NGS Approaches to Epigenomics

Yuzhen Ye (yye@indiana.edu)
School of Informatics & Computing, IUB
Contents

- Background: chromatin structure & DNA methylation
- Epigenomic projects
  - ENCODE
  - modENCODE
  - Human epigenome atlas
- Techniques
  - ChIP-Seq
  - MeDIP (MeDIP-chip & MeDIP-seq)
- Data analysis
  - Peak detection
  - Decoding problem (I529)
Views of genomes beyond strings of ATCG

- Genomic DNA is packaged into a complex molecular structure known as chromatin. This structure mediates the interaction between the genome and all types of regulatory and transcriptional molecules.

- In vertebrate genomes, methylation at position 5 of the cytosine in CpG dinucleotides is a heritable “epigenetic” mark that has been connected with both transcriptional silencing and imprinting
  - Ref: DNA methylation patterns and epigenetic memory (Genes & Dev. 2002. 16: 6-21)
A few terminologies

- **Epigenetics**
  - “the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being.” (Conrad Waddington, 1940s)
  - “the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence.” (Epigenetics: Definition, Mechanisms and Clinical Perspective, 2010)
  - “An epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence.” (An operational definition of epigenetics, 2009)

- **Epigenetic modifications**
  - histone variants, posttranslational modifications of amino acids on the amino-terminal tail of histones, and covalent modifications of DNA bases

- **Epigenome**
  - refers to the different chromatin states at the whole genome level

- **Epigenomics**—studies of epigenomes
The epigenetic pathway.

Epigenator  
- e.g. Differentiation signals;  
- Temperature variations

Epigenetic Initiator  
- e.g. DNA binding factors;  
- Non-coding RNAs

Epigenetic Maintainer  
- e.g. Histone/DNA modifiers;  
- Histone variants

Extracellular  

Cytoplasm  

Nucleus

ENCODE

- Encyclopedia of DNA Elements
  - “The ENCODE Consortium is integrating multiple technologies and approaches in a collective effort to discover and define the functional elements encoded in the human genome, including genes, transcripts, and transcriptional regulatory regions, together with their attendant chromatin states and DNA methylation patterns.”

- Initial phase launched in 2003—1% of the human genome
  - Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project (Nature, June 13, 2007)
Figure 1. The Organization of the ENCODE Consortium.

A.

http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046
### Table 1. Experimental assays used by the ENCODE Consortium.

<table>
<thead>
<tr>
<th>Region/Feature</th>
<th>Method</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene annotation</td>
<td>GENCODE</td>
<td>Wellcome Trust</td>
</tr>
<tr>
<td>PolyA+ coding regions</td>
<td>RNA-seq; tiling DNA microarrays; PET</td>
<td>CSHL; Stanford/Yale//Harvard; Caltech</td>
</tr>
<tr>
<td>Total RNA coding regions</td>
<td>RNA-seq; tiling DNA microarrays; PET</td>
<td>CSHL</td>
</tr>
<tr>
<td>Coding regions in subcellular RNA fractions (e.g. nuclear, cytoplasmic)</td>
<td>PET</td>
<td>CSHL</td>
</tr>
<tr>
<td>Small RNAs</td>
<td>short RNA-seq</td>
<td>CSHL</td>
</tr>
<tr>
<td>Transcription initiation (5’-end) and termination (3-end”) sites</td>
<td>CAGE; dITAGs</td>
<td>RIKEN, GIS</td>
</tr>
<tr>
<td>Full-length RNAs</td>
<td>RACE</td>
<td>University of Geneva; University of Lausanne</td>
</tr>
<tr>
<td>Protein-bound RNA coding regions</td>
<td>RIP; CLIP</td>
<td>SUNY-Albany; CSHL</td>
</tr>
</tbody>
</table>

#### Transcription Factors/Chromatin

<table>
<thead>
<tr>
<th>Elements/Regions</th>
<th>Method(s)</th>
<th>Group(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription Factor Binding Sites (TFBS)</td>
<td>ChIP-seq</td>
<td>Stanford/Yale/UC-Davis/Harvard; HudsonAlpha/Caltech; Duke/UT-Austin; UW; U. Chicago/Stanford</td>
</tr>
<tr>
<td>Chromatin structure (accessibility, etc.)</td>
<td>DNasel hypersensitivity; FAIRE</td>
<td>UW; Duke; UNC</td>
</tr>
<tr>
<td>Chromatin modifications (H3K27ac, H3K27me3, H3K36me3, etc.)</td>
<td>ChIP-seq</td>
<td>Broad; UW</td>
</tr>
<tr>
<td>DNasel footprints</td>
<td>Digital genomic footprinting</td>
<td>UW</td>
</tr>
</tbody>
</table>

#### Other Elements/Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Method(s)</th>
<th>Group(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA methylation</td>
<td>RRBS; Illumina Methyl27; Methyl-seq</td>
<td>HudsonAlpha</td>
</tr>
<tr>
<td>Chromatin interactions</td>
<td>5C; CHIA-PET</td>
<td>UMass; UW; GIS</td>
</tr>
<tr>
<td>Genotyping</td>
<td>Illumina 1M Duo</td>
<td>HudsonAlpha</td>
</tr>
</tbody>
</table>

http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046
ENCODE data

chr21:33,031,597-33,041,570 9,974 bp.

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
Figure 4. ENCODE chromatin annotations in the HLA locus.

http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046
Figure 5. Occupancy of transcription factors and RNA polymerase 2 on human chromosome 6p as determined by ChIP-seq.


http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046
http://www.modencode.org/

“The modENCODE Project will try to identify all of the sequence-based functional elements in the *Caenorhabditis elegans* and *Drosophila melanogaster* genomes.”
Successive releases of the Atlas will provide progressively more detailed insights into locus-specific epigenomic states, including histone marks and DNA methylation marks across specific tissues and cell types, developmental stages, physiological conditions, genotypes, and disease states.
ChIP-Seq

- Chromatin immunoprecipitation (ChIP) followed by high-throughput DNA sequencing (ChIP-seq) has become a valuable and widely used approach for mapping the genomic location of transcription-factor binding and histone modifications in living cells.
  - Genome-Wide Mapping of in Vivo Protein-DNA Interactions (Science, 2007); 1946 binding sites of the Neuron-restrictive silencer factor (NRSF) were mapped at ~50bp resolution

- There are considerable differences in how these experiments are conducted, how the results are scored and evaluated for quality, and how the data and metadata are archived for public use.
  - Genome Res. 2012 Sep;22(9):1813-31
Barcoded ChIP-Seq

Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing (BMC Genomics 2009, 10:37)
DNase-seq

- DNase digestion followed by sequencing.
- DNase I hypersensitive sites (DHS), short regions of chromatin that are highly sensitive to cleavage by DNase I, typically occur in nucleosome free (nucleosome-depleted) regions as a result of transcription factor binding.
- DNA sequence motif analysis on DHS data was proposed as a method for discovering the binding sites of multiple transcription factors in a single experiment.
- DNase-seq profile resemble to some extent the data from ChIP-seq, with important differences (Front. Genet., 31 October 2012)
Genome-wide DNA methylation profiling

- **Restriction enzyme-based methods**
  - Use one or more enzymes that will restrict DNA only if it is unmethylated (e.g. HpaII or NotI), or methylated (e.g. McrBC).
  - Limited to the analysis of CpG sites located within the enzyme recognition site(s).

- **Bisulfite-conversion based approaches**
  - Unmethylated cytosines are converted to uracil; offer single CpG resolution; the gold standard for DNA methylation analysis
  - Con: reduction of sequence complexity following bisulfite conversion (Bi-chip) & Bi-seq approach is expensive.
  - Align BS-treated reads to a reference genome

- **Immunoprecipitation-based methods**
  - Use either 5-methylcytosine-specific antibodies (MeDIP) or methyl-binding domain proteins, to enrich for the methylated (or unmethylated) fraction of the genome.
Methylation analysis by DNA immunoprecipitation (MeDIP)

MeDIP can be coupled with microarray or high-throughput sequencing

Image from wikipedia
Long-range chromatin interaction

Long-range Chromatin interactions: Chromosome Conformation Capture Carbon Copy (5C)

CHIP-Seq: peak detection
Chromatin-state decoding

- Automated mapping of large-scale chromatin structure in ENCODE

- ChromHMM: automating chromatin-state discovery and characterization
NCBI Epigenomics: What’s new for 2013