RNA-Seq with DNA Subway
Part I

Jason Williams
Cold Spring Harbor Laboratory, DNA Learning Center
williams@cshl.edu
@JasonWilliamsNY
DNALC Live

This is an experiment, give us feedback on what you would like to see!
DNALC Live

• Provide genetics, molecular biology, and bioinformatics learning resources

• Laboratory and computer demos, short online courses for middle school, high school, and the general public

• Interviews with scientists, help for teachers

• At-home activities, social media contests, and more
DNALC Website and Social Media

dnalc.cshl.edu

dnalc.cshl.edu/dnalc-live
DNALC Website and Social Media

youtube.com/DNALearningCenter

facebook.com/cshldnalc

@dnalc

@dna_learning_center
RNA-Seq with DNA Subway
Part I
Who is this course for?

• Audience(s):
  • Undergraduate biology 200 level and up
  • (advanced AP Bio/graduate)

• Format: 3 sessions (1 per week); ~ 45 minutes each

• Exercises: Follow along with our online bioinformatics tool DNA Subway

• Learning resources: Slides and resource sheets available
Course Learning Goals

• Understand the rationale of an RNA-Seq experiment and its design

• Understand how we obtain DNA sequence and access its quality

• Use *DNA Subway* (*FastQC/FastX*) to QC sequence data

• Use *DNA Subway* (*Kallisto*) to (pseudo)align reads

• Use *DNA Subway* (*Sleuth*) to explore RNA-Seq results
Lab Setup

- We will be using DNA Subway – You can get a free account at cyverse.org (required)
RNA-Seq with *DNA Subway*

Part I

(background and sequence quality)
Steps for today’s session

• Introduction to RNA-Seq

• Learn about our example data set

• Learn about high-throughput sequencing and data sources

• Examine DNA sequence quality and QC
Introduction to RNA-Seq
What is RNA-Seq? - measuring the transcriptome

• To understand what genes are active, and under what circumstances, we must know what genes are being transcribed into messenger RNA
What is RNA-Seq? - measuring the transcriptome

• To understand what genes are active, and under what circumstances, we must know what genes are being transcribed into messenger RNA

• A cell in the liver has the same DNA instructions as a neuron in the brain. However the genes being expressed differ greatly between these cells
What is RNA-Seq? - measuring the transcriptome

- RNA-Seq allows us to measure the transcriptome – take an account of all transcription occurring in a cell/tissue
What is RNA-Seq? - measuring the transcriptome

• RNA-Seq allows us to measure the transcriptome – take an account of all transcription occurring in a cell/tissue

• We use the abundance of an RNA transcript as a proxy for the activity of some cellular process (e.g. protein synthesis, regulatory activity)
What is RNA-Seq? - measuring the transcriptome

• RNA-Seq allows us to measure the transcriptome – take an account of all transcription occurring in a cell/tissue

• We use the abundance of an RNA transcript as a proxy for the activity of some cellular process (e.g. protein synthesis, regulatory activity)

• We analyze these data to compare samples (e.g. cancerous vs. non-cancerous)
What is RNA-Seq?
What is RNA-Seq?
What is RNA-Seq?
What is RNA-Seq?
What is RNA-Seq?
What is RNA-Seq?
What can expression tell you?

• CYP1A/1B – Cytochrome p450 family, involved in drug metabolism including processing toxins

Photo Credit:
Effects of Tobacco Smoke on Gene Expression and Cellular Pathways in a Cellular Model of Oral Leukoplakia
Zeynep H. Gümüş, Baoheng Du, Ashutosh Kacker, Jay O. Boyle, Jennifer M. Bocker, Piali Mukherjee, Kotha Subbaramaiah, Andrew J. Dannenberg and Harel Weinstein
Cancer Prev Res July 1 2008 (1) (2) 100-111; DOI: 10.1158/1940-6207.CAPR-08-0007
What can expression tell you

Highly expressed in lung
Key Concept: Variation vs. Difference
In an RNA-Seq experiment we need to find “true” differences between samples (while subtracting trivial differences)
Spot the difference

Photo Credit: https://www.rd.com/culture/spot-10-differences-picture/
Spot the difference

Found it? Great! Now let's try something increadable.
Find the missing star

If you can find the missing star, you're awsome.

Photo Credit:
https://9gag.com/gag/3934072
Spot the difference – biological variation

Which of these is chocolate cake?

Photo Credit:
https://sallysbakingaddiction.com/triple-chocolate-layer-cake/
https://www.chefsteps.com/activities/ultimate-chocolate-cake
https://www.lifeloveandsugar.com/raspberry-chocolate-layer-cake/
Spot the difference – biological variation

Photo Credit:
Spot the difference – experimental variation
Spot the difference – replication

Photo credit: https://hbctraining.github.io/intro-to-rnaseq-hpc-salmon/lessons/experimental_planning_considerations.html
Introduction to our data set
RNA-Seq of hNPC – Zika Virus

Zika Virus
RNA-Seq of hNPC – Zika Virus

Zika infection of neural progenitor cells perturbs transcription in neurodevelopmental pathways
Lynn Yi, Harold Pimentel, Lior Pachter
Published: April 27, 2017 • https://doi.org/10.1371/journal.pone.0175744 • See the preprint

Zika virus (ZIKV) is associated with congenital microcephaly and adult Guillain-Barre Syndrome

Photo credit:

Photo credit:
RNA-Seq of hNPC – Zika Virus

- Zika virus (ZIKV) infects human embryonic cortical neural progenitor cells (hNPCs)
- ZIKV-infected hNPCs produce infectious ZIKV particles
- ZIKV infection leads to increased cell death of hNPCs
- ZIKV infection dysregulates cell cycle and transcription in hNPCs
<table>
<thead>
<tr>
<th>Sample</th>
<th>Accession Number</th>
<th>Condition</th>
<th>Seq method</th>
<th>Seq machine</th>
<th>Reads</th>
<th>No. Fragments / weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mock1-1</td>
<td>SRR3191542</td>
<td>mock</td>
<td>paired-end</td>
<td>MiSeq</td>
<td>15855554</td>
<td>7927777</td>
</tr>
<tr>
<td>Mock2-1</td>
<td>SRR3191543</td>
<td>mock</td>
<td>paired-end</td>
<td>MiSeq</td>
<td>14752152</td>
<td>7391076</td>
</tr>
<tr>
<td>ZIKV1-1</td>
<td>SRR3191544</td>
<td>zika</td>
<td>paired-end</td>
<td>MiSeq</td>
<td>14723054</td>
<td>7361527</td>
</tr>
<tr>
<td>ZIKV2-1</td>
<td>SRR3191545</td>
<td>zika</td>
<td>paired-end</td>
<td>MiSeq</td>
<td>15242694</td>
<td>7621347</td>
</tr>
<tr>
<td>Mock1-2</td>
<td>SRR3194428</td>
<td>mock</td>
<td>single-end</td>
<td>NextSeq</td>
<td>72983243</td>
<td>72983243</td>
</tr>
<tr>
<td>Mock2-2</td>
<td>SRR3194429</td>
<td>mock</td>
<td>single-end</td>
<td>NextSeq</td>
<td>94729809</td>
<td>94729809</td>
</tr>
<tr>
<td>ZIKV1-2</td>
<td>SRR3194430</td>
<td>zika</td>
<td>single-end</td>
<td>NextSeq</td>
<td>71055823</td>
<td>71055823</td>
</tr>
<tr>
<td>ZIKV-2-2</td>
<td>SRR3194431</td>
<td>zika</td>
<td>single-end</td>
<td>NextSeq</td>
<td>66528035</td>
<td>66528035</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pone.0175744.t001
Sequence data from NCBI
Generation of sequence data
Preparation of the library – cDNA synthesis
Preparation of the library – cDNA synthesis

Photo credit:
https://media.springernature.com/full/springer-static/image/art%3A10.1038%2Fnmeth.f.280/MediaObjects/41592_2009_Article_BFnmethf280_Fig1_HTML.jpg
Preparation of the library – cDNA synthesis

Photo credit:
https://media.springernature.com/full/springer-static/image/art%3A10.1038%2Fnmeth.f.280/MediaObjects/41592_2009_Article_BFnmethf280_Fig1_HTML.jpg
Illumina Sequencing

A. Library Preparation

Genomic DNA → Fragmentation → Adapters → Ligation → Sequencing Library

NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

Photo credit: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf
Illumina Sequencing

B. Cluster Amplification

Flow Cell

Bridge Amplification Cycles

Clusters

Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

Photo credit:
C. Sequencing

Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated “n” times to create a read length of “n” bases.
Illumina Sequencing

D. Alignment and Data Analysis

Reads

ATGCCATTGCAATTGTACAT
TGGCATTGCAATTGG
AGATGGTTATTG
GATGCCATTGCAA
GCATTGCAATTGTAC
ATGCCATTGCAATT
AGATGGCATTGCAATTGG

Reference Genome

AGATGGTTATTGCAATTGTACAT

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.
Lab: RNA-Seq of neural progenitor stem cells infected with Zika virus
Lab: Creating a DNA Subway Project
Working on DNA Subway Green Line

DNA Subway ties together key bioinformatics tools and databases to assemble gene models, investigate genomes, work with phylogenetic trees and analyze DNA barcodes. Roll over the "stations" on the subway map to find out more about the analysis steps. Analyze your own data or sample data provided. To start a project, select one of the "lines" (red, yellow, blue, green, purple). Register and login to be able to save and share your results.
Working on DNA Subway Green Line

[Diagram showing processes related to data management and tools like Manage data and FastX Toolkit]
Key Concept: Sequence Quality
Examining quality with FastQC
### Phred scores...

<table>
<thead>
<tr>
<th>Phred Score</th>
<th>Error (bases miscalled)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1,000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10,000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100,000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>
FastQ Format

@SEQ_ID
GATTTGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCAGTTT+

!' * (((**++)%%%++) (%%%%) .1***-++' ) ) **55CCF>>>>C65

• Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description.
• Line 2 is the raw sequence letters.
• Line 3 begins with a '+' character and is *optionally* followed by the same.
• Line 4 encodes the quality values for the sequence in Line 2.

Photo and text credit:
https://en.wikipedia.org/wiki/FASTQ_format
FastQ Format

Photo and text credit:
https://en.wikipedia.org/wiki/FASTQ_format
FastQ Format

Photo and text credit:
https://en.wikipedia.org/wiki/FASTQ_format
Examining quality with FastQC

- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality
Examining quality with FastQC

Fail: most frequently observed mean quality is below 20 (1% error rate)
Next time:
Sequence alignment to reference
DNALC Website and Social Media

dnalc.cshl.edu

dnalc.cshl.edu/dnalc-live