Introduction to first homework

Kaur Alasoo
21 February 2019
Sequencing is the new imaging!

- **In computing:** Dedicated hardware devices replaced by generic hardware + software application (barcode readers, pedometers)

- **In biology:** Dedicated biochemical experiments replaced by high-throughput measurements which are often cheaper and/or more accurate.

- Sequencing is the new imaging! Image sensor capture light, sequencers capture DNA.
Deciphering gene function

- **Strategy 1: ‘guilty by association’**
  - Differences in gene expression over time, between various conditions.

- **Strategy 2: common genetic variation**
  - QTL mapping, colocalisation, mendellian randomisation

- **Strategy 3: direct perturbation**
  - CRISPR screens
The human genome
Solely 2% of the human genome encodes proteins.
Genes can code for multiple different *alternatively spliced* transcripts (mRNA)
### Human genome

<table>
<thead>
<tr>
<th>Type</th>
<th>Count (incl readthrough)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding genes</td>
<td>20,338 (incl 562 readthrough)</td>
</tr>
<tr>
<td>Non coding genes</td>
<td>22,521</td>
</tr>
<tr>
<td>Small non coding genes</td>
<td>5,363</td>
</tr>
<tr>
<td>Long non coding genes</td>
<td>14,720 (incl 238 readthrough)</td>
</tr>
<tr>
<td>Misc non coding genes</td>
<td>2,222</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>14,638 (incl 6 readthrough)</td>
</tr>
<tr>
<td>Gene transcripts</td>
<td>200,310</td>
</tr>
</tbody>
</table>

[Link to Human genome information](http://www.ensembl.org/Homo_sapiens/Info/Annotation)

### Mouse genome

<table>
<thead>
<tr>
<th>Type</th>
<th>Count (incl readthrough)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding genes</td>
<td>22,612 (incl 251 readthrough)</td>
</tr>
<tr>
<td>Non coding genes</td>
<td>15,402</td>
</tr>
<tr>
<td>Small non coding genes</td>
<td>5,531</td>
</tr>
<tr>
<td>Long non coding genes</td>
<td>9,308 (incl 63 readthrough)</td>
</tr>
<tr>
<td>Misc non coding genes</td>
<td>563</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>12,363 (incl 6 readthrough)</td>
</tr>
<tr>
<td>Gene transcripts</td>
<td>133,944</td>
</tr>
</tbody>
</table>

[Link to Mouse genome information](http://www.ensembl.org/Mus_musculus/Info/Annotation)
Explore the human genome!

- Open IGV.
- Change the genome version to hg38 (top left corner)
- Search for the Sonic Hedgehog (SHH) gene.
- Right click on the gene and select Expanded.
- How many annotated transcript does SHH gene have?
- How far is the closest neighbouring gene?
- How about the lactase (LCT) gene that allows you to drink milk?
DNA sequencing
Illumina sequencing
Immobilizing single molecules

1) Prepare Sequencing Library
2) Seed Flow Cell with Single Molecules

Fragment DNA

Add adaptors, size select, PCR

denature
Problem: Half of the human genome is comprised of repeats

(first bit of human chromosome 1)
Paired-end sequencing: A molecular hack to sequence longer fragments

Genomic DNA

Shear to desired length (~400bp)

DNA fragments

Ligate adapters, size select

Sequencing library

Illumina GA2 clusters on a flow-cell

Millions to billions of paired-end reads (readpairs)

Slide from Ira Hall
Aligning pairs of reads to the reference genome

1) a readpair

5’ GGTGTA GTA TATTTTCCTTTTACACTCCTTGACCACC C5’

| ATGTCGAATTTAAATTTGGTATCAATGGTTTTGGTCGTATCGGCCGTATCGTATTCCGTGCAGACAACAC |
| TATGCAATTTAAATTTGGTATCAATGGTTTTGGTCGTATCGGCCGTATCGTATTCCGTGCAGCACAACAC |
| GGACTGAAACTTCATCTGTCTTTATAGATATGCGTGCAGC |

2) a reference genome on a computer

3) Alignment of the readpair to the reference genome gives coordinates describing where in the human genome the readpair came from
Formats
(Mostly) all technologies yield DNA sequences in FASTQ format

@seq1
ACCTTCGAACGGCGGGGGTTACAA
+
!''*(((**+))%%%++).1***
@seq2
TGGAACCGAACGGCCCCGGTTACAT
+
!''*!!!!****))++++++).1***
And so on...
The FASTQ format. Welcome to a minor hell.

A “standard” format for storing and defining sequences from next-generation sequencing technologies.

Sequence ID  @SEQ_ID
Sequence   GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
Quality scores  +

http://en.wikipedia.org/wiki/FASTQ_format
If (almost) all human cells have the same DNA, then why do they look so different?
Figure 1: Overview of hematopoiesis, UKBB GWAS, and fine-mapping.

(a) Schematic of the human hematopoietic hierarchy showing the primary cell types analyzed in this work. ATAC-seq and RNA-seq were collected for all sorted cell types except granulocytes (ATAC-seq, RNA-seq) and megakaryocytes (RNA-seq). Colors used in this schematic are consistent throughout all figures. Mono, monocyte; gran, granulocyte; ery, erythroid; mega, megakaryocyte; CD4, CD4+ T cell; CD8, CD8+ T cell; B, B cell; NK, natural killer cell; mDC, myeloid dendritic cell; pDC, plasmacytoid dendritic cell. The 16 terminal blood traits that were genetically fine-mapped are shown below the hierarchy.

(b) Schematic of UKBB GWAS and fine-mapping approach. Briefly, blood traits from ~113K individuals were fine-mapped allowing for multiple causal variants and using imputed genotype dosages as reference LD.

(c) Number of fine-mapped 3MB regions for each trait with the best posterior probability for a variant being causal indicated.

(d) Kernel density of the expected number of causal variants for each region.

Interrogation of human hematopoiesis at single-cell and single-variant resolution.
Macrophages eat other cells

The macrophage marches on its phagosome: dynamic assays of phagosome function
Russel et al (2009)
Macrophages eat other cells and pathogens

The macrophage marches on its phagosome: dynamic assays of phagosome function
Russel et al (2009)
Macrophages eat other cells and pathogens

(a) Resting macrophage
- Nucleus
- Apoptotic body
- No inflammatory signals
- No inflammation

(b) Activated macrophage
- Pathogen
- LPS (lipopolysaccharide)
- Inflammatory signals such as LPS and IFNγ
- Activated macrophage
- Superoxide burst
- Proteolysis
- MHC class II
- Inflammation and enhanced antigen presentation

The macrophage marches on its phagosome: dynamic assays of phagosome function
Russel et al (2009)
LPS induces strong response in macrophages

Transcriptional control of the inflammatory response
Medzhitov et al (2009)
Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response

Kaur Alasoo1,4*, Julia Rodrigues1, Subhankar Mukhopadhyay1, Andrew J. Knights1, Alice L. Mann1, Kousik Kundu1,3, HIPSCI Consortium2, Christine Hale1, Gordon Dougan1 and Daniel J. Gaffney1*
Experimental design

Macrophages

- N: Naive
- I: IFNγ (18h)
- S: Salmonella (5h)
- I+S: IFNγ (18h) + Salmonella (5h)

Sample sizes

<table>
<thead>
<tr>
<th>RNA</th>
<th>ATAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td>42</td>
</tr>
<tr>
<td>84</td>
<td>41</td>
</tr>
<tr>
<td>84</td>
<td>31</td>
</tr>
<tr>
<td>84</td>
<td>31</td>
</tr>
</tbody>
</table>
A typical RNA-seq experiment