RNA-Seq Tutorial 1

Kevin Silverstein, Ying Zhang
Research Informatics Solutions, MSI
October 18, 2016

Slides available at www.msi.umn.edu/tutorial-materials
RNA-Seq Tutorials

https://www.msi.umn.edu/events

- Lectures
  - RNA-Seq experiment design and analysis
  - PacBio Sequencing (Iso-Seq): **Nov. 29**

- Hands-on tutorials
  - QC of data with Galaxy
    - Tuesdays: Oct. 11, **Nov. 15**
  - QC of data at the command line
    - Thursdays: Oct. 13, **Nov. 17**
  - Analyzing human RNA-Seq data with Galaxy
    - Tuesday: **Oct. 25**
  - Analysis of PacBio sequencing Data using SMRT Portal
    - Dec. 1
Outline

• Part I – Kevin Silverstein
  – Introduction
  – Experimental Design
  – Sequencing Design

• Part II – Ying Zhang
  – Data Quality Control
  – Read Mapping
  – Differential Gene Expression
Part I

• Introduction
• Experimental Design
• Sequencing Design
Introduction

• RNA-Seq
  – High-throughput sequencing of RNA
  – Differential expression analysis (Illumina)
    • Quantitative measurement of transcript expression
  – Transcriptome assembly (PacBio)
    • Qualitative identification of expressed sequence
Sample 1

mRNA isolation

Fragmentation
RNA -> cDNA

Sequence fragment end(s)

Map reads

<table>
<thead>
<tr>
<th></th>
<th>Gene A</th>
<th>Gene B</th>
<th>Total mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (bp)</td>
<td>1000</td>
<td>2000</td>
<td></td>
</tr>
</tbody>
</table>

1. **Count**: # of Reads (mapped to the gene)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Gene A</th>
<th>Gene B</th>
<th>Total mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>4</td>
<td>4</td>
<td>1.1 x 10^6</td>
</tr>
<tr>
<td>Sample 2</td>
<td>4</td>
<td>8</td>
<td>0.9 x 10^6</td>
</tr>
</tbody>
</table>

2. **CPM**: # of Reads Per Million (mapped reads)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Gene A</th>
<th>Gene B</th>
<th>Total mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>3.6</td>
<td>3.6</td>
<td>1.1 x 10^6</td>
</tr>
<tr>
<td>Sample 2</td>
<td>4.4</td>
<td>8.9</td>
<td>0.9 x 10^6</td>
</tr>
</tbody>
</table>

3. **TPM**:\ # of Reads per length of all expressed Transcripts (in kb) Per Million*

<table>
<thead>
<tr>
<th>Sample</th>
<th>Gene A</th>
<th>Gene B</th>
<th>Total mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>6.7</td>
<td>3.3</td>
<td>1.1 x 10^6</td>
</tr>
<tr>
<td>Sample 2</td>
<td>5</td>
<td>5</td>
<td>0.9 x 10^6</td>
</tr>
</tbody>
</table>

4. **RPKM / FPKM**: # of Reads / Fragments Per Kb of exon of one gene per Million

<table>
<thead>
<tr>
<th>Sample</th>
<th>Gene A</th>
<th>Gene B</th>
<th>Total mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>3.6</td>
<td>1.8</td>
<td>1.1 X 10^6</td>
</tr>
<tr>
<td>Sample 2</td>
<td>4.4</td>
<td>4.4</td>
<td>0.9 X 10^6</td>
</tr>
</tbody>
</table>

*Scaled by 10^6 for readability

RNA-Seq: reliable and reproducible for gene expression analysis

RNA-Seq Workflow

- **Design Experiment**: Set up the experiment to address your questions.
- **RNA Preparation**: Isolate and purify input RNA.
- **Prepare Libraries**: Convert the RNA to cDNA; add sequencing adapters.
- **Sequence**: Sequence cDNAs using a sequencing platform.
- **Analysis**: Analyze the resulting short-read sequences.

http://rnaseq.uoregon.edu/
Experimental Design

- Qualitative (Annotation) vs Quantitative (Differential expression)
- Biological comparison(s)
- Replicates
- Pooling Samples
- Strand-specific sequencing
- single-cell Sequencing
Experimental design: Goal

