ENCODE: Understanding the Genome

Michael Snyder

May 8, 2013

Conflicts: Personalis, Genapsys, Illumina
Slides From Ewan Birney, Marc Schaub, Alan Boyle
Encyclopedia of DNA Elements (ENCODE)

- NHGRI-funded consortium
- Goal: delineate all “functional” elements in the human genome
- Wide array of experimental assays
- Three Phases: 1) Pilot 2) Scale Up 1.0 3) Scale up 2.0


Project website: [http://encodeproject.org](http://encodeproject.org)
The ENCODE Consortium Phase 2

Brad Bernstein (Eric Lander, Manolis Kellis, Tony Kouzarides)
Ewan Birney (Jim Kent, Mark Gerstein, Bill Noble, Peter Bickel, Ross Hardison, Zhiping Weng)
Greg Crawford (Ewan Birney, Jason Lieb, Terry Furey, Vishy Iyer)
Jim Kent (David Haussler, Kate Rosenbloom)
John Stamatoyannopoulos (Evan Eichler, George Stamatoyannopoulos, Job Dekker, Maynard Olson, Michael Dorschner, Patrick Navas, Phil Green)
Mike Snyder (Kevin Struhl, Mark Gerstein, Peggy Farnham, Sherman Weissman)
Rick Myers (Barbara Wold)
Scott Tenenbaum (Luiz Penalva)
Tim Hubbard (Alexandre Reymond, Alfonso Valencia, David Haussler, Ewan Birney, Jim Kent, Manolis Kellis, Mark Gerstein, Michael Brent, Roderic Guigo)
Tom Gingeras (Alexandre Reymond, David Spector, Greg Hannon, Michael Brent, Roderic Guigo, Stylianos Antonarakis, Yijun Ruan, Yoshihide Hayashizaki)
Zhiping Weng (Nathan Trinklein, Rick Myers)

Additional ENCODE Participants: Elliott Marguiles, Eric Green, Job Dekker, Laura Elnitski, Len Pennachio, Jochen Wittbrodt

.. and many senior scientists, postdocs, students, technicians, computer scientists, statisticians and administrators in these groups

NHGRI: Elise Feingold, Mike Pazin, Peter Good
The ENCODE Consortium Phase 3

Brad Bernstein (Eric Lander, Manolis Kellis, Tony Kouzarides)

Ewan Birney (Jim Kent, Mark Gerstein, Bill Noble, Peter Bickel, Ross Hardison, Zhiping Weng)

Greg Crawford (Ewan Birney, Jason Lieb, Terry Furey, Vishy Iyer)

Jim Kent (David Haussler, Kate Rosenbloom) Mike Cherry

John Stamatoyannopoulos (Evan Eichler, George Stamatoyannopoulos, Job Dekker, Maynard Olson, Michael Dorschner, Patrick Navas, Phil Green)

Mike Snyder (Kevin Struhl, Mark Gerstein, Peggy Farnham, Sherman Weissman)

Rick Myers (Barbara Wold)

Scott Tenenbaum (Luiz Penalva)

Tim Hubbard (Alexandre Reymond, Alfonso Valencia, David Haussler, Ewan Birney, Jim Kent, Manolis Kellis, Mark Gerstein, Michael Brent, Roderic Guigo)

Tom Gingeras (Alexandre Reymond, David Spector, Greg Hannon, Michael Brent, Roderic Guigo, Stylianos Antonarakis, Yijun Ruan, Yoshihide Hayashizaki)

Zhiping Weng (Nathan Trinklein, Rick Myers)

Brenton Graveley (John Rinn, Others)

.. and many senior scientists, postdocs, students, technicians, computer scientists, statisticians and administrators in these groups

NHGRI: Elise Feingold, Mike Pazin, Peter Good
Experimental Assays

Chip-seq (180 TFs + Histone marks; 1770 data sets)
RNA-seq (418)
DNAseq-seq (318)
RNA-Sequencing

Functional data: ChIP-seq

ChIP-exo
Histone Marks

Immunoprecipitation

Sequence and align

ChIP-seq Peak 300-500 bp

Motif (8-12 bp)

Transcription Factor

Antibody
Functional data: DNase-seq

Sequence and align

DNasel hypersensitivity peak

Region of open chromatin

Histone

Histone

Transcription Factor

DNasel

DNase-seq Region of open chromatin
Functional data: DNase footprints

Sequence and align → DNase Footprint

Transcription Factor → Region of open chromatin

DNase I

Histone

Histone
Examples of Signal Tracks

chr5: 39500000
chr5: 39,274,501-40,819,500 (1,545,000 bp)

C9
DAB2
BC026261

PTGER4
TTC33
OSRF
PRKAA1

HUVEC GATA2
HUVEC cFOS
HUVEC Input

TFs

HUVEC

Jurkat

DNase I

Th1

Th2
ENCODE Dimensions

- 200 Cell Lines/Tissues
- 200 Assays (~165 different TFs)
- 2,886 Experiments
- ~10 TeraBases
- ~3000x of the Human Genome

Methods/Factors

Cells

Histone Mods

RNA

Open chromatin

TFs

Methylation
## Mouse Datasets Produced and Released

<table>
<thead>
<tr>
<th></th>
<th># tissue or cell types</th>
<th># experiments</th>
<th># of data sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histone Modifications</td>
<td>33</td>
<td>157</td>
<td>310</td>
</tr>
<tr>
<td>Transcription Factor</td>
<td>29</td>
<td>109</td>
<td>299</td>
</tr>
<tr>
<td>RNA-Seq</td>
<td>69</td>
<td>104</td>
<td>193</td>
</tr>
<tr>
<td>DNase-Seq</td>
<td>55</td>
<td>55</td>
<td>127</td>
</tr>
<tr>
<td>Replication Timing</td>
<td>18</td>
<td>18</td>
<td>33</td>
</tr>
<tr>
<td>ChIP Controls</td>
<td>34</td>
<td>36</td>
<td>108</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>123</strong></td>
<td><strong>479</strong></td>
<td><strong>1070</strong></td>
</tr>
</tbody>
</table>
## ENCODE Uniform Analysis Pipeline

