How can we “read” the human genome sequence?

- No instruction manual/punctuation marks
- Evolutionary conservation helps to identify functionally important regions
  - ~5% conserved; ~1.5% protein coding
- Moderately good at identifying protein-coding regions, but fine structures difficult to predict from sequence
- Regulatory regions can be very far away from genes
- Need unbiased experimental investigation to identify all functional regions
Goal: To compile a *comprehensive encyclopedia* of all of the sequence features in the human genome and in the genomes of selected model organisms

**ENCODE**
- **Pilot Project Phase (9/03 – 9/07)**
  Studied defined 1% of the human genome sequence using existing technologies
- **Production Phase (9/07 – 9/11)**
  New/continued pilot projects and expansion to whole genome studies in human

**modENCODE (5/07 – 5/11)**
- Production projects to comprehensively identify functional elements in the genomes of *C. elegans* and *D. melanogaster*

**Mouse ENCODE (9/09 – 9/11 with ARRA funds)**
- Limited production projects to identify functional elements in the mouse genome to inform annotation of human genome

**Technology Development (9/03 -9/10)**
- Focused on less well-studied functional elements; funded solicitations in 2003, 2004, 2007
Lots of data and data types...

...... generated by:

- RNA-seq
- RNA-array
- TF ChIP-seq
- Histone modif ChIP-seq
- DNaseHS-seq
- FAIRE-seq
- Methyl-seq
- Methyl27-bisulfite
- 1M SNP genotyping
  (+ WGS for GM12878)
Progress

• Large-scale data production ongoing
  – ENCODE: 959 datasets submitted
  – modENCODE: 1170 datasets submitted

• Analysis requires development of:
  – Common data reporting formats
  – Data standards
  – Analytical tools

• Integrative analyses for each species ongoing
  – Long-term plans for integration of fly/worm; fly/worm/human

• Follow up and expand on findings from pilot
  ➢ Human genome is pervasively transcribed
  ➢ Many functional elements are seemingly unconstrained across mammalian evolution
How will the catalogs of functional elements be used?

1. Enhance understanding of regulation of gene expression on a spatial, temporal and quantitative level
   - Who are the players?
   - How do they interact?
   - How do variants affect gene expression?
   - Can we predict gene expression from sequence?
   - Can we manipulate gene expression?
2. Enhance understanding of genetic basis of disease
   - Many genome-wide association studies (GWAS) find SNPs in non-coding regions
   - How do SNPs/mutations in non-coding regions alter gene expression and contribute to disease?

3. Enhance understanding of epigenetic contributions to disease
   - Epigenomics (NIH Common Fund)
Functional Element Variation

• Genome-wide differences in transcription factor bindings sites between individual
  • RNA polymerase II: 25% difference
  • NF Kappa B: 7.5%
    – Binding differences frequently associated with SNPs and SVs, and differences in gene expression
    – Suggests functional consequences of binding variation

• Individual-specific and allele-specific chromatin signatures in humans
  • 10% active chromatin sites individual specific
  • 10% active chromatin sites allele-specific
  • Presence of individual-specific DHS site near TSS correlated with expression
    – Strong genetic component for individual and allele-specific differences

Multiple regions in 8q24 have alleles predisposing to many cancers (e.g., prostate, breast and colon)

Regions far from annotated genes; unknown biological function

Profiled risk region (RNA expression, histone modifications, binding sites for Pol II & androgen receptor)

Several enhancers identified

SNP found in one enhancer within FoxA1 TF binding site

Prostate cancer risk allele facilitating stronger FoxA1 binding and stronger androgen response

Pilot Project Findings

- The human genome is pervasively transcribed.
- Many novel non-protein-coding transcripts and transcription start sites identified.
- Regulatory sequences that surround transcription start sites are symmetrically distributed, with no bias towards upstream regions.
- Chromatin accessibility and histone-modification patterns are highly predictive of both the presence and activity of transcription start sites.
- Distal DNaseI hypersensitive sites have characteristic histone modification patterns that reliably distinguish them from promoters; some of these distal sites show marks consistent with insulator function.
- 5% of the bases in the genome can be confidently identified as being under evolutionary constraint in mammals.
  - For ~60% of these constrained bases, there is evidence of function based on the results of the experimental assays.
- Many functional elements are seemingly unconstrained across mammalian evolution.
## ENCODE Funding

(4 years)

<table>
<thead>
<tr>
<th>Project</th>
<th>ENCODE</th>
<th>modENCODE</th>
<th>Mouse ENCODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production /Pilot Grants</td>
<td>$81.5M</td>
<td>$57.2M</td>
<td>$4.2M</td>
</tr>
<tr>
<td></td>
<td>$7.5M</td>
<td>$5.2M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$89.0M</td>
<td>$62.4M</td>
<td></td>
</tr>
<tr>
<td>Data Coord. Center</td>
<td>$5.0M</td>
<td>$5.0M</td>
<td></td>
</tr>
<tr>
<td>Data Analysis Center</td>
<td>$5.0M</td>
<td>$2.8M (2 yrs)</td>
<td>*2 year ARRA funds</td>
</tr>
<tr>
<td>RM Epigenomics Program</td>
<td>ENCODE (NHGRI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Goal</strong></td>
<td>Generate comprehensive catalog of functional elements in the human genome</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cell types</strong></td>
<td>7 common cell types (cell lines, primary tissues and one hES cell line); Addnl ~9-80 cell sources depending on functional element</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Epigenetic marks</strong></td>
<td>Focus on Histone modifications, DNA methylation, and small ncRNAs</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Disease Relevance</strong></td>
<td>Data is resource to be mined by researchers, no disease focus</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- RM Epigenomics Program
  - Understanding epigenetic basis of disease
  - >80 normal human cell types, selected based on relevance to disease
  - Focus on Histone modifications, DNA methylation, and small ncRNAs
  - RFA on Epigenomics of Human Health and Disease
Epigenomics of Health and Disease
Investigates variety of human diseases

- Alzheimer’s disease
- cognitive decline
- tumor stem cells
- breast cancer
- glaucoma
- bipolar
- autism/environment
- breast cancer/BPA
- insulin resistance
- Barrett’s esophagus
- chronic kidney disease
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- insulin resistance
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosis/vascular disease
- asthma
- schizophrenia
- IUGR/LGA
- FSHD
- SLE (lupus)
- atherosclerosi