• What are my goals?
  – Expression changes in response to disturbance?
  – Rare or novel transcripts?
  – Regulatory RNAs: Short or non-coding transcripts

• What are the characteristics of my system?
  – Large, complex genome?
  – Introns and high degree of alternative splicing?
  – Gene-dense and overlapping? (strand-specific)
  – Hybrid cell populations? (single-cell)
Experimental design

• Simple designs (Pairwise comparisons)

- Two group
  Drug effect
  Control
  Experimental (drug applied)

- Two group
  Drug effect
  Control
  Experimental (drug applied)

• Complex designs — Consult a statistician

- Two factor
  Cancer type X drug
  Cancer sub-type 1
  With drug
  Cancer sub-type 1
  With drug

- Two factor
  Cancer type X drug
  Cancer sub-type 2
  With drug
  Cancer sub-type 2
  With drug

- Matched-pair
  Normal
  Cancer
Experimental design

• Technical replicates
  – Not needed: low technical variation
    • Minimize batch effects
    • Randomize sample order

• Biological replicates
  – Essential for differential expression analysis

• Pooling samples
  – Limited RNA obtainable
    • Multiple pools per group required
  – Transcriptome assembly
Experimental Design

- Why go Strand-Specific library?
  - Gene-dense genomes (bacteria, archaea, lower eukaryotes)
  - Antisense transcription (higher eukaryotes)

- When to consider single-cell technology?
  - Heterogeneous cell types
  - Cell-specific signature expression
Experimental design

Cost = library prep ($129/sample) + sequencing

<table>
<thead>
<tr>
<th>UMGC HiSeq 2500 High-output Rates</th>
<th>10 million reads (1/25 lane)</th>
<th>20 million reads (1/12 lane)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-read (1x50 cycles)</td>
<td>$170</td>
<td>$214</td>
</tr>
<tr>
<td>$1,025/lane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paired-end read (2x50 cycles)</td>
<td>$194</td>
<td>$265</td>
</tr>
<tr>
<td>$1,642/lane</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

-Valid Spring, 2016

10 million reads per sample, 50bp single-end reads
- Small genomes with no alternative splicing

20 million reads per sample, 50bp paired-end reads
- Mammalian genomes (large transcriptome, alternative splicing, gene duplication)
Experimental design

Calculating Sample Size Estimates for RNA Sequencing Data:
Steven Hart, Terry Therneau, Yuji Zhang, Gregory Poland, Jean-Pierre Kocher.
Journal of Computational Biology 2013, 10(12): 970-978

RnaSeqSampleSize
https://cqs.mc.vanderbilt.edu/shiny/RnaSeqSampleSize/

RNA-seq: technical variability and sampling
Lauren M McIntyre, Kenneth K Lopiano, Alisson M Morse, Victor Amin, Ann L Oberg, Linda J Young and Sergey V Nuzhdin
BMC Genomics 2011, 12:293

Statistical Design and Analysis of RNA Sequencing Data
Paul L. Auer and R. W. Doerge

Calculating Sample Size Estimates for RNA Sequencing Data

subSeq: Determining Appropriate Sequencing Depth Through Efficient Read Subsampling.
David G. Robinson and John D. Storey, Bioinformatics first published online September 3, 2014

Focus on RNA sequencing quality control (SEQC)
Nature Biotechnology 32, vii (2014)
Sequencing

- Platforms
- Library preparation
  - Paired-end vs single end reads
  - Read length
  - Sequencing depth
  - Remove Ribosomal RNA
Sequencing

- **Illumina sequencing by synthesis**
  - HiSeq2500 (high output or rapid run)
  - MiSeq
    - low throughput, longer reads (2x300), fast turnaround

- **PacBio SMRT single-molecule sequencing**
  - Very long reads, low throughput, high cost
  - Great option for transcriptome assembly

Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study

Sequencing

- Library preparation (Illumina TruSeq protocol for HiSeq/MiSeq)
  - RNA isolation
  - Poly-A purification
  - Fragmentation
  - cDNA synthesis using random primers
  - Adapter ligation
  - Size selection
  - PCR amplification (~16 cycles)
Sequencing

