Week 14, Lecture 27

István Albert

Biochemistry and Molecular Biology and Bioinformatics Consulting Center

Penn State
Quantifying with sequencing

1. Sample consists of varying amounts of DNA/RNA representing different genomic locations

2. Observed abundance → sample abundance → biological significance
Complications

• Typically a population of cells is being sequenced

• Each cell may be in different states

• What we observe is the average state over all cells

• It is possible to sample a single cell → but is it representative?
Quantifying technologies

- Chip-Seq – isolates certain locations in the genome

- Metagenomics – samples multiple genomes of different abundances

- RNA-Seq – samples transcripts that are present in different abundances
Question: X-Seq: Can You Provide A Listing Of The Various Applications For Sequencing Technologies?

Could we compile a growing list of all the different sequencing applications? This could be accompanied by a short description for tyros.

RNA-Seq involves sequencing of transcriptomes consisting of different species of RNA (e.g. mRNA) as cDNA. It provides a good picture of transcript abundance as well as transcript-processing events. Wikipedia page

ChiP-Seq combines chromatin immunoprecipitation (ChiP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. Wikipedia page

ChiP-Exo is a modification of the ChiP-seq protocol, improving the resolution of binding sites from hundreds of base pairs to less than one base pair. It employs the use of exonucleases to degrade strands of the protein-bound DNA in the 5'→3' direction to within a small number of nucleotides of the protein binding site. Wikipedia page

miRNA-Seq Wikipedia page

DNA-Seq SEQWiki page

Methyl-Seq Wikipedia page

RAD-Seq University of Edinburgh page

GRO-Seq is the name for Global Run-On sequencing. See more details here: ngsbioinformatics.com page

NET-Seq is short for Native Elongating Transcript sequencing. It involves immunoprecipitation of RNA polymerase followed by sequencing bound RNA fragments Nature paper

FAIRE-Seq is the name for Formaldehyde-Assisted Isolation of Regulatory Elements following by sequencing. Check out a description here: Wikipedia page

Ribo-Seq consists of freezing ribosomes to capture bound mRNA in the process of translation. After the ribosomes are primed open the mRNA fragments are sequenced. Science page

Update: Here's a much better list: Lior Pachter's list of *Seq
Effective lengths and coverage

Boundary conditions introduce artifacts – coverage is a lagging indicator.
RNA-Seq concepts

There are three major, quite independent steps

1. **Alignment** → produces a BAM file

2. **Quantitation (abundance estimation)** starts with a BAM file produces tabular files

3. **Differential Expression Analysis (statistical inference)** → statistically significant changes between samples

4. **Result Annotation** → what is the biological significance of results
RNA-Seq – other considerations

• Annotation information lacking:
  – Missing altogether → transcriptome assembly
  – Incomplete → new transcripts based on known exons

• Transcript level versus exon level analysis
RNA-Seq interpretation strategies

For each step one can mix and match the various components. Can be overwhelming.

1. The results will be different depending on what method you choose

2. Major/strong variations stay the same (usually)

3. Smaller changes may only show up with some tools/techniques.

4. The size of a variation does not correlate with biological significance!

5. How do we know what is an artifact and what is not? ...
What makes RNA-Seq difficult?

1. Going in without a hypothesis to check

2. Expecting RNA-Seq to provide both the question and the correct answer to a unexplained biological phenomena
Asking the right questions

Wrong way to pose the question:

“I need to explain a biological phenomena. I am going to run RNA-Seq experiments over many time points and conditions to identify which genes change.”

Right way to pose the question:

“I need to explain a biological phenomena. I formulated one or more hypotheses and I am going to design RNA-Seq experiments to validate these.”
RNA-seq alignment

1. Align against a known transcriptome:
   - **good**: efficient, well defined answers
   - **bad**: unable to discover novel transcripts, may align reads that would map better in noncoding regions

2. Align against genome:
   - **good**: discover novel transcripts
   - **bad**: more false positives, more uncertainty

Some (most) methods try to make use of a combination of both
Align against the known transcriptome

Simplest approach:

• Produce transcripts sequences and treat them as the “reference”

• Use an aligner to align against this reference

• Post-process the results
Align RNA-seq reads against genome

Typical methodology implemented internally by tools

1. Separate reads (read-pairs) that map “correctly” to exonic locations

2. Re-align reads that did not map in step 1 to “potential” junction sites
   - create a putative transcriptome by fusing sequences at the border of mapped reads
   - identify intron splicing indictor base pairs: GT --- AG, etc.
   - train machine learning algorithms to predict junction sites
Download and install **tophat** and **cufflinks**

- The so-called Tuxedo suite is a consistently good performer, tools that work immediately with minimal fuss and therefore
  - **bowtie** → short read mapper
  - **tophat** → RNA-seq mapping
  - **cufflinks** → isoform assembly and quantification
    - **cuffdiff** → establish differential expression
    - **Cuffcompare** → compare assembled transcripts against reference
    - **cuffmerge** → merge experiments
  - **cummeRbund** → R package to facilitate RNA-Seq analysis
RNA-seq data analysis bootcamp

This workshop is directed toward life scientists with little to no experience with statistical computing or bioinformatics. This interactive workshop will introduce both the Linux/UNIX operating system and the R statistical computing environment, with a focus on a biological application - analyzing RNA-seq data for differentially expressed genes. The morning session will introduce basic operation in a UNIX environment, and will cover the first steps in an RNA-seq analysis including QC, alignment, and quantitation. The afternoon will introduce the R statistical computing environment, and will cover differential gene expression analysis using Bioconductor. By the end of the workshop, participants will:

1. Be familiar with the UNIX shell, including navigating the filesystem, creating/examining/removing files, getting help, and batch operations.
2. Know how to align and quantitate gene expression with RNA-seq data.
3. Become familiar with the R statistical computing environment, including data types, variables, array manipulation, functions, data frames, data import/export, visualization, and using packages.
4. Know what packages to use and what steps to take to analyze RNA-seq data for differentially expressed genes.

Participants will also be exposed to operating in a virtual environment and/or provisioning their own cloud computing resources. This course is sponsored by the Claude Moore Health Sciences Library, and borrows some materials from the Software Carpentry and Data Carpentry projects.

Date: Monday, November 10, 2014  
Time: 8:00 am (sharp!) - 5:00 pm  
Location: Carter classroom, first floor Health Sciences Library

By Stephen Turner see http://bioconnector.github.io/workshops/
Code repository

- Follow the instructions in the code repository
- Install the tools
- Obtain and prepare the data
Bwa-mem alignment: gene RPS3A

Reads align well even though it is not a splice aware aligner – soft clips ends but does not follow to the other ise
A spliced alignment in IGV
Homework 27

• Select one dataset from the lecture downloads

• Use tophat to align the reads.

• Show the commands that you ran and a screenshot of the alignments.

• How many of the exons are part of junctions? (hint: intersect exons with junctions)