RNA-Seq Analysis & Interpretation

Kathleen Fisch, Ph.D.
Executive Director, Center for Computational Biology & Bioinformatics, University of California, San Diego, La Jolla, CA, USA

Email: Kfisch@ucsd.edu
Website: compbio.ucsd.edu
Outline

• RNA-Seq Background
  • Overview
  • Rationale & analysis goals
  • Library prep
  • Experimental design

• RNA-Seq Analysis
  • Overview
  • QC
  • Alignment
  • Gene & Transcript quantification
  • Normalization
  • Differential expression

• Downstream Analysis & Interpretation
  • Hypergeometric test & Overrepresentation analysis
  • Functional enrichment analysis
  • Pathway analysis
  • Visualization with IGV
  • Network Analysis
UC San Diego
Center for Computational Biology & Bioinformatics

Epigenomics

Whole Genome Analysis

Gene Expression & Regulation

Networks & Systems Biology

Biomarkers & Therapeutic Targets

Compbio.ucsd.edu
RNA-Seq Overview

Samples of interest
- Condition 1 (normal colon)
- Condition 2 (colon tumor)

Isolate RNAs

Generate cDNA, fragment, size select, add linkers

Sequence ends

Map to genome, transcriptome, and predicted exon junctions

Downstream analysis

100s of millions of paired reads
10s of billions bases of sequence

Goldsby 2015 NESCent Academy

Compbio.ucsd.edu
RNA sequencing Rationale

Griffith & Griffith 2013

Griffith & Griffith 2013

Short read is split by intron when aligning to reference Genome

Compbio.ucsd.edu
RNA-Seq Analysis Goals

- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- IncRNA

- Allele specific expression
- Mutation discovery
- Fusion detection
- RNA editing
- miRNASeq
- Single cell
RNA-Seq Experimental Design

1. RNA extraction protocol
   • Poly(A) selection vs deplete rRNA
2. Stranded protocols
3. Single-end (SE) vs paired-end (PE) reads
4. Sequencing Depth aka library size
5. Number of replicates

Martin et al. 2014. Scientific Reports
RNA-Seq Experimental Design: Types of Replicates

- **Technical Replicate**
  - Multiple instances of sequence generation
    - Flow Cells, Lanes, Indexes

- **Biological Replicate**
  - Multiple isolations of cells showing the same phenotype, stage or other experimental condition
  - Some example concerns/challenges:
    - Environmental Factors, Growth Conditions, Time

Griffith & Griffith 2013
RNA-Seq Experimental Design: Number of Replicates

**Table 1** Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

<table>
<thead>
<tr>
<th>Effect size (fold change)</th>
<th>Replicates per group</th>
<th>3</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25</td>
<td>17 %</td>
<td>25 %</td>
<td>44 %</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>43 %</td>
<td>64 %</td>
<td>91 %</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>87 %</td>
<td>98 %</td>
<td>100 %</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequencing depth (millions of reads)</th>
<th>3</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>19 %</td>
<td>29 %</td>
<td>52 %</td>
</tr>
<tr>
<td>10</td>
<td>33 %</td>
<td>51 %</td>
<td>80 %</td>
</tr>
<tr>
<td>15</td>
<td>38 %</td>
<td>57 %</td>
<td>85 %</td>
</tr>
</tbody>
</table>

Example of calculations for the probability of detecting differential expression in a single test at a significance level of 5 %, for a two-group comparison using a Negative Binomial model, as computed by the RNASeqPower package of Hart et al. [190]. For a fixed within-group variance (package default value), the statistical power increases with the difference between the two groups (effect size), the sequencing depth, and the number of replicates per group. This table shows the statistical power for a gene with 70 aligned reads, which was the median coverage for a protein-coding gene for one whole-blood RNA-seq sample with 30 million aligned reads from the GTEx Project [214].
RNA-Seq Analysis Overview

Conesa et al. 2016 Genome Biology
RNA-Seq Analysis Overview

Lin 2015 NGS Workshop

1. Hypothesis
2. Experimental Design
   - RNA Purification
     - Single-end/Paired-end
     - Un-stranded/Stranded
   - Library Preparation
   - Sequencing

3. Raw Read 1, Raw Read 2
   - QC of Raw Reads
     - FASTQC
   - Read Alignment
     - Annotation-based Genome-guide assembly
     - De novo assembly
   - Quantification
     - Gene-level
     - Transcript-level
   - Differential Expression Testing
     - Biological Insight
     - Visualization
RNA-Seq Analysis QC

