

# RNA Sequencing Examples

C. Ryan Campbell

Duke University

*c.ryan.campbell@duke.edu*

3 Oct 2017

# Overview

## 1 Goals

## 2 Experimental Design

- Number of Replicates

## 3 Data Analysis

- Handling Reads
- Transcripts
- Differential Expression
- Functional Profiling

# Today's Goals

- What are RNAseq best practices?

# Today's Goals

- What are RNAseq best practices?
- How did the assigned papers address these recommendations?

# Today's Goals

- What are RNAseq best practices?
- How did the assigned papers address these recommendations?
- Jigsaw Activity

# Review to Read

- This is a summary from:

# Review to Read

- This is a summary from:
- Conesa et al. 2016. A survey of best practices for RNA-seq data analysis.

# Review to Read

- This is a summary from:
- Conesa et al. 2016. A survey of best practices for RNA-seq data analysis.
- Good overview of current “best practices”



# Jigsaw Activity

- Each of your papers took a different approach to this problem

# Jigsaw Activity

- Each of your papers took a different approach to this problem
- As I'm covering these practices make a note of what your paper did

# Jigsaw Activity

- Each of your papers took a different approach to this problem
- As I'm covering these practices make a note of what your paper did
- Compare and contrast during the activity

# Experimental Design

- Enrichment method

# Experimental Design

- Enrichment method
  - Deplete rRNA

# Experimental Design

- Enrichment method
  - Deplete rRNA
  - Enrich mRNA via polyA selection

# Experimental Design

- Enrichment method
  - Deplete rRNA
  - Enrich mRNA via polyA selection
- Library Type (single v paired-end)

# Experimental Design

- Enrichment method
  - Deplete rRNA
  - Enrich mRNA via polyA selection
- Library Type (single v paired-end)
- Sequencing Depth



# Experimental Design

- Enrichment method
  - Deplete rRNA
  - Enrich mRNA via polyA selection
- Library Type (single v paired-end)
- Sequencing Depth
- Number of Replicates

# Enrichment

- Deplete rRNA

# Enrichment

- Deplete rRNA
- Enrich mRNA via polyA selection

# Enrichment

- Deplete rRNA
  - Requires good quality RNA
- Enrich mRNA via polyA selection

# Enrichment

- Deplete rRNA
  - Requires good quality RNA
  - High RIN (RNA Integrity Number)
- Enrich mRNA via polyA selection

# Enrichment

- Deplete rRNA
  - Requires good quality RNA
  - High RIN (RNA Integrity Number)
  - Often not possible with tissue
- Enrich mRNA via polyA selection

# Enrichment

- Deplete rRNA
  - Requires good quality RNA
  - High RIN (RNA Integrity Number)
  - Often not possible with tissue
- Enrich mRNA via polyA selection
  - Have to use with bacterial samples (lack polyA)

# Library Type

- Single End



# Library Type

- Single End
- Paired End

# Library Type

- Single End
  - Better for well-annotated organisms
- Paired End

# Library Type

- Single End
  - Better for well-annotated organisms
  - Cost can go towards more reads, better coverage
- Paired End

# Library Type

- Single End
  - Better for well-annotated organisms
  - Cost can go towards more reads, better coverage
- Paired End
  - More crucial for *de novo*

# Library Type

- Single End
  - Better for well-annotated organisms
  - Cost can go towards more reads, better coverage
- Paired End
  - More crucial for *de novo*
  - Improve mappability to “dicey” genomes

# Sequencing Depth

- Deeper is better

# Sequencing Depth

- Deeper is better
- Lower end ranges from:

# Sequencing Depth

- Deeper is better
- Lower end ranges from:
  - 5mil reads (for common mRNA)



# Sequencing Depth

- Deeper is better
- Lower end ranges from:
  - 5mil reads (for common mRNA)
  - 100mil reads (for rare mRNA)

# Sequencing Depth

- Deeper is better
- Lower end ranges from:
  - 5mil reads (for common mRNA)
  - 100mil reads (for rare mRNA)
- What else would effect necessary depth?

# Sequencing Depth

- Deeper is better
- Lower end ranges from:
  - 5mil reads (for common mRNA)
  - 100mil reads (for rare mRNA)
- What else would effect necessary depth?

# Sequencing Depth

- Deeper is better
- Lower end ranges from:
  - 5mil reads (for common mRNA)
  - 100mil reads (for rare mRNA)
- What else would effect necessary depth?
- Use a “Saturation Curve” to assess results at given depth

# Number of Replicates

- Depends on:

# Number of Replicates

- Depends on:
  - 1 Technical Variability

# Number of Replicates

- Depends on:
  - ① Technical Variability
    - Technical replicates should result in an R-squared  $> 0.9$

# Number of Replicates

- Depends on:
  - ① Technical Variability
    - Technical replicates should result in an R-squared  $> 0.9$
  - ② Biological Variability



# Number of Replicates

- Depends on:
  - 1 Technical Variability
    - Technical replicates should result in an R-squared  $> 0.9$
  - 2 Biological Variability
  - 3 Desired Statistical Power

# Number of Replicates

- Depends on:
  - 1 Technical Variability
    - Technical replicates should result in an R-squared  $> 0.9$
  - 2 Biological Variability
  - 3 Desired Statistical Power
- Minimum of 3 replicates per group