• Common Library types:
  – Polyadenylated RNA > 200bp (standard method)
  – Small RNA
  – Strand-specific
  – Stranded with rRNA reduction

• Other considerations
  – Single End vs Paired End
  – Low input
  – Total RNA
  – Targeted capture
  – Ribosomal Reduction
  – Degraded RNA
Sequencing

Optimal library fragment size depends on goals and organism: **exon size**

- **Adjacent connectivity**
  - Insert size = exon size

- **Minimal connectivity**
  - Insert size << exon size

- **Long-range connectivity**
  - Insert size >> exon size

One size doesn’t fit all: organisms can differ in exon size distribution
## Genome size characteristics (iGenomes)

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of genes</th>
<th>Transcriptome size (Mbp)</th>
<th>Mode Avg</th>
<th>Intron size range (1%</th>
<th>99%)</th>
<th>% genome repetitive</th>
<th>% genes in families*</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>29230</td>
<td>70.1</td>
<td>100</td>
<td>300</td>
<td>77</td>
<td>107000</td>
<td>47</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>24080</td>
<td>61.4</td>
<td>100</td>
<td>300</td>
<td>78</td>
<td>100000</td>
<td>44</td>
</tr>
<tr>
<td><em>Gallus gallus</em>*</td>
<td>4906</td>
<td>11.1</td>
<td>100</td>
<td>230</td>
<td>73</td>
<td>120000</td>
<td>10</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>18436</td>
<td>30.1</td>
<td>150</td>
<td>450</td>
<td>30</td>
<td>25000</td>
<td>32</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>23933</td>
<td>28.0</td>
<td>110</td>
<td>220</td>
<td>43</td>
<td>8000</td>
<td>4</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>27278</td>
<td>51.1</td>
<td>70</td>
<td>300</td>
<td>46</td>
<td>4900</td>
<td>9</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>6692</td>
<td>8.9</td>
<td>75</td>
<td>1200</td>
<td>20</td>
<td>2600</td>
<td>1</td>
</tr>
<tr>
<td><em>Escherichia coli</em>**</td>
<td>4290</td>
<td>0.6</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>3</td>
<td>52</td>
</tr>
</tbody>
</table>

* % genes with at least one paralog in the COG database (unicellular) or included in the COG lineage specific expansion (LSE) list. (These percentages are likely systematic underestimates)

** Poor annotation is suspected for iGenomes UCSC-based Gallus gallus (galGal3)

*** http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/E/Esch.coli.html; ecocyc; Gur-Arie, Genome Res 2000;.
Break!
Part II

- Data Quality Control
- Read Mapping
- Differential Gene Expression
Data Quality Control

- Quality assessment
- Trimming and filtering
Data Quality Assessment

• Evaluate read library quality
  – Identify poor/bad samples
  – Sequencing errors
  – Contamination

• Software
  – FastQC (recommended)
    • Command-line, Java GUI, or Galaxy
  – RSeQC
  – UMGC standards
    • Default to UMGC customers

RSeQC assessment of RNA degradation
Data Quality Assessment

• Recommendations
  – Generate quality plots for all read libraries
  – Trim and/or filter all data if needed
    • Always trim and filter for de novo transcriptome assembly
  – Regenerate quality plots after trimming and filtering to determine effectiveness

• Software:
  – Trimmomatic
  – Cutadapt
  – Galaxy, many options including cutadapt (NGS: QC and manipulation)
  – Many others: http://omictools.com/whole-genome-resequencing-category
UMGC QC report

• Gopher-pipelines:
  – https://bitbucket.org/jgarbe/gopher-pipelines/wiki/Home

• Sample Document:
Read Mapping

- Pipeline
- Software
- Input
- Output
Mapping Reads

- Millions of short reads
  - Splice-aware aligner
  - Reference Genome
    - Reads aligned to genome
      - SAM/BAM
  - Reference Transcriptome
    - Reads aligned to transcriptome
      - SAM/BAM

- Abundance estimation and Differential expression analysis

© Research Informatics Solutions (RIS)
Minnesota Supercomputing Institute
Aligner

