Mouse-Human ENCODE Revisited

ENCODE User’s Meeting
Washington, DC
July 1, 2015

Thomas R. Gingeras
Cold Spring Harbor Laboratory
Human Transcriptome:
- 15 cell lines: nucleus-cytosol
- 70% of ~50K annotated genes

Mouse vs. Human

Study Design

• 18 human cell lines (ENCODE) vs. 25 mouse tissue samples in 5 developmental stages
• Two bio-replicates per sample
• Only data passing IDR at 90%> reproducibility (5 read min)
• Poly A+ from total RNA extracted from each sample used to make Illumina libraries consisting of PE 100mers (400 million reads/replica)
• “Conservation” is not used in this study in an evolutionary sense (i.e., it does not mean that the similarity of any feature shared by the compared genes found in the two species has been maintained by purifying selection)

Key Points to Remember

1. The difference in sample types and species underscores the significance of any similarities
2. Conserved features highlighted are not dependent upon common sequences
Distribution of RNAs Within Individual Human Foreskin Fibroblasts

Hox D10

JazF1

Arjun Raj (U. Penn)
Completing the Mouse Genome Annotation

<table>
<thead>
<tr>
<th>Species</th>
<th>Annotated transcripts</th>
<th>Novel transcripts</th>
<th>Total transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse</td>
<td>90,100</td>
<td>200,032</td>
<td>290,132</td>
</tr>
<tr>
<td>Human</td>
<td>164,174</td>
<td>151,761</td>
<td>315,935</td>
</tr>
</tbody>
</table>
# Supplementing Mouse Genome Annotation

## (A) Mouse

<table>
<thead>
<tr>
<th>Gene category</th>
<th>Exons</th>
<th>Transcripts</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>detected</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number</td>
<td>% of Total</td>
<td></td>
</tr>
<tr>
<td>All long</td>
<td>345,616</td>
<td>327,381</td>
<td>94.7</td>
</tr>
<tr>
<td>Protein-coding</td>
<td>320,024</td>
<td>309,131</td>
<td>96.6</td>
</tr>
<tr>
<td>LncRNAs</td>
<td>16,107</td>
<td>12,964</td>
<td>80.5</td>
</tr>
<tr>
<td>Other</td>
<td>9,599</td>
<td>5,390</td>
<td>56.2</td>
</tr>
<tr>
<td>Novel</td>
<td>201,388</td>
<td>0.58</td>
<td></td>
</tr>
</tbody>
</table>

## (B) Human

<table>
<thead>
<tr>
<th>Gene category</th>
<th>Exons</th>
<th>Transcripts</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>detected</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number</td>
<td>% of Total</td>
<td></td>
</tr>
<tr>
<td>All long</td>
<td>509,579</td>
<td>406,630</td>
<td>79.8</td>
</tr>
<tr>
<td>Protein-coding</td>
<td>432,261</td>
<td>375,287</td>
<td>86.8</td>
</tr>
<tr>
<td>LncRNAs</td>
<td>49,513</td>
<td>20,839</td>
<td>42.1</td>
</tr>
<tr>
<td>Other</td>
<td>29,635</td>
<td>12,183</td>
<td>41.1</td>
</tr>
<tr>
<td>Novel</td>
<td>75,118</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>
Correlation of Expression across the Mouse and Human Genomes

100 bp bins

Whole Genomes
cc=0.67

Alignable Intergenic Regions
cc=0.37
Comparison of Dynamic Range of Expression Levels of Mouse and Human Orthologous Genes

The dynamic range (DNR) of gene expression in a cell line or tissue sample can be up to 6 orders of magnitudes.