- Quality-control checkpoints
  - Raw reads
  - Read alignment
  - Quantification
  - Reproducibility
RNA-Seq Analysis QC

• Quality-control checkpoints
  • Raw reads
    • Sequence quality, GC content, presence of adaptors, overrepresented k-mers, and duplicated reads
  • Tools: FASTQC
  • Goals
    • Detect sequencing errors
    • PCR artifacts or contaminations
RNA-Seq Analysis QC

• Quality-control checkpoints
  • Read alignment
    • % Mapped reads
    • Uniformity of read coverage on exons and mapped strand
    • Ideal: 70-90% mapped to human genome
  • Tools: Picard, RSeQC, Qualimap
  • Goals:
    • Global indicator of overall sequencing accuracy and presence of contaminating DNA
    • Non-uniformity may indicate low RNA quality in starting material

RSeQC

Uniform

Low RNA Quality 3’ Bias
RNA-Seq Analysis QC

• Quality-control checkpoints
  • **Quantification**
    • Check GC content and gene length bias
    • Tools: Bioconductor packages – NOISEq or EDA-Seq
  • Goal:
    • Apply correcting normalization methods, if necessary

*Li et al. 2014. Nature Biotechnology*
RNA-Seq Analysis QC

- Quality-control checkpoints
  - **Reproducibility**
    - Checking on reproducibility among replicates and for possible batch effects
    - Tool: Principal component analysis (PCA)/Multidimensional Scaling
  - Goal: Assess global quality of RNA-seq dataset
RNA-Seq Analysis -- Alignment
RNA-Seq Analysis
Alignment – Transcriptome Mapping

(b) Transcriptome mapping

Reads

Ungapped mapper → Bowtie → Mapping to transcriptome

RSEM, Kallisto → Transcript identification & counting

Conesa et al. 2016 Genome Biology
RNA-Seq Analysis
Alignment – Genome Mapping

Conesa et al. 2016 Genome Biology
RNA-Seq Analysis
Alignment -- Important Parameters

• Strandedness of the RNA-seq library
• # of mismatches to accept
• Read length
• Type of reads (SE or PE)
• Fragment length
RNA-Seq Analysis – Transcript Discovery

Reference-based assembly

De novo assembly

Align reads to the genome

Reference-based assembly of aligned reads

Unaligned reads

De novo assembly

RNA-seq reads

Assemble-then-align

Scaffold contigs

Unassembled reads

Extend contigs

Comprehensive assembly

RNA-Seq Analysis
Gene-level Quantification

- Aggregation of raw counts of mapped reads
  - HTSeq-count or featureCounts
  - Gene-level approach based on GTF gene coordinates
  - Discard multimappers

<table>
<thead>
<tr>
<th></th>
<th>union</th>
<th>intersection_strict</th>
<th>intersection_nonempty</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>gene_A</td>
<td>no_feature</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>gene_A</td>
<td>no_feature</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>gene_A</td>
<td>ambiguous</td>
<td>ambiguous</td>
<td>ambiguous</td>
</tr>
<tr>
<td>gene_A</td>
<td>ambiguous</td>
<td>ambiguous</td>
<td>ambiguous</td>
</tr>
<tr>
<td>gene_A</td>
<td>ambiguous</td>
<td>ambiguous</td>
<td>ambiguous</td>
</tr>
</tbody>
</table>
RNA-Seq Analysis

Transcript-level Quantification

• Transcript-level expression algorithms
  • Allocate multi-mapping reads among transcript and output within-sample normalized values corrected for sequencing biases.
  • RSEM (RNA-Seq by Expectation Maximization)
  • Cufflinks, eXpress, Kallisto
RNA-Seq Analysis Normalization

Dillies et al. 2013
Briefings in Bioinformatics
RNA-Seq Analysis Normalization

Summary of comparison results for the seven normalization methods under consideration

<table>
<thead>
<tr>
<th>Method</th>
<th>Distribution</th>
<th>Intra-Variance</th>
<th>Housekeeping</th>
<th>Clustering</th>
<th>False-positive rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>UQ</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>–</td>
</tr>
<tr>
<td>Med</td>
<td>++</td>
<td>++</td>
<td>–</td>
<td>++</td>
<td>–</td>
</tr>
<tr>
<td>DESeq</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>TMM</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>Q</td>
<td>++</td>
<td>–</td>
<td>+</td>
<td>++</td>
<td>–</td>
</tr>
<tr>
<td>RPKM</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

A ‘–’ indicates that the method provided unsatisfactory results for the given criterion, while a ‘+’ and ‘++’ indicate satisfactory and very satisfactory results for the given criterion.