Anshul Kundaje, Qunhua Li, Michael Hoffman, Jason Ernst, Joel Rozowsky, Pouya Kheradpour

<table>
<thead>
<tr>
<th>Gene</th>
<th>Rep1</th>
<th>Rep2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensembl/Havana</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Genes</td>
<td>SPT1 &gt;</td>
<td>DERM1 &gt;</td>
</tr>
</tbody>
</table>

**Mapped reads from production (Bam)**

**Uniform Peak Calling Pipeline (SPP, PeakSeq)**

**Signal Generation**
(read extension and mappability correction)

**IDR Processing, QC and Blacklist Filtering**

**Motif Discovery**

**Stats, GSC enrichments, etc.**

**Signal Aggregation over peaks**

**Segmentation**

**ChromHMM/Segway**

**Self Organising Maps**

<table>
<thead>
<tr>
<th>UW DNase Gm12878 replicate 2</th>
<th>UW DNase Gm12878 replicate 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>22q12.2</td>
<td></td>
</tr>
</tbody>
</table>

**Poor reproducibility**

**Good reproducibility**
# Raw genome coverage of elements

<table>
<thead>
<tr>
<th>Element Type</th>
<th>Coverage</th>
<th>Cumulative Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exons</td>
<td>3%</td>
<td>3%</td>
</tr>
<tr>
<td>Chip-seq bound motifs</td>
<td>4.5%</td>
<td>5%</td>
</tr>
<tr>
<td>DNaseI Footprints</td>
<td>5.7%</td>
<td>9%</td>
</tr>
<tr>
<td>Chip-seq bound regions</td>
<td>8.1%</td>
<td>12%</td>
</tr>
<tr>
<td>DNaseI HS regions</td>
<td>15.2%</td>
<td>19.4%</td>
</tr>
<tr>
<td>Histone Modifications (*)</td>
<td>44%</td>
<td>49%</td>
</tr>
<tr>
<td>RNA</td>
<td>62%</td>
<td>80%</td>
</tr>
<tr>
<td>(* excluding broad marks)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*(Union over all experiments and cell types)*
Saturation

Most aggressive fit for saturation suggests a maximum of 50% of elements discovered

Likely to be lower due to inaccessible cell types etc
ENCODE Integrative Segmentations

~7 Major genome segments
25 “elaborations”
1,000s of details

Well Known: TSS, Gene Start, Gene Bodies

New Info: “Enhancers” (2 states), Insulators

Unexpected: Specific Gene End
Experimental Confirmation of New Enhancers

Jason Gertz, Barbara Wold, Rick Myers, Len Pennacchio

Mann Whitney 0.003 HMM vs Background
1e-7, HMM vs Naïve or Biologist picks
Myers Lab

53% hit rate in Mouse Assay
Pennacchio Lab
Many other stories…

TF Co association and Regulatory Code
Mike Snyder + Mark Gerstein

DNaseI footprints – John Stam.
DNA Methylation – Rick Myers

Splicing/Histone interaction (Roderic Guigo)

RNA landscape
Tom Gingeras
What’s Next- Enhancing ENCODE

1) Deep analysis of six cell lines/tissues
   Cancer: K562, MCF7
   Diploid: GM12878, H1 ES Cells
   Tissues: Liver, Brain

2) More limited coverage of hundreds of other cell types and tissues (RNA-Seq, DNAaseHS, etc)

3) Some mouse data

Many investigators same as ENCODE2 + Brenton Graveley-a few others
What’s Next- Species Comparisons

1) modENCODE/ENCODE Comparison
   Worms, Flies, Human
   Hundreds of worm and fly datasets (e.g. >250 C. elegans TF ChIP-Seq datasets)

2) MouseENCODE-humanENCODE Comparison
The ENCODE Consortium

Brad Bernstein (Eric Lander, Manolis Kellis, Tony Kouzarides)

Ewan Birney (Jim Kent, Mark Gerstein, Bill Noble, Peter Bickel, Ross Hardison, Zhiping Weng)

Greg Crawford (Ewan Birney, Jason Lieb, Terry Furey, Vishy Iyer)

Jim Kent (David Haussler, Kate Rosenbloom)

John Stamatoyannopoulos (Evan Eichler, George Stamatoyannopoulos, Job Dekker, Maynard Olson, Michael Dorschner, Patrick Navas, Phil Green)

Mike Snyder (Kevin Struhl, Mark Gerstein, Peggy Farnham, Sherman Weissman)

Rick Myers (Barbara Wold)

Scott Tenenbaum (Luiz Penalva)

Tim Hubbard (Alexandre Reymond, Alfonso Valencia, David Haussler, Ewan Birney, Jim Kent, Manolis Kellis, Mark Gerstein, Michael Brent, Roderic Guigo)

Tom Gingeras (Alexandre Reymond, David Spector, Greg Hannon, Michael Brent, Roderic Guigo, Stylianos Antonarakis, Yijun Ruan, Yoshihide Hayashizaki)

Zhiping Weng (Nathan Trinklein, Rick Myers)

Additional ENCODE Participants: Elliott Marguiles, Eric Green, Job Dekker, Laura Elnitski, Len Pennachio, Jochen Wittbrodt

... and many senior scientists, postdocs, students, technicians, computer scientists, statisticians and administrators in these groups

NHGRI: Elise Feingold, Mike Pazin, Peter Good