Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project


Science (in press)
<table>
<thead>
<tr>
<th>Integration</th>
<th>Highlights from Worm AWG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td><strong>Timecourse</strong> Data + Inferred Elements</td>
</tr>
<tr>
<td>RNA</td>
<td><strong>7K ncRNAs</strong> from a statistical model; Small number of dramatic splicing changes; <strong>Correlated changes in expression</strong> &amp; binding</td>
</tr>
<tr>
<td>TF Elements</td>
<td>Relating levels of <strong>TF heirarchy</strong> &amp; <strong>HOT regions</strong> to gene properties; Integrated <strong>miRNA-TF network motifs</strong></td>
</tr>
<tr>
<td>Chromatin</td>
<td><strong>Large-scale repression on arms</strong>; Differences in X v autosomes; <strong>Chromatin models predict gene expression</strong> &amp; <strong>TFBS</strong> (and even miRNA expr.)</td>
</tr>
<tr>
<td>Comparative</td>
<td><strong>Most</strong> of constrained sequence <strong>accounted for</strong> by anotation</td>
</tr>
<tr>
<td>vs. Human ENCODE</td>
<td>Initial stats on overall binding &amp; transcription <strong>similar</strong></td>
</tr>
</tbody>
</table>
## Expression, TF-binding, & Chromatin Data Collected over a Developmental Timecourse

<table>
<thead>
<tr>
<th>Experiment Type</th>
<th>Isolated Developmental Stages</th>
<th>Mixed Stages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Embryos (6 substages)</td>
<td>L1 (3 subconditions)</td>
</tr>
<tr>
<td></td>
<td>2 / 139M</td>
<td>1 / 29M</td>
</tr>
<tr>
<td>RNA-seq reads</td>
<td>2 / 111M</td>
<td>1 / 33M</td>
</tr>
<tr>
<td>(experiments(mapped reads)</td>
<td>8 / 51M</td>
<td>1 / 10M</td>
</tr>
<tr>
<td></td>
<td>1 / 0.5M</td>
<td>1 / 0.4M</td>
</tr>
<tr>
<td>ChIP-seq, TFs</td>
<td>6 substages</td>
<td>L1 (3 subconditions)</td>
</tr>
<tr>
<td></td>
<td>2 / 139M</td>
<td>1 / 29M</td>
</tr>
<tr>
<td></td>
<td>2 / 111M</td>
<td>1 / 33M</td>
</tr>
<tr>
<td></td>
<td>8 / 51M</td>
<td>1 / 10M</td>
</tr>
<tr>
<td></td>
<td>1 / 0.5M</td>
<td>1 / 0.4M</td>
</tr>
<tr>
<td>Nucleosome</td>
<td>6 substages</td>
<td>L1 (3 subconditions)</td>
</tr>
<tr>
<td></td>
<td>2 / 139M</td>
<td>1 / 29M</td>
</tr>
<tr>
<td></td>
<td>2 / 111M</td>
<td>1 / 33M</td>
</tr>
<tr>
<td></td>
<td>8 / 51M</td>
<td>1 / 10M</td>
</tr>
<tr>
<td></td>
<td>1 / 0.5M</td>
<td>1 / 0.4M</td>
</tr>
</tbody>
</table>