# Number of Replicates

- Depends on:
  - 1 Technical Variability
    - Technical replicates should result in an R-squared  $> 0.9$
  - 2 Biological Variability
  - 3 Desired Statistical Power
- Minimum of 3 replicates per group
- Conduct a power analysis

# Number of Replicates

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

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

# Handling Reads

- Clean reads with FASTX-Toolkit or Trimmomatic

# Handling Reads

- Clean reads with FASTX-Toolkit or Trimmomatic
- Expect 70-90% of reads aligning (model organism)

# Handling Reads

- Clean reads with FASTX-Toolkit or Trimmomatic
- Expect 70-90% of reads aligning (model organism)
- Reads accumulating at the 3' end of transcripts could be a sign of poor RNA quality (polyA enrichment only)

# Transcripts

- If a genome/annotation exists you can map to it



# Transcripts

- If a genome/annotation exists you can map to it
- but, you can only quantify expression

# Transcripts

- If a genome/annotation exists you can map to it
- but, you can only quantify expression
- Discovery of new transcripts must be done separately

# Transcript Discovery

- Need high coverage to discover new transcripts

# Transcript Discovery

- Need high coverage to discover new transcripts
- Paired end data helps, hard to get complete transcript (see: IsoSeq)

# Transcript Discovery

- Need high coverage to discover new transcripts
- Paired end data helps, hard to get complete transcript (see: IsoSeq)
- Several software to tackle this question:

# Transcript Discovery

- Need high coverage to discover new transcripts
- Paired end data helps, hard to get complete transcript (see: IsoSeq)
- Several software to tackle this question:
- Cufflinks, iReckon, SLIDE, StringTie

# Transcript Discovery

- Need high coverage to discover new transcripts
- Paired end data helps, hard to get complete transcript (see: IsoSeq)
- Several software to tackle this question:
- Cufflinks, iReckon, SLIDE, StringTie
- Montebello - quantification AND isoforms! (project idea?)

# Transcript Quantification

- Count transcripts to measure expression



# Transcript Quantification

- Count transcripts to measure expression
- Raw counts of mapped reads are converted...

# Transcript Quantification

- Count transcripts to measure expression
- Raw counts of mapped reads are converted...

# Transcript Quantification

- Count transcripts to measure expression
- Raw counts of mapped reads are converted...
- Not necessary when comparing within the same gene across samples

# Transcript Quantification

- Count transcripts to measure expression
- Raw counts of mapped reads are converted...
- Not necessary when comparing within the same gene across samples
- Is necessary for correctly ranking gene expression levels

# Differential Expression

- Find which genes are expressed at different levels

# Differential Expression

- Find which genes are expressed at different levels
- Possible software to use:

# Differential Expression

- Find which genes are expressed at different levels
- Possible software to use:
- Cufflinks, TMM, DESeq, PoissonSeq, UpperQuartile

# Differential Expression

- Find which genes are expressed at different levels
- Possible software to use:
- Cufflinks, TMM, DESeq, PoissonSeq, UpperQuartile
- COMBAT and ARSyN can be used to correct batch effects



# Differential Expression

- Find which genes are expressed at different levels
- Possible software to use:
- Cufflinks, TMM, DESeq, PoissonSeq, UpperQuartile
- COMBAT and ARSyN can be used to correct batch effects
- Not necessarily memory intensive software (DESeq runs in R)

# Functional Profiling

- Genes are nice, but what do they do?

# Functional Profiling

- Genes are nice, but what do they do?
- Characterize the function of the differentially expressed genes

# Functional Profiling

- Genes are nice, but what do they do?
- Characterize the function of the differentially expressed genes
- Also compare functional pathways for over or under-expression

# Functional Profiling

- Genes are nice, but what do they do?
- Characterize the function of the differentially expressed genes
- Also compare functional pathways for over or under-expression
- Tools:

# Functional Profiling

- Genes are nice, but what do they do?
- Characterize the function of the differentially expressed genes
- Also compare functional pathways for over or under-expression
- Tools:
- GOseq, Blast2GO, Gene Set Variation Analysis, SeqGSEA

# Jigsaw Activity

- 1 Form a group with everyone else who read the same paper

# Jigsaw Activity

- 1 Form a group with everyone else who read the same paper
- 2 Discuss your paper by answering the questions on the next slide – **10-15 mins**



# Jigsaw Activity

- 1 Form a group with everyone else who read the same paper
- 2 Discuss your paper by answering the questions on the next slide – **10-15 mins**
- 3 Get together with a new group, with a single representative from each paper covered

# Jigsaw Activity

- 1 Form a group with everyone else who read the same paper
- 2 Discuss your paper by answering the questions on the next slide – **10-15 mins**
- 3 Get together with a new group, with a single representative from each paper covered
- 4 Share your results and discuss the differences – **15-20 mins**

# Jigsaw Activity

- **Mouse Olf** - Sisi, Kevin, Alan
  - **Lemurs** - Alvin, Rahul, Helena
  - **AD** - Austin, Othmane, Jenn
  - **Obese Chickens** - Hank, Nayib, Jake
- 1 Number (and type) of replicates
  - 2 Software used
  - 3 What functional analysis was done?
  - 4 What conclusions were drawn?

# The End