasets identified by different types of evidence: validated in transgenic mice (VISTA\textsuperscript{89}), FANTOM5 tested enhancers, FANTOM5 CAGEtags\textsuperscript{90}, and EnhancerFinder predictions\textsuperscript{84}. Better performance gives higher values on each axis. EnhancerFinder is only shown for FANTOM data, as it is trained using VISTA, and it is used as a reference in the 4th panel.

Yu Zhang
Thanks

Standing: Garam Han, Chris Morrissey, Belinda Giardine, Ross, Cheryl Keller, Susan Magargee, Maria Long
Sitting: Marta Byrska-Bishop, Deepti Jain, Tejaswini Mishra, Kuan-Bei Chen

Nergiz Dogan, Penn State University
Yong Cheng, Stanford University
Olgert Denas, Emory University
Feng Yue, Penn State University
James Taylor, Johns Hopkins

Bing Ren, UCSD
Mike Snyder, Stanford University
Len Pennacchio, LBNL
Axel Visel, LBNL
TAL1 OSs also carry many epigenetic signatures of enhancers
Enhancers predicted by conserved GATA1 binding are active in tissues with paralogous GATA factors

<table>
<thead>
<tr>
<th>GATA factor</th>
<th>Erythroid, Megakaryocyte</th>
<th>T-lymphocytes</th>
<th>Heart</th>
<th>Brain</th>
<th>Vasculature</th>
<th>Liver</th>
<th>Pancreas</th>
<th>Lung</th>
<th>Intestine</th>
<th>Ovary</th>
<th>Testis</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATA1</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATA2</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATA3</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>GATA4</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATA5</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATA6</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>FOG1</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FOG2</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>

Hypothesis: The same GATA factor-dependent enhancer is used in erythroid (GATA1), heart (GATA4) and brain (GATA3) for different targets.
Non-erythroid function of GATA1-bound sites could result from binding of paralogs (e.g. GATA4) to same site in other tissues.
Integrate histone modifications as distinct chromatin segments
Regulatory regions in genomic DNA are marked by epigenetic features

- Genomic DNA:
  - Clusters of motifs that could be bound by transcription factors (TFs)
- Epigenetic marks on chromatin
  - DNase hypersensitive sites
  - Transcription factor and polymerase occupancy
  - Histone modifications
  - H3K4me3: promoters
  - H3K4me1, H3K27ac: enhancers
  - Binding by co-activators or co-repressors
  - DNA methylation

Naturally occurring isoforms of GATA1

GATA1 full length transcript producing:
- full length form of GATA1
- GATA1s

wtGATA1: GATA1fl & GATA1s

GATA1 short transcript producing only GATA1s

Sankaran et al. JCI 122(7): 2439-2443, 2012
Mutations in human GATA1

83 aa

<table>
<thead>
<tr>
<th>NAD</th>
<th>Nf</th>
<th>Cf</th>
</tr>
</thead>
</table>

Deletion (GATA1s mutations)

Missense mutations

With Trisomy 21 (Down syndrome):
- Transient myeloproliferative disorder (TMD)
- Acute megakaryoblastic leukemia (AMKL)

With normal karyotype:
- Congenital anemia with trilineage dysplasia
- Diamond-Blackfan anemia

Anemia +/- thalassemia
Thrombocytopenia +/- platelet dysfunction
Porphyria

Sankaran et al. JCI 122(7): 2439-2443, 2012
Weiss et al. JCI 122(7): 2346-2349, 2012
iPSCs: Model to study the role of GATA1s in hematopoiesis

MNC cells from peripheral blood:

- GATA1s + euploid
- GATA1s + trisomy 21
- wtGATA1 + euploid
- wtGATA1 + trisomy 21

OCT4, SOX2, KLF4, MYC

2/26/2015
Conservation and repurposing of TFos

10,079 MAX os in MEL cells

- SeqCons
  
  Orthologous sequence in human is bound by MAX in K562 cells

- LineageSpec
  
  Any TF, any cell type

- Orthologous sequence in human is bound but different TF and/or cell type
  
  FuncActive

- Orthologous sequence in human is bound by MAX in K562 cells
  
  FuncCons

Denas, Sandstrom, Cheng, Beal, Herrero, Hardison, Taylor (2015) BMC Genomics; bioRxiv

See also Vierstra et al. (2014) Science