An Introduction to RNA-Seq Data and Computation

Jason Knight
March 19th, 2014
Overview

● Biology Overview
● Sequencing Technology
● Tutorial
  − QC
  − Mapping
  − Quantification
● Online Resources

NB: Very brief, but very basic for diverse audiences.
Biology Overview
Biology 101 – Cells

• Human:
  - $10^{13}$ cells\(^1\)
  - Each with 46 chromosomes (23 pairs)
  - Genome is ~3 billion nucleotides long

\(^1\) http://www.ncbi.nlm.nih.gov/pubmed/23829164
DNA well suited for:

- Reliable Storage of information
- Replication into DNA and RNA
- Having specific binding sites for proteins
Biology 101 – The Central Dogma

Simple right?

Not so fast...
Biology 201 – Alternative Splicing

But we're not done yet...
Biology 201 – mRNA Structure

Many other types of RNA too:
- siRNAs
- miRNAs
- IncRNAs
- tRNAs
- rRNAs
- Etc... but those won't concern us too greatly today
Humans have ~30k genes
- Genes overlap, run in different directions and are gapped
- ~300k mRNA molecules per cell
Sequencing Technology
What is sequencing?

• The ability to 'read' the nucleotides of DNA/RNA molecules
• This turns out to be phenomenally powerful/flexible:
  – DNA Sequencing
    • Human genome project / 1000 genomes project
    • Epigenetics: ChIP-Seq
  – RNA Sequencing
    • “Snapshot” of the action inside the cell
Some motivation...
Different Techniques

Developments in High Throughput Sequencing

- SOLID
- Proton
- MiSeq
- Hiseq 2000
- Hiseq 2500 RR
- GA II
- PGM
- GS FLX
- GS Junior
- PacBio RS
- Sanger

Gigabases per run (log scale)

Read length (log scale)
Illumina

- 2/3 of sequencing market share
- Fluorescent chemistry gives large numbers of short (50-250 bp) reads
Illumina – More Details

- Illumina HiSeq
- Single sequencing run = 1 “flow cell”
- 1 flow cell = 8 “lanes”
- 1 lane = ~180M reads
Barcoding

- Barcode = index = 4-8 mer 'fingerprint'
- A form of multiplexing
- More samples in a lane
- Separate samples bioinformatically
The Tutorial
(AKA: Nuts and Bolts)
Pipeline

Wet lab sample prep

Computational

RNA-ase Inhibitors
Lyse Cells
Centrifugation
Column Filtration
Add primers and adapters
PCR
QC
Raw Reads
Fastq
QC
Processing
Aligning/Mapping
BAM/SAM
QC
Normalization
Differential Expression
Isoform Analysis
Quantification
Wet Lab Pipeline

RNA-ase Inhibitors → Biological Sample

Lyse Cells → Centrifugation

Column Filtration → Add primers and adapters

PCR → QC
Computational Pipeline

Raw Reads
- ACTGAGTGATACGTGATG
- ACACTGATACGTGATG
- ATAACTGATACGTGATG
- ACTGATACGTGATG
- CCTGACTGAGTGA TACTG

Fastq
- QC

Processing
- Aligning/Mapping
  - BAM/SAM
  - QC

Normalization

Differential Expression

Isoform Analysis

Quantification
Raw Data

- From sequencing machine
- Fastq file format
- ASCII Text representing reads
- 4 lines per read
- Contains quality scores (Phred = -log_{10}(Probability of incorrect base))

DEMO
Raw Data QC

- Usually provided by sequencing center
- FastQC package
- And Spreadsheets with number of reads and other metrics
- Check for:
  - Lane/Multiplex Balance
  - Quality vs. nucleotide position
  - Diversity

DEMO
Processing

- Combining multiplexed groups of reads spread across lanes
- Trimming (dependent on QC)
- Entropy filtering
Aligning/Mapping

- Sometimes done by sequencing center
- Software (DNA)
  - Bowtie, BWA, SNAP
- Software (RNA-Seq)
  - TopHat, STAR
- Requires a choice of reference genomes and annotations

Raw Reads

ACTGAGTGA TACGTGA TG
ACGTGA TACGTGA TG
ATAAGTGA TACGTGA TG
ACTGTGA TACGTGA TG
CCTGACTGAGTGA TACTG

Fastq

BAM/SAM

QC

Processing

Aligning/Mapping

QC

Normalization

Differential Expression

Isoform Analysis

Quantification

DEMO
Mapping QC

- Now consider balance between samples
- Visually inspect alignments
  - 3' or 5' biases
- Diversity of aligned reads
- Consider alternative organism mappings
- ERCC Calibration

**DEMO**
Mapping QC

![Graph showing the relationship between observed counts and true concentration before dilution (attomoles/μL).]

