Comparative network analysis of ENCODE and modENCODE data:

Inter-phyla Comparison of Regulation & Transcription

Mark Gerstein
Yale

Material “tweetable” (via @markgerstein)
Inter-phyla Comparison of Regulation & Transcription

• Previous functional genomics analyses on animals have focused on
  – Integrated (cross data type) analyses of single organisms (eg ChIP-seq & RNA-seq just in humans)
  – Comparisons of a 1 data type within closely related organisms (e.g. RNA-seq within mammals, Brawand et al. '11)

• Here, integrated comparison across phyla
  – Of extensive, matched functional genomics data (on human, worm & fly)
  – Revealing ancient “features” of transcription & regulation
ENCODE/modENCODE data

Transcriptomes

Regulomes

Also, large-scale chromatin data

575 different experiments containing >67B reads,
Inter-phyla Comparisons of Regulation & Transcription

Even though human, fly and worm are highly divergent they share many ancient features of transcription:

• **Transcription:** Expression clustering
  – 16 Conserved co-expression modules
  – Inter- and intra-species hourglass behavior of 12 of these modules
  – Expression clustering, particularly of hourglass modules, able to align the developmental stages of worm & fly

• **Regulation:** Hierarchical regulatory network with many FFLs

• **Both:** Statistical Models "Predicting" Expression
  – Small numbers of TF can reliably "predict"
  – Possible to construct a universal HM model for gene expression which works for mRNAs and ncRNAs
Cross-Species Co-expression Clustering

Use Potts model (generalized Ising model) to simultaneously cluster co-expressed genes within an organism as well as ~2000 orthologs shared between organisms.

ncRNAs can also be assigned to modules base of co-expression.
Hourglass Behavior

Canonical Inter-organism Behavior
• “Hourglass hypothesis”: all organisms go through a particular stage in embryonic development ("phylotypic" stage) where inter-organism expression differences of orthologous genes are smallest.
• We identify modules (12 out of 16) which have this behavior at the phylotypic stage.

Intra-organism Behavior also Present
• We observe that the expression of genes across 12 modules are the most tightly coordinated at the phylotypic stage (fly).
• Strongly correlated correlation at phylotypic stage (worm).

[ENCODE-modencode Transcriptome paper, submitted]
Alignment of Developmental Time-Course

For worm & fly find stage-specific genes

We can align developmental stages using fraction of shared orthologs between worm and fly amongst these

Reuse of genes from LE in worm in fly pupa

[ENCODE-modencode Transcriptome paper, submitted]
Alignment of Developmental Time-Course

Using only orthologs in 12 "hourglass" modules show stronger alignment except for absence of genes at the phylotypic stage

– By definition genes in hourglass modules are not phylotypic stage specific, hence the gap

[ENCODE-modencode Transcriptome paper, submitted]
Algorithms for Hierarchy Inference

A Definition

\[ HS = \frac{N_a + N_h}{N_a + N_h} \]

B Simulated annealing

maximize HS

C HSM algorithm

L1 L2 L3

N1 0 0 1
N2 0.4 0.6 0
N3 0 0.84 0.16
N4 0 0.56 0.44
N5 0 0 1
N6 0 1 0
N7 0.46 0.54 0
N8 1 0 0
N9 1 0 0
N10 1 0 0

Repeat k times

Discretized hierarchy network

Probabilistic hierarchy network
Inter-phyla Comparisons of Regulation & Transcription

Even though human, fly and worm are highly divergent they share many ancient features of transcription:

• **Transcription**: Expression clustering
  – 16 Conserved co-expression modules
  – Inter- and intra-species hourglass behavior of 12 of these modules
  – Expression clustering, particularly of hourglass modules, able to align the developmental stages of worm & fly

• **Regulation**: Hierarchical regulatory network with many FFLs

• **Both**: Statistical Models "Predicting" Expression
  – Small numbers of TF can reliably "predict"
  – Possible to construct a universal HM model for gene expression which works for mRNAs and ncRNAs
Conservation of regulatory networks rewiring