- Tophat
  - One of the earliest splice-aware aligner
- HiSAT2
  - Fast and memory-efficient
- STAR
  - Ultra fast, but memory-intensive
- Kallisto
  - Not a mapper, just a quantification program
  - Super fast, and robust to errors (highly accurate)
Mapping

• Input
  – Fastq files
  – Index of genome/transcriptome
  – Annotation file (optional for some, but required for others)

• Output
  – SAM (text) / BAM (binary) alignment files
    • SAMtools – SAM/BAM file manipulation
    • Picard-tools – SAM/BAM file manipulation
  – Summary statistics (per read library)
    • % reads with unique alignment
    • % reads with multiple alignments
    • % reads with no alignment
    • % reads properly paired (for paired-end libraries)
Differential Expression

- Abundance estimation
- Differential expression analysis
Abundance Estimation

- Exon 1
- Exon 2
- Exon 3

Splice form 1

Splice form 2

Definitely splice form 1

Definitely splice form 2

Ambiguous
Abundance Estimation

- Raw gene-level read counts
  - easy, powerful, inaccurate for some genes
- FPKM transcript
  - Corrects for gene length and library size
  - Directly comparable between different genes within the same dataset
  - Uses ambiguously mapped reads
- Transcripts per million (TPM)
  - Normalizes to transcript copies instead of reads
  - Corrects for cases where the average transcript length differs between samples
  - Uses ambiguously mapped reads
Abundance Estimation

• Raw read counts
  – Htseq
  – Subread FeatureCounts

• Transcript abundance estimation
  – Cufflinks (cuffquant, cuffnorm)
  – RSEM
  – Salmon - includes alignment; requires reference transcriptome
  – Kallisto - includes alignment; requires reference transcriptome

• Other
  – EDGE-pro - prokaryotic alignment and abundance estimation
Differential Expression

• Cuffdiff (Cufflinks package)
  – Pairwise comparisons
  – Differential gene, transcript, and primary transcript expression; differential splicing and promoter use
  – Easy to use, well documented
  – Input: transcriptome, SAM/BAM read alignments (abundance estimation built-in)

• EdgeR
  – Complex experimental designs using generalized linear model
  – Information sharing among genes (Bayesian gene-wise dispersion estimation)
  – Difficult to use R package
  – Input: raw gene/transcript read counts (calculate abundance with separate software not cufflinks)

• Ballgown
  – Differential gene, transcript, and exon expression
  – Complex experimental designs: multi-group, timecourse, continuous covariates; adjust for confounding factors
  – R package, works with Cufflinks
Differential Expression

• Others
  – DESeq - R package, generates nice plots
  – DEXseq – R package
  – Slueth
    • Works with Kallisto and Salmon
    • Transcript-level differential expression
    • Incorporates expression estimate uncertainty
Further Reading

RNA-seqlopedia
http://rnaseq.uoregon.edu/

Bioinformatics for High Throughput Sequencing
Rodríguez-Ezpeleta, Naiara.; Hackenberg, Michael.; Aransay, Ana M.;
SpringerLink New York, NY : Springer c2012

RNA sequencing: advances, challenges and opportunities
Fatih Ozsolak1 & Patrice M. Milos1
Nature Reviews Genetics 12, 87-98 (February 2011)

Computational methods for transcriptome annotation and quantification using RNA-seq
Manuel Garber, Manfred G Grabherr, Mitchell Guttman & Cole Trapnell
Nature Methods 8, 469–477 (2011)

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks
Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David Kelley, Harold Pimentel, Steven Salzberg, John L Rinn & Lior Pachter

SEQanswers.com
biostar.stackexchange.com
omicstools.com - bioinformatics software directory
Analyze your data

• Easy
  – UMII: free analysis of simple experiments
  – RIS: fee-based analysis of complex experiments
    • UMII Updraft funding available http://www.research.umn.edu/umii/funding/index.html

• Basic bioinformatics skills
  – Galaxy
    • Tutorials

• Advanced bioinformatic skills
  – Command-line analysis pipelines: http://bmgc-docs.readthedocs.org
Acknowledgements

• The lecture slides are adapted from previous RIS tutorial of RNA-Seq created by John Garbe (UMGC).
Questions / Discussion

Free consultations with bioinformatics experts at RIS:
Email request to help@msi.umn.edu