Introduction to Homework 2

Kaur Alasoo
26 February 2020
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.
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
- Denaturation
- Load flow cell

Slide from Ira Hall
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

1. **Genomic DNA**
2. Shear to desired length (~400bp)
3. **DNA fragments**
4. Ligate adapters, size select
5. **Sequencing library**
6. **Illumina GA2**
7. Clusters on a flow-cell
8. **Millions to billions of paired-end reads (readpairs)**

---

**DNA Fragments**

```
5'  GGTGTACGAATAGTTTCCTTTTACACTCCTTGACCATCCTAGC
--------------------------------\/---------------------------------
5'  GGACTGAAACTTCATCTGTCTTTATAGATATGCGTGCAGC
```

---

**Illumina GA2**

---

**Clusters on a flow-cell**

---

**Millions to billions of paired-end reads (readpairs)**
Aligning pairs of reads to the reference genome

1) a readpair

5' GGTGTCAGATAGTTTTCCCTTACACTCCTTGACCCTCCTAGC 5'  
-----------------------------------  
GGACTGAACCTTCATCTGTCTTTATAGATATGC

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

---

Slide from Aaron Quinlan
Formats
(Mostly) all technologies yield DNA sequences in FASTQ format

@seq1
ACCTTCGAACGGCGGGGGTTACAA+
"*"(((**))%%%++).1***
@seq2
TGGAACCGAACGGCCCCGTTACAT+
"*"((((+++))+++++)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  GATTTGGGGTTCAAAGCAGTATCGATCAAAATAGTAAATCCATTGTTCAACTCACAGTTT
Quality scores  +
                    !'('*(((**++)%%++)%(%%).1***-+*'))*55CCF>>>>>CCCCCCCC65

http://en.wikipedia.org/wiki/FASTQ_format
A typical RNA-seq experiment

RNA-Seq: a revolutionary tool for transcriptomics
The Missing Semester of Your CS Education

Classes teach you all about advanced topics within CS, from operating systems to machine learning, but there's one critical subject that's rarely covered, and is instead left to students to figure out on their own: proficiency with their tools. We'll teach you how to master the command-line, use a powerful text editor, use fancy features of version control systems, and much more!

Students spend hundreds of hours using these tools over the course of their education (and thousands over their career), so it makes sense to make the experience as fluid and frictionless as possible. Mastering these tools not only enables you to spend less time on figuring out how to bend your tools to your will, but it also lets you solve problems that would previously seem impossibly complex.

https://missing.csail.mit.edu/