Regulatory mapping and the genetics of human disease

Matthew Maurano
Senior Fellow, Stamatoyannopoulos lab
Dept. of Genome Sciences
University of Washington

From July:
Assistant Professor
NYU Dept. of Pathology & Institute for Systems Genetics
Overview

1. Introduction to mapping the regulatory genome

2. Four key characteristics of regulatory DNA

3. Practical aspects of studying regulatory variation
The living genome
The living genome

Chromatin

Genes

RNA

Regulatory DNA

Transcription factors
The living genome

Chromatin

Genes

RNA

Regulatory DNA

Transcription factors
The living genome

- promoters
- enhancers
- silencers
- insulators
- etc.
Mapping regulatory DNA

- promoters
- enhancers
- silencers
- insulators
- etc.
Mapping regulatory DNA

DNase I hypersensitive site (DHS)

Regulatory DNA
Mapping regulatory DNA

An expansive atlas of regulatory DNA landscapes is now available

>400 cell/tissue types and developmental states now available

~95% from primary cells and tissues
The living genome c.2015

~55K
~21K protein coding
~34K long noncoding
>1M smaller non-coding RNAs

>3 million

>20 million
What do all these elements do?
Some have assay-defined functions…

- promoters
- enhancers
- silencers
- insulators
- etc.
...but most have complex activities that are still being sorted out

Mercer et al., Nature Genetics 2013
Overview

1. Introduction to mapping the regulatory genome

2. Four key characteristics of regulatory DNA

3. Practical aspects of studying regulatory variation
1) Regulatory DNA is densely populated by transcription factors (TFs)
DHSs reflect the cumulative action of transcription factor binding
2) Actuation of regulatory DNA is extremely cell-selective
Actuation of regulatory DNA is extremely cell selective
Actuation of regulatory DNA is extremely cell selective

Most of the genome looks like this...
Actuation of regulatory DNA is extremely cell selective

...not this
3) Regulatory landscapes encode extensive memories of prior cell states.
Regulatory DNA also shows pronounced lineage-selectivity

Stergachis et al. Cell 2013
4) The function of regulatory elements is finely tuned to both cell type and genomic location
The function of regulatory elements is finely tuned to both cell type and genomic location.

Spatial tuning of gene expression is a major feature of vertebrate genomes. Known dependencies include:

- Distance from target gene(s)
- Presence of intervening gene(s)
- Presence of intervening regulatory elements (e.g., insulators)

For any given element, the requirement for cooperative TF occupancy imposes numerous constraints including:

- Specific combination of TFs present
- Spacing and orientation of TFs
- Modification state of surrounding nucleosomes
Mapping functional regulatory variants \textit{in vivo}

Of 50M variants in dbSNP138, only 483,415 are strongly predicted to affect TF occupancy.
Overview

1. Introduction to mapping the regulatory genome

2. Four key characteristics of regulatory DNA

3. Technical aspects of genome-wide chromatin profiling
   - 4 key characteristics of regulatory genomics datasets
Library complexity is a cardinal sign of high-quality data

- Low-complexity libraries caused by inadequate input material or inefficient library construction
- Recognized during amplification: 8-10 PCR cycles is optimal, >18 is excessive
- Duplicate read rate (samtools and Picard) should be <2% for initial sequencing
- Can attenuate effect by removing duplicates (but ideally use high-complexity data to start with)
High-quality data must be adequately sequenced

Sources of unexpectedly low effective depth:
• Low enrichment
• Mitochondrial contamination
• Adapter contamination

➔ Tools like FastQC

<table>
<thead>
<tr>
<th>Compartment</th>
<th>Num. sites</th>
<th>Width (bp)</th>
<th>Genome covered</th>
<th>Sequencing depth (reads)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNase I hypersensitive sites</td>
<td>180,000</td>
<td>150</td>
<td>0.90%</td>
<td>40M-400M</td>
</tr>
<tr>
<td>Histone ChIP-seq (focal marks)</td>
<td>50,000</td>
<td>300</td>
<td>0.50%</td>
<td>20M-40M</td>
</tr>
<tr>
<td>TF ChIP-seq peaks</td>
<td>2,000</td>
<td>150</td>
<td>0.01%</td>
<td>5M-10M</td>
</tr>
</tbody>
</table>
Enrichment is measured by signal-to-noise ratio

Determinants of enrichment:

• Cell manipulation: minimizing shock response and degradation
• Inherent characteristics of method (esp. antibody for ChIP-seq)
• Enrichment can be quantified as SPOT score (signal proportion of tags) or as fragment length
  – regulatory elements are enriched in short <125bp fragments relative to nucleosomal DNA
Resolution to detect focal TF binding

A SNP must directly perturb TF recognition sequence
Summary

• Adequate library complexity – % duplicates
• Adequate sequencing depth
• Strong enrichment – signal-to-noise
• Resolution of protein-DNA binding interface
• High replicate concordance
• Appropriate cell type and culture conditions
Further reading

Data sources

• https://www.encodeproject.org/
• http://www.roadmapepigenomics.org/
• NCBI GEO, GTEx, EU Blueprint, etc.

Browsing raw data

• UCSC Genome Browser - https://genome.ucsc.edu/
• many others

Working with raw data, peak calls, SNPs, etc:
• https://bedops.readthedocs.org/en/latest/
The NYU Institute for Systems Genetics is recruiting computational and experimental students, postdocs, and faculty

Funding: NHGRI (ENCODE), NIH Common Fund (Roadmap Epigenomics), NIMH (NRSA)