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Worm</th>
<th>Fly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of TFs (interactions)</td>
<td>155 (1331)</td>
<td>79 (496)</td>
<td>25 (50)</td>
</tr>
<tr>
<td>Fraction of feedback edges</td>
<td>30%</td>
<td>22%</td>
<td>4%</td>
</tr>
<tr>
<td>% of nodes in Bottom, Middle, Top</td>
<td>30, 37, 33</td>
<td>32, 55, 13</td>
<td>46, 47, 7</td>
</tr>
</tbody>
</table>

Network motif enrichment

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Worm</th>
<th>Fly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>

[ENCODE-modencode Regulation paper, submitted]
Inter-phyla Comparisons of Regulation & Transcription

Even though human, fly and worm are highly divergent they share many ancient features of transcription:

- **Transcription**: Expression clustering
  - 16 Conserved co-expression modules
  - Inter- and intra-species hourglass behavior of 12 of these modules
  - Expression clustering, particularly of hourglass modules, able to align the developmental stages of worm & fly

- **Regulation**: Hierarchical regulatory network with many FFLs

- **Both**: Statistical Models "Predicting" Expression
  - Small numbers of TF can reliably "predict"
  - Possible to construct a universal HM model for gene expression which works for mRNAs and ncRNAs
TF Model for Gene Expression

Construct non-linear model for predicting gene expression from signals of multiple TFs binding in promoter

- Correlations of individual TFs with expression (activators & repressors)
- Integrating correlations into model
- Good accuracy for mRNAs & ncRNAs
- TF model accuracy only needs a small number of TFs for high accuracy (>90%)

[ENCODE-modencode Transcriptome paper, submitted]
Models for Predicting Gene Expression from Histone Marks

Scaled Correlation with Expression

-2Kb  TSS  +2Kb

Histone Modifications (HM)

H3K4me2
H3K4me3
H3K27me3
H3K36me3
H3K27ac
H3K4me1
H4K20me1

Universal  Human  Worm  Fly

Relative Importance

0%  20%  40%  60%  80%  100%

Human  Worm  Fly

Model

[ENCODE-modencode Transcriptome paper, submitted]
Universal, cross-organism Model

Comparison of TF model and HM models around TSS

We construct single universal HM model (one set of parameters)
  - works almost as well as species specific models
  - works for both mRNAs and ncRNAs
Inter-phyla Comparisons of Regulation & Transcription

Even though human, fly and worm are highly divergent they share many ancient features of transcription:

• **Transcription**: Expression clustering
  – 16 Conserved co-expression modules
  – Inter- and intra-species hourglass behavior of 12 of these modules
  – Expression clustering, particularly of hourglass modules, able to align the developmental stages of worm & fly

• **Regulation**: Hierarchical regulatory network with many FFLs

• **Both**: Statistical Models "Predicting" Expression
  – Small numbers of TF can reliably "predict"
  – Possible to construct a universal HM model for gene expression which works for mRNAs and ncRNAs
Acknowledgements

modENCODE/ENCODE Transcriptome Group


modENCODE/ENCODE Regulation Group

Alan P. Boyle, Carlos L. Araya, Cathleen Brdlik, Philip Cayting, Chao Cheng, Yong Cheng, Kathryn Gardner, LaDeana Hillier, Judith Janette, Lixia Jiang, Dionna Kasper, Trupti Kawli, Pouya Kheradpour, Anshul Kundaje, Jingyi Jessica Li, Lijia Ma, Wei Niu, E. Jay Rehm, Joel Rozowsky, Matthew Slattery, Rebecca Spokony, Robert Terrell, Dionne Vafeados, Daifeng Wang, Peter Weisdepp, Yi-Chieh Wu, Dan Xie, Koon-Kiu Yan, Elise A. Feingold, Peter J. Good Michael J. Pazin, Haiyan Huang, Peter J. Bickel, Steven E. Brenner, Valerie Reinke, Robert H. Waterston, Kevin P. White, Manolis Kellis, Michael Snyder
TAR Characterization

• We identify a small novel transcripts using a machine learning algorithm trained on known ncRNAs using RNA-Seq data and more.

• Identify TARs that are significantly correlated and anti-correlated with genes in the 16 modules.