- Raw Reads
- Fastq
- QC
- Processing
- Aligning/Mapping
- BAM/SAM
- QC
- Normalization
- Differential Expression
- Isoform Analysis
- Quantification
Analyses

- What are you looking for?
- What do you want to show?
- What comparisons do you want to make?
Normalization

- Sometimes normalization is built into the analysis/software
- Most popular: RPKM/FPKM (reads/fragments per kilobase of gene length per megabase of mapped reads)
- Dillies et al (2012)¹: RPKM is inferior to simpler methods such as upper quartile

¹ “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis”
Quantification

- How do we want to compare?
- Downstream Analyses?
- Many software packages:
  - Scripture, Cufflinks, MMSEQ, Alexa-Seq, MISO, IsoLasso, CEM, cuffdiff 2, RSEM, FDM, IsoformEx, NSMAP, IQSEQ, DEXSeq, DSGSeq, IsoEM, Isolnfer, iReckon, BitSeq, eXpress, SLIDE, DiffSplice, SplicingCompass, HTSeq-count ....

- Notably RSEM is used by TCGA
- Although Cufflinks is “tried and true”

DEMO
Quantification

http://arxiv.org/abs/1304.5952

* novel discovery
B favorite
/ didn’t test to completion
Oh, look at the time...

• Many, many things not covered here
  – Mapping and Quantification are each additional presentations
  – Visualization (e.g. CummeRbund)
  – Data management (checksums, RAID≠backup)
  – Software best practices (Github, version control)
  – Linux command line magic
  – Gene ID Conversion
  – Etc...

• So where to find more help?
Online Resources

• Q&A Community
  – Biostars.org
  – Seqanswers.com

• Tools (see previous bullet)

• Repositories
  – Sequence/Annotations: UCSC, RefSeq
  – Experimental datasets: NCBI SRA, GEO
## Online Resources

### NCBI/EBI/UCSC Summary Table

<table>
<thead>
<tr>
<th></th>
<th>NCBI</th>
<th>EBI</th>
<th>UCSC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tools</strong></td>
<td>BLAST</td>
<td>BLAST</td>
<td>BLAT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exonerate</td>
<td>In-Silico PCR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ClustalW2</td>
<td></td>
</tr>
<tr>
<td><strong>Data Repository</strong></td>
<td>GenBank</td>
<td>ArrayExpress</td>
<td>ENCODE</td>
</tr>
<tr>
<td></td>
<td>GEO</td>
<td>ENA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SRA</td>
<td>PDBe</td>
<td></td>
</tr>
<tr>
<td><strong>DNA/Genome</strong></td>
<td>Genome</td>
<td>Ensembl</td>
<td>Ideogram</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Recombination Rate</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GC Content</td>
</tr>
</tbody>
</table>
### Online Resources

#### NCBI/EBI/UCSC Summary Table (Cont’d)

<table>
<thead>
<tr>
<th></th>
<th>NCBI</th>
<th>EBI</th>
<th>UCSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA/Gene</td>
<td>Gene</td>
<td>Ensembl</td>
<td>UCSC Genes GENCODE RefSeq Genes</td>
</tr>
<tr>
<td>RNA</td>
<td>RefSeq</td>
<td>Ensembl</td>
<td>mRNAs ESTs UniGene</td>
</tr>
<tr>
<td>Proteins</td>
<td>Protein, RefSeq Conserved Domain</td>
<td>UniProt InterPro PRIDE</td>
<td></td>
</tr>
</tbody>
</table>
Summary – Words of Advice

- Fast changing field
- Use internet resources/recent literature to find your way
- Don't believe everything you read
- Never delete data (unless you need to)
- Write clean code (you're going to need it again)
- Organize! Otherwise all is lost
Contact Information

• Jason Knight
  - Jknight@tamu.edu
  - Github.com/chapkinlab
Thanks!