

# Everything you wanted to know about RNAseq

C. Ryan Campbell

Duke University

*c.ryan.campbell@duke.edu*

12 Sept 2017

# Overview

1

## RNAseq

- Goals
- Introduction
- Central Dogma
- Measuring RNA

2

## Project Groups

# Today's Goals

- How does RNAseq work?
- What are the benefits?
- What are the challenges?
  
- Meet your group!

# Dotstorming

- What questions did you have?
- Dotstorming Board
- Write your question
- Vote on others
- (Take 5 mins)

# A Parable

- Steve and Susan wanted to know how water affects plant growth



# A Parable

- Steve and Susan wanted to know how water affects plant growth
- They decide to grow the same vine in 6 different pots by the window in their classroom



# A Parable

- Each day they give the pots 2 cups of water



# A Parable

- Each day they give the pots 2 cups of water
- At the end of 3 months they measure the length of the vine





# A Parable

- Each day they give the pots 2 cups of water
- At the end of 3 months they measure the length of the vine
- What is wrong with this “experiment”?



# Central Dogma

- DNA codes for

# Central Dogma

- DNA codes for
- RNA codes for

# Central Dogma

- DNA codes for
- RNA codes for
- Protien

- Static, non-changing

# DNA

- Static, non-changing
- You're stuck with the genome you have

# DNA

- Static, non-changing
- You're stuck with the genome you have
- Good for investigating large scale and inflexible impacts of environment

- Static, non-changing
- You're stuck with the genome you have
- Good for investigating large scale and inflexible impacts of environment
- A written record of selection over time
  - (or population size changes)
  - (or species histories)



- Dynamic, ever-changing

- Dynamic, ever-changing
- Transcribed “as needed” from DNA

- Dynamic, ever-changing
- Transcribed “as needed” from DNA
- Only exons, introns are spliced out

# Protein

- Final product of central dogma

# Protein

- Final product of central dogma
- “Action item” from the list

# Protein

- Final product of central dogma
- “Action item” from the list
- Performs a task within the cell

- Sequencing the RNA molecules in a cell

- Sequencing the RNA molecules in a cell
  - Use as a proxy for identifying active proteins



- Sequencing the RNA molecules in a cell
  - Use as a proxy for identifying active proteins
- Why not DNA?

- Sequencing the RNA molecules in a cell
  - Use as a proxy for identifying active proteins
- Why not DNA?
  - Static, doesn't change based on cell function or status

- Sequencing the RNA molecules in a cell
  - Use as a proxy for identifying active proteins
- Why not DNA?
  - Static, doesn't change based on cell function or status
  - Doesn't tell us where the genes are (annotation)

- Sequencing the RNA molecules in a cell
  - Use as a proxy for identifying active proteins
- Why not DNA?
  - Static, doesn't change based on cell function or status
  - Doesn't tell us where the genes are (annotation)
- Why not Proteins?

- Sequencing the RNA molecules in a cell
  - Use as a proxy for identifying active proteins
- Why not DNA?
  - Static, doesn't change based on cell function or status
  - Doesn't tell us where the genes are (annotation)
- Why not Proteins?
  - Unstable

- Sequencing the RNA molecules in a cell
  - Use as a proxy for identifying active proteins
- Why not DNA?
  - Static, doesn't change based on cell function or status
  - Doesn't tell us where the genes are (annotation)
- Why not Proteins?
  - Unstable
  - Difficult to collect and quantify

# Measuring RNA

## 1 Generate cDNA

# Measuring RNA

- 1 Generate cDNA
- 2 Prepare a sequencing library



# Measuring RNA

- 1 Generate cDNA
- 2 Prepare a sequencing library
- 3 (Next-Gen) Sequence!!!

# Generate cDNA

- Isolate RNA from cells and use deoxyribonuclease (DNase) to remove accidental DNA



# Generate cDNA

- Isolate RNA from cells and use deoxyribonuclease (DNase) to remove accidental DNA
- Select for poly-A tails to enrich for mature RNA (often with magnetic beads)



# Generate cDNA

- Isolate RNA from cells and use deoxyribonuclease (DNase) to remove accidental DNA
- Select for poly-A tails to enrich for mature RNA (often with magnetic beads)
- Reverse transcribe the RNA into cDNA (lasts longer)



# Generate cDNA

- Isolate RNA from cells and use deoxyribonuclease (DNase) to remove accidental DNA
- Select for poly-A tails to enrich for mature RNA (often with magnetic beads)
- Reverse transcribe the RNA into cDNA (lasts longer)
  - Lose strand info



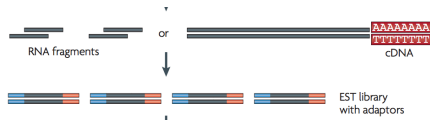
# Generate cDNA

- Isolate RNA from cells and use deoxyribonuclease (DNase) to remove accidental DNA
- Select for poly-A tails to enrich for mature RNA (often with magnetic beads)
- Reverse transcribe the RNA into cDNA (lasts longer)
  - Lose strand info
  - Can be kept with labeling



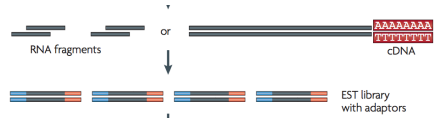
# prepare a sequencing library

- Take the now-cDNA into a standard library prep



# prepare a sequencing library

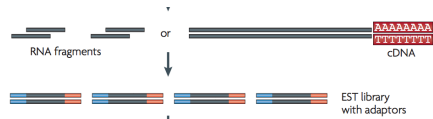
- Take the now-cDNA into a standard library prep
- Add adapters





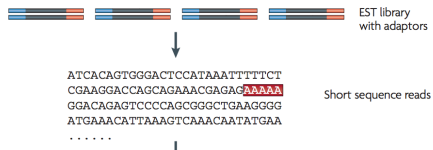
# prepare a sequencing library

- Take the now-cDNA into a standard library prep
- Add adapters
- Amplify (with PCR) and sequence



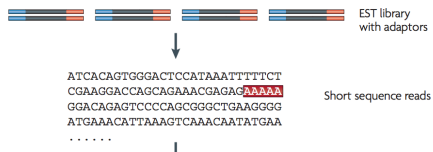
# NGS!!!

- Single-end or paired-end



# NGS!!!

- Single-end or paired-end
  - Because genes are often shorter than the full length of paired-end, data is analyzed one end at a time



# NGS!!!

- Single-end or paired-end
  - Because genes are often shorter than the full length of paired-end, data is analyzed one end at a time
- Generate billions of reads



# NGS!!!