Each dot is the DNR using all expressed orthologs found in each of the mouse and human samples.
Mouse expressed PC genes: 20,494

Human expressed PC genes: 18,341

Orthologs in 6 species: 5,971

Expressed Orthologous Genes Expressed In Human and Mouse: 14,984

Species: human, mouse, macaque, rat, chicken, cow
Correlation of $\log_{10} \text{Mean}, \text{Max and Min}$

RPKM vs. Dynamic Range of Expression
Distribution of # Genes and $\log_{10}$ Max and Min RPKM Values

$\log_{10}$ Max

$\log_{10}$ Min

$\sim 1$ RPKM
2 dimensional plot of log mean of expression vs DNR

Deconvolved plots of gene number vs DNR
Two Gene Populations with Conserved Unconstrained and Constrained Variation in Levels of Expression

The dynamic range (DNR) of a gene expression levels across multiple sample types (cell lines and tissues) in human and mouse has a bimodal distribution, identifying two populations of genes with constrained (DNR=<2) and unconstrained (DNR=>2 levels of expression).
Constrained Genes Provide Substantial Fraction of Cell’s/ Tissue’s Total RNA Output

- Approximately 40% of RNA mass is attributed to the 17% of all annotated genes

- This RNA output is smaller for less differentiated cells (embryonic liver cells)
Number of Constrained Orthologous Protein Coding (PC) Genes in Six Species

Mouse expressed PC genes: 20,494

Orthologs in 6 species: 5,971

Constrained: 6,636

Constrained in 6 species: 2,591

1-to-1 orthologs expressed in human and mouse: 14,984

Human expressed PC genes: 18,341
Conclusions

- 73% and 81% of human and mouse 1:1 orthologue genes are expressed comparing cell lines vs tissues.
- 40% of orthologue genes expressed in mouse and human are expressed in 4 other species (macaque, rat, chicken, cow)
- 44% of expressed mouse and human orthologues have constrained expression (<2 log variation in expression)
- 17% of orthologue genes expressed in mouse and human are constrained in their expression
- 39% of expressed mouse and human orthologue genes constrained in their expression are constrained in 4 other species
Correlation of Expression across the Mouse and Human Genomes

100 bp bins

Whole Genomes
cc=0.67

Alignable Intergenic Regions
cc=0.37
Constrained Genes are Drivers of the Correlation in Levels of Gene Expression seen for All Orthologous Genes
<table>
<thead>
<tr>
<th>HK gene set</th>
<th>identifier</th>
<th>Technique used</th>
<th>Number of genes in Gencode v10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fantom5, Nature, 2014</td>
<td>F5</td>
<td>CDNA 5' end sequencing</td>
<td>6,560</td>
</tr>
<tr>
<td>Eisenberg et al., Trends in Genetics, 2013</td>
<td>E-L</td>
<td>RNA-seq</td>
<td>3,664</td>
</tr>
<tr>
<td>Chang et al., PLoS One, 2011</td>
<td>Chang</td>
<td>microarray</td>
<td>1,989</td>
</tr>
<tr>
<td>She et al., BMC Genomics, 2009</td>
<td>She</td>
<td>microarray</td>
<td>1,382</td>
</tr>
<tr>
<td>Intersection</td>
<td></td>
<td></td>
<td>429</td>
</tr>
</tbody>
</table>
Proposal: Principled Definition of Housekeeping Genes

Genes that have the variation in expression levels constrained irrespective of the tissue or species in which they are active.
**Possible Controls of the Conserved Constrained Gene Expression**

- Constraint in gene expression is not reflected by sequence conservation
- Constrained set of genes have patterns of histone modification different from unconstrained genes
- Using human and mouse ENCODE epigenetic data for all cell types studied, stronger histone modification signals (H3K4me3, H3K27ac and H3K36me3) for constrained vs. unconstrained genes (controlling for levels of gene expression sample by sample)
- Suggesting constrained vs. unconstrained gene are under different epigenetic regulatory programs
Other Questions

• Mechanism(s) responsible for establishing, maintaining and inheriting the restricted variation in expression

• What genes are constrained at 1 RPKM in what cell/tissue types
• Are there unconstrained genes that determine cell type and to what levels of expression are they in different cell types
• Do these properties extend to Inc-RNA genes
• What about non-orthologous genes?
Acknowledgements

Cold Spring Harbor
Functional Genomics Group

C. Davis
A. Dobin
J. Drenkow
A. Scavelli
L. H. See
C. Zaleski

CRG, Barcelona
Computational Genomics Group

R. Guigo
A. Breschi
S. Djebali
J. Lagarde
D. Pervouchine