Examples of how ENCODE facilitates biomedical research

ENCODE Users Meeting
Potomac, MD
July 1, 2015

Richard M. Myers
rmyers@hudsonalpha.org
http://www.hudsonalpha.org/
Disclosure

Our group has been part of the ENCODE Consortium since it began in 2003

Rick Myers
Barbara Wold
Ross Hardison

Eric Mendenhall
Ali Mortazavi
Tim Reddy
Goals of ENCODE

Annotate the human genome

Disseminate data to researchers everywhere
5 examples of how we use ENCODE data to help in our research on human diseases
1. Discovering the causes of undiagnosed genetic diseases
Childhood genetic disorders

1.5-3% of kids worldwide are born with 1 or more of:

- intellectual disability
- developmental delay
- heart defects
- craniofacial and skeletal abnormalities
- severe autism
- seizures

The vast majority of these problems have genetic causes
Diagnostic challenges for childhood genetic disorders

Inaccurate or undetermined causes (i.e., diagnoses) are a major hardship:

- Years of expensive, invasive, and futile testing
- Impossible to predict disease progression, symptoms
- Treatment decisions are complicated
- Slows research into developing new therapies
- Impacts family planning
- Results in feelings of parental guilt and lack of control

Thus, identifying the root genetic causes is essential
HudsonAlpha Pediatric Genetics Project

Sequence whole genomes of 500 children with developmental/intellectual delay of unknown etiology (and both parents’ genomes too)
**Exome results so far**

**Exome sequencing completed for 171 families**

Definitive genetic diagnosis in 25% of the children
- Pitt-Hopkins syndrome
- Dravet syndrome
- Rett syndrome
- Rubinstein-Taybi syndrome
- Noonan-like syndrome
- Many never-described causes

>20% of families receive uncertain genetic findings that will likely be definitively diagnostic in the future
Whole genome sequencing of trios

Illumina X Ten sequencers: $ of 30X WGS = $ of exome

We have completed WGS of 30 trios in our Childhood Genetics Project

Results:

Diagnostic rate is higher

Identified at least 3 cases where regulatory mutations were the causes

We relied heavily on ENCODE data to identify functional regulatory segments
Annotating genetic variants

Problem:

HUGE number of sequence variants in each individual

Most are not important

How to find which variants have an effect on:
  The molecular/biochemical function of the gene
  The organism
A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher\textsuperscript{1,5}, Daniela M Witten\textsuperscript{2,5}, Preeti Jain\textsuperscript{3,4}, Brian J O’Roak\textsuperscript{1,4}, Gregory M Cooper\textsuperscript{3} & Jay Shendure\textsuperscript{1}

CADD integrates many features to give a single pathogenicity score.
Promoter mutations that cause B-thalassemia
Enhancer mutations that cause pancreatic agenesis
Enhancer mutations that cause limb defects
Use the CADD webservice!

http://cadd.gs.washington.edu
2. Understanding renal cell carcinoma

ENCODE data were instrumental in helping us identify regions of the genome that are ~100% accurate diagnostic markers for kidney cancer

(and even for prognosis of different subtypes)
We measured DNA methylation and copy number variants in 135 kidney tumors and matched non-tumor kidney tissues
Top 20 DNA methylation markers

All subtypes

Kidney Tumor
Normal Tissue
Clear Cell
Other Subtypes
Normal Tissue

20 CpGs

Cytosine
5-Methyl Cytosine
DNA methylation patterns are highly accurate at predicting patients with renal cell carcinoma

ROC curves of DNA methylation results from 135 tumor and matched non-tumor samples from RCC patients

HAIB AUC = 0.990  
TCGA AUC = 0.972

Apply these assays to urine or blood as a routine screening for early detection of kidney cancer

Schwarzenbach et al. Nature Reviews Cancer 11, 426-437
Integrating genomic signatures

Copy Number Variation (CNV)

DNA Methylation (DNAme)

CNV and DNAme
Example: MSH4 gene

~16% of patients

~28% of patients

~44% of patients total
3. Using ENCODE TF data to prioritize cancer genetics and functional genomics data
Using ENCODE data to find cancer regulators

Genomic assays often reveal thousands of dysregulation events in cancer.

These widespread genomic changes may be regulated by a few key transcription factors.
Differentially expressed genes in cancer are enriched for particular TFs

Genes differentially expressed in prostate cancer compared to normal prostate tissue are enriched for **EZH2**, **SUZ12**, and **CTPB2** binding sites (adjusted p-value < 0.05) and actual binding events (from ENCODE ChIP data).
Intersect transcription factor binding sites from the ENCODE Project with genomic regions specifically unmethylated in basal breast cancer

- 70 cell lines
- 149 transcription factors

Scale:
- chr7:
  - Unmethylated in Basal
  - Unmethylated in Luminal

Methylated levels:
- 0%
- 50%
- 100%
- 25%

Genomic regions:
- EGFR
- Luminlal
- 7p11.2
- ERG
- E2F1
- TBP
- Pol2
- Poiz2
- Poiz2(phosphiSEZ)
- E2F6
- Mxi1(lb-HH)
- Max
- GTF2F1(RAP74)
- CEBPB
- EzF1
- SMC3(_db9263)
- STAT3

Richard M. Myers
Master regulators (?) of different breast cancer subtypes

Intersect gene regulatory regions containing subtype-associated methylation with binding sites of 149 transcription factors in ENCODE datasets

Significantly enriched binding sites:

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>Fold Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estrogen Receptor</td>
<td>6.9</td>
</tr>
<tr>
<td>FOXA1</td>
<td>8.1</td>
</tr>
<tr>
<td>GATA3</td>
<td>10.3</td>
</tr>
</tbody>
</table>

Percent Methylated

100 83 66 50 34 17 0
Master regulators (?) of different breast cancer subtypes

Intersect gene regulatory regions containing subtype-associated methylation with binding sites of 149 transcription factors in ENCODE datasets.

Significantly enriched binding sites:

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>Fold Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estrogen Receptor</td>
<td>6.9</td>
</tr>
<tr>
<td>FOXA1</td>
<td>8.1</td>
</tr>
<tr>
<td>GATA3</td>
<td>10.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>Fold Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAT3</td>
<td>4.8</td>
</tr>
<tr>
<td>GR (glucocorticoid receptor)</td>
<td>4.2</td>
</tr>
</tbody>
</table>
4. Using RNA-seq to identify drug targets
Transcript fusions in cancer

Breast Cancer Res Treat
DOI 10.1007/s10549-014-3019-2

PRECLINICAL STUDY

Recurrent read-through fusion transcripts in breast cancer

Katherine E. Varley · Jason Gertz · Brian S. Roberts · Nicholas S. Davis · Kevin M. Bowling · Marie K. Kirby · Amy S. Nesmith · Patsy G. Oliver · William E. Grizzle · Andres Forero · Donald J. Buchsbaum · Albert F. LoBuglio · Richard M. Myers

K-T Varley with collaborators at UAB Comprehensive Cancer Center
3 fusion transcripts produce fusion proteins located in the cell membrane

Potential therapeutic: Use drug-antibody complexes to direct a cellular toxin exclusively to cancer cells
5. Which TF binding events are functionally important?
Using expression assays to identify functional transcription elements

(eespecially long-distance ones)

Dan Savic, Brian Roberts, Chris Partridge, Barak Cohen, Greg CooperJay Gertz, Rick Myers

Test thousands of ENCODE-identified putative elements (based on TF binding, chromatin marks, etc.) in an ultra-high throughput reporter assay
Massively parallel reporter assay
Cis-Regulatory Element sequencing (CRE-seq)

Barcode abundance (sequence count) is a proxy for test sequence activity
Findings

RNAP2 at promoter-distal TF sites is a very strong mark of active regulatory elements
3. Using genomics to predict which patients will respond to various treatments
Clinical trial of a novel combination of drugs in ER+ breast cancer

Gene expression patterns in **responders** and **non-responders** during clinical trial of Letrozole (anti-estrogen) and Avastin (anti-angiogenesis)

145 genes q-value < 0.05

K-T Varley, Andres Forero, Al LoBuglio, Don Buchsbaum, Rick Myers