Dillies et al. 2013 Briefings in Bioinformatics
TMM – Trimmed Mean of M values

Attempts to correct for differences in RNA *composition* between samples

E.g. if certain genes are very highly expressed in one tissue but not another, there will be less “sequencing real estate” left for the less expressed genes in that tissue and RPKM normalization (or similar) will give biased expression values for them compared to the other sample.

Equal sequencing depth -> orange and red will get lower RPKM in RNA population 1 although the expression levels are actually the same in populations 1 and 2.

RNA-Seq – Differential Expression Analysis Overview

http://sourceforge.net/projects/trinityrnaseq/files/misc/RNASEQ_WORKSHOP/r
Figure 4 Power to detect true differential expression. Bars show the total number of genes that are detected as statistically significant (FDR < 0.1) (a) with equal library sizes and (b) with unequal library sizes. The blue segments show the number of true positives while the red segments show false positives. 200 genes are genuinely differentially expressed. Results are averaged over 100 simulations. Height of the blue bars shows empirical power. The ratio of the red to blue segments shows empirical FDR. FDR, false discovery rate.
RNA-Seq – Differential Expression Analysis – Bioconductor RNAseq123

Data pre-processing
- Transformations from the raw-scale
- Removing genes that are lowly expressed
- Normalising gene expression distributions
- Unsupervised clustering of samples

Differential expression analysis
- Creating a design matrix and contrasts
- Removing heteroscedascity from count data
- Fitting linear models for comparisons of interest
- Examining the number of DE genes
- Examining individual DE genes from top to bottom
- Useful graphical representations of differential expression results

https://www.bioconductor.org/help/workflows/RNAseq123/#normalising-gene-expression-distributions
RNA-Seq – Differential Expression Analysis: Data Pre-processing

1. Transform raw counts into counts per million (CPM) or log2-counts per million (log-CPM)
2. Remove genes that are lowly expressed (CPM > 1)
RNA-Seq – Differential Expression Analysis: Data Pre-processing

3. Normalize gene expression distributions (TMM)

4. Unsupervised clustering of samples

Example data: Boxplots of log-CPM values showing expression distributions for unnormalised data (A) and normalised data (B) for each sample in the modified dataset where the counts in samples 1 and 2 have been scaled to 5% and 500% of their original values respectively.
RNA-Seq – Differential Expression Analysis

1. Create design matrix and contrasts

```r
# Create design matrix
design <- model.matrix(~0+group+lane)
colnames(design) <- gsub("group", ",", colnames(design))
design

# Design matrix
## Basal  LP  ML  laneL006  laneL008
## 1  0  1  0  0  0
## 2  0  0  1  0  0
## 3  1  0  0  0  0
## 4  1  0  0  1  0
## 5  0  0  1  1  0
## 6  0  1  0  1  0
## 7  1  0  0  1  0
## 8  0  0  1  0  1
## 9  0  1  0  0  1
```

```r
# Make contrasts
contr.matrix <- makeContrasts(
    BasalvsLP = Basal-LP,
    BasalvsML = Basal-ML,
    LPvsML = LP-ML,
    levels = colnames(design))
contr.matrix

# Contrasts
## Levels  BasalvsLP  BasalvsML  LPvsML
## Basal     1       1        0
## LP        -1       0        1
## ML        0       -1       -1
## laneL006  0       0        0
## laneL008  0       0        0
```
RNA-Seq – Differential Expression Analysis

2. Remove heteroscedasticity from count data
RNA-Seq – Differential Expression Analysis

3. Fitting linear models for comparisons of interest – limma

4. Examining the number of DE genes

```
summary(decideTests(efit))
```

<table>
<thead>
<tr>
<th></th>
<th>BasalvsLP</th>
<th>BasalvsML</th>
<th>LPvsML</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>4127</td>
<td>4338</td>
<td>2895</td>
</tr>
<tr>
<td>0</td>
<td>5740</td>
<td>5655</td>
<td>8825</td>
</tr>
<tr>
<td>1</td>
<td>4298</td>
<td>4172</td>
<td>2445</td>
</tr>
</tbody>
</table>

![Venn diagram showing overlap of DE genes across conditions]
RNA-Seq – Differential Expression Analysis