### Inferred Genomic Elements

<table>
<thead>
<tr>
<th>Inferred Genomic Elements</th>
<th>Representative Developmental Stages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding transcripts</td>
<td>36,067</td>
</tr>
<tr>
<td></td>
<td>34,598</td>
</tr>
<tr>
<td></td>
<td>34,373</td>
</tr>
<tr>
<td></td>
<td>34,373</td>
</tr>
<tr>
<td></td>
<td>34,494</td>
</tr>
<tr>
<td></td>
<td>34,598</td>
</tr>
<tr>
<td></td>
<td>34,373</td>
</tr>
<tr>
<td>TSSes</td>
<td>12,654</td>
</tr>
<tr>
<td></td>
<td>14,211</td>
</tr>
<tr>
<td></td>
<td>14,211</td>
</tr>
<tr>
<td></td>
<td>14,211</td>
</tr>
<tr>
<td></td>
<td>10,240</td>
</tr>
<tr>
<td></td>
<td>10,240</td>
</tr>
<tr>
<td></td>
<td>10,240</td>
</tr>
<tr>
<td>TARs</td>
<td>39,328</td>
</tr>
<tr>
<td></td>
<td>42,421</td>
</tr>
<tr>
<td></td>
<td>41,791</td>
</tr>
<tr>
<td></td>
<td>41,791</td>
</tr>
<tr>
<td></td>
<td>41,791</td>
</tr>
<tr>
<td></td>
<td>41,791</td>
</tr>
<tr>
<td>mRNAs</td>
<td>34,295</td>
</tr>
<tr>
<td></td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>72</td>
</tr>
</tbody>
</table>

### Inferred Genomic Elements

<table>
<thead>
<tr>
<th>Inferred Genomic Elements</th>
<th>Representative Developmental Stages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding transcripts</td>
<td>36,067</td>
</tr>
<tr>
<td></td>
<td>34,598</td>
</tr>
<tr>
<td></td>
<td>34,373</td>
</tr>
<tr>
<td></td>
<td>34,373</td>
</tr>
<tr>
<td></td>
<td>34,494</td>
</tr>
<tr>
<td></td>
<td>34,598</td>
</tr>
<tr>
<td></td>
<td>34,373</td>
</tr>
<tr>
<td>TSSes</td>
<td>12,654</td>
</tr>
<tr>
<td></td>
<td>14,211</td>
</tr>
<tr>
<td></td>
<td>14,211</td>
</tr>
<tr>
<td></td>
<td>14,211</td>
</tr>
<tr>
<td></td>
<td>10,240</td>
</tr>
<tr>
<td></td>
<td>10,240</td>
</tr>
<tr>
<td></td>
<td>10,240</td>
</tr>
<tr>
<td>TARs</td>
<td>39,328</td>
</tr>
<tr>
<td></td>
<td>42,421</td>
</tr>
<tr>
<td></td>
<td>41,791</td>
</tr>
<tr>
<td></td>
<td>41,791</td>
</tr>
<tr>
<td></td>
<td>41,791</td>
</tr>
<tr>
<td></td>
<td>41,791</td>
</tr>
<tr>
<td>mRNAs</td>
<td>34,295</td>
</tr>
<tr>
<td></td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>72</td>
</tr>
</tbody>
</table>

### RNA-seq reads

- PolyA selected: 2 / 139M, 2 / 111M, 1 / 33M, 1 / 9M, 1 / 0.8M, 1 / 0.5M, 1 / 0.4M
- Small RNA selected: 8 / 51M, 1 / 10M, 1 / 9M, 1 / 0.8M, 1 / 0.5M, 1 / 0.4M

### ChIP-chip, Pol II

- 3 experiments: 6 substages
- 147M reads
- 40.1M reads
- 54M reads
- 29M reads

### ChIP-seq, TFs

- 6 experiment types
- 17,147 factors
- 26,944 factors
- 8,004 factors
- 13,960 factors
- 6,944 factors

### ChIP-chip, Histone marks

- 22 experiments: 39M reads
- 15M reads
- 25M reads

### ChIP-seq, Chromatin factors

- 31 experiments: 42M reads
- 88M reads
- 88M reads

### Nucleosome

- 1 experiment: 42M reads
- 88M reads

### Inferred Genomic Elements

- Coding transcripts: 36,067
- TSSes: 12,654
- TARs: 39,328
- mRNAs: 34,295
- Other elements: 17,117