5. Examining individual DE genes from top to bottom

```r
treat <- topTreat(tfit, coef=1, n=Inf)
treat2 <- topTreat(tfit, coef=2, n=Inf)
head(treat)

##   ENTREZID SYMBOL TXCHROM logFC AveExpr   t  P.Value adj.P.Val
## 1     12759  C1u            chr14  -5.44   8.86 33.4  3.99e-10  2.7e-06
## 2     53624  C1dn7           chr11  -5.51   6.30 32.9  4.50e-10  2.7e-06
## 3    242505  Rasef           chr4   -5.92   5.12 31.8  6.06e-10  2.7e-06
## 4    67451  Pk2              chr16  -5.72   4.42 30.7  8.01e-10  2.7e-06
## 5    228543  Rhov             chr2   -6.25   5.49 29.5  1.11e-09  2.7e-06
## 6    70350  Basp1           chr15  -6.07   5.25 28.6  1.38e-09  2.7e-06

head(treat2)

##   ENTREZID SYMBOL TXCHROM logFC AveExpr   t  P.Value adj.P.Val
## 1     242505  Rasef           chr4  -6.51   5.12 35.5  2.57e-10  1.92e-06
## 2     53624  C1dn7           chr11  -5.47   6.30 32.5  4.98e-10  1.92e-06
## 3     12521  Cd82            chr2  -4.67   7.07 31.8  5.80e-10  1.92e-06
## 4     71740  Nectin4         chr1  -5.56   5.17 31.3  6.76e-10  1.92e-06
## 5     20661  Sort1            chr3  -4.91   6.71 31.2  6.76e-10  1.92e-06
## 6     15375  Foxa1            chr12   -5.75   5.63 28.3  1.49e-09  2.28e-06
```
RNA-Seq – Differential Expression Analysis

6. Useful graphical representations of differential expression
RNA-Seq – Differential Expression Analysis

6. Useful graphical representations of differential expression
Downstream Analysis & Interpretation

• Hypergeometric test and overrepresentation analysis
• Functional Gene Set Enrichment Analysis
• Pathway Analysis
• Visualize Alignments with IGV
• Network Analysis
Hypergeometric test

- Uses hypergeometric distribution to measure the probability of having drawn a **specific number of successes** (out of a total number of draws) from a population

- Example:

  Imagine that there are 4 green and 16 red marbles in a box.

  You close your eyes and draw 5 marbles **without replacement**

  **What is the probability that exactly 2 of the 5 are green?**
Hypergeometric Test for Overrepresentation Analysis

- **Development (1842 genes)**
  - Hoxa5
  - Hoxa11
  - Ltbp3
  - Sox4
  - Foxc1
  - Edn1
  - Ror2
  - Gnag
  - Smad3
  - Wdr5
  - Trp63
  - Sox9
  - Pax1
  - Acd
  - Rai1
  - Pitx1
  - ......

- **Differentially expressed genes (581 genes)**
  - Sash1
  - Cd24a
  - Agt
  - Psrc1
  - Ctl42a
  - Angptl4
  - Depdc7
  - Sorbs1
  - Macrod1
  - Enpp2
  - Tmem176a
  - ......

- **Compare**
  - Observe: 581 genes
  - Expect: 1842 genes
  - 98 genes

- **Is the observed overlap significantly larger than the expected value?**
Downstream Analysis & Interpretation: Functional Enrichment

• Gene list enrichment analysis (Hypergeometric test) based on functional annotations

• Tools
  • ToppGene, GSEA, Webgestalt, DAVID
Downstream Analysis & Interpretation: Pathway Analysis

- **Databases**
  - Ex. KEGG, WikiPathways, Reactome, PathwayCommons, BioCarta

- **Tools**
  - Ex. Webgestalt, Signaling Pathway Impact Analysis, ToppGene, WikiPathways

Werner 2008 Current Opinion Biotechnology
Downstream Analysis & Interpretation: Pathway Analysis

Tool: Bioconductor Pathview
Downstream Analysis & Interpretation: Visualization with IGV and/or GenePattern

Wilkening et al. 2013 Nucleic Acids Research
Network Analysis

• Databases
  • PPI
  • Physical interactions
  • Indirect associations
  • Coexpression
  • Literature
  • Experimental

• Tools
  • Cytoscape, StringDB, GeneMania, NetworkX

Farina et al. 2012 Skeletal Muscle
Recommended Reading

A survey of best practices for RNA-seq data analysis

Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szczęśniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang and Ali Mortazavi


Review Article | Published: 24 July 2020

mRNAs, proteins and the emerging principles of
gene expression control

Christopher Buccitelli & Matthias Selbach

*Nature Reviews Genetics* 21, 630–644(2020) | Cite this article