### RNA-seq reads

- PolyA selected: 2 / 139M, 2 / 111M, 1 / 33M, 1 / 9M, 1 / 0.8M, 1 / 0.5M, 1 / 0.4M
- Small RNA selected: 8 / 51M, 1 / 10M, 1 / 9M, 1 / 0.8M, 1 / 0.5M, 1 / 0.4M

### ChIP-chip, Pol II

- 3 experiments: 6 substages
- 147M reads
- 40.1M reads
- 54M reads
- 29M reads

### ChIP-seq, TFs

- 6 experiment types
- 17,147 factors
- 26,944 factors
- 8,004 factors
- 13,960 factors
- 6,944 factors

### ChIP-chip, Histone marks

- 22 experiments: 39M reads
- 15M reads
- 25M reads

### ChIP-seq, Chromatin factors

- 31 experiments: 42M reads
- 88M reads
- 88M reads

### Nucleosome

- 1 experiment: 42M reads
- 88M reads
~7K Novel ncRNA candidates

(Identified by statistical model with >90% PPV, not by individual features)
One type of ncRNA: Transcribed Pseudogenes

Total number of pseudogenes: 1198

Number of pseudogenes with evidence of transcription: 323

Number of pseudogenes that are independently transcribed from the parent gene (ruling out mapping artifacts): 194
Splicing Changes over the Timecourse
(~280 changes/pair-of-stages)

$$\text{Diff}_{\text{gene}_i}(\Theta(\text{Stage}_1), \Theta(\text{Stage}_2)) = \frac{\sum_{k=1}^{K} (\theta_k^{\text{Stage}_1} - \theta_k^{\text{Stage}_2})^2}{K}$$
Dynamics of Correlated Expression Changes over Timecourse
Dynamics of Correlated Expression & Pol2 Binding Changes over Timecourse
Relating a TF Hierarchy with Gene Properties

Max signal

<table>
<thead>
<tr>
<th>Other TFs</th>
<th>GEI-11</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Target fraction

Ratio of noncoding to coding

0 5 10 15 20 25

TF

Essential TF

HOX TF

Gene

miRNA ->target

TF ->target

positively correlated

negatively correlated

(25276)

(3999)

(7426)

(6469)
Similar Patterns in More Elaborate Hierarchies in Yeast

[Yu et al., PNAS ('06); Bhardwaj et al., PNAS ('10)]

Fraction of Essential Genes

P < $10^{-10}$

P < 0.02

P = 0.42
Integrated miRNA-TF Network (with over-represented motifs)
HOT regions of clustered binding

Percent essential** genes

Specific targets

HOT regions

All genes

n.s. 5% 10% 15% 20% 25%

$\rho < 10^{-40}$
Large-scale Chromatin Structure

A

Chromosome III

Group 1

Group 2

Group 3

B

Chromosome X

Gene Density
Repeat Density
EE HCP-3
EE H3K27me3
L3 H3K27me3
L3 LEM-2
l3 H3K9me3
EE H3K9me3
EE H3K9me1
L3 H3K9me1
L3 H3K9me2
EE H3K9me2
EE H4K20me1
L3 H4K20me1
L3 SDC-3
MxE MIX-1
MxE DPY-27
L3 RNAseq
EE RNAseq
EE H3K9ac
EE H3K4me2
EE H3K4me3
EE H3K27ac
L3 H3K4me2
L3 H3K4me3
L3 H3K27ac
EE H4tetra-ac
EE RNA Pol II
EE H3K79me1
EE H3K79me2
EE H3K79me3
EE MES-4
EE H3K36me2
EE H3K36me3
L3 H3K36me2
L3 H3K36me3

Color Key

Value

-1.5 0 1.5
Large-scale Chromatin Structure

A Chromosome III

B Chromosome X

C RefSeq Genes

D H3K9 methylation regions

Repeat regions

Transcribed genes with RNA Pol II and active histone marks
Histone Modification Signals around the TSS and TTS: Relationship to Gene Expression
Chromatin Model Accurately Predicts Expression Levels of coding-genes and miRNAs.
Chromatin Model also Predicts TF Binding Sites & HOT Regions
Conservation of Functional Elements: Most of the Constrained Bases are Annotated
Comparison between Human ENCODE and modENCODE C. elegans Results

**Aggregate Pol II ChIP-Seq over TSSs**

---

**Comparison of TF Binding (modENCODE vs Pilot ENCODE)**

<table>
<thead>
<tr>
<th>Pilot ENCODE Human TF Binding</th>
<th>Total Binding (bp)</th>
<th>Percentage of Pilot Regions (29.96Mb) Bound</th>
<th>Intergenic Binding (bp)</th>
<th>Percentage of Intergenic Pilot Regions (13.15Mb) Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>268,520</td>
<td>0.90%</td>
<td>236,458</td>
<td>1.80%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>modENCODE C. elegans TF Binding</th>
<th>Total Binding (bp)</th>
<th>Percentage of Genome Bound</th>
<th>Intergenic Binding (bp)</th>
<th>Percentage of Intergenic Regions (40.89Mb) Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>499,052</td>
<td>0.50%</td>
<td>452,667</td>
<td>1.11%</td>
</tr>
</tbody>
</table>

---

**Comparison of Transcription (modENCODE vs Pilot ENCODE)**

<table>
<thead>
<tr>
<th>Pilot ENCODE Human Transcription</th>
<th>Pilot Regions Transcribed (bp)</th>
<th>Genic Regions Transcribed (bp)</th>
<th>Percentage Genic Transcription</th>
<th>Intergenic Regions Transcribed (bp)</th>
<th>Percentage Intergenic Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placenta PolyA</td>
<td>484,629</td>
<td>407,959</td>
<td>84.2%</td>
<td>76,670</td>
<td>15.8%</td>
</tr>
<tr>
<td>HeLa PolyA</td>
<td>905,973</td>
<td>507,108</td>
<td>56.0%</td>
<td>398,865</td>
<td>44.0%</td>
</tr>
<tr>
<td>momENCODE C. elegans Transcription</td>
<td>Genome transcribed (bp)</td>
<td>Genic Regions Transcribed (bp)</td>
<td>Percentage Genic Transcription</td>
<td>Intergenic Regions Transcribed (bp)</td>
<td>Percentage Intergenic Transcription</td>
</tr>
<tr>
<td>L2 PolyA</td>
<td>20,421,924</td>
<td>17,244,358</td>
<td>84.2%</td>
<td>3,177,566</td>
<td>15.6%</td>
</tr>
</tbody>
</table>
## Highlights from Worm AWG

<table>
<thead>
<tr>
<th>Integration</th>
<th><strong>Data</strong></th>
<th><strong>Timecourse</strong> Data + Inferred Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RNA</strong></td>
<td>7K ncRNAs from a statistical model; Small number of dramatic splicing changes; <strong>Correlated changes in expression</strong> &amp; binding</td>
<td></td>
</tr>
<tr>
<td><strong>TF Elements</strong></td>
<td>Relating levels of <strong>TF heirarchy</strong> &amp; <strong>HOT regions</strong> to gene properties; Integrated <strong>miRNA-TF network motifs</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Chromatin</strong></td>
<td><strong>Large-scale repression on arms</strong>; Differences in X v autosomes; <strong>Chromatin models predict gene expression</strong> &amp; <strong>TFBS</strong> (and even miRNA expr.)</td>
<td></td>
</tr>
<tr>
<td><strong>Comparative</strong></td>
<td><strong>Most</strong> of constrained sequence <strong>accounted for</strong> by annotation</td>
<td></td>
</tr>
<tr>
<td>vs. Human ENCODE</td>
<td>Initial stats on overall binding &amp; transcription <strong>similar</strong></td>
<td></td>
</tr>
</tbody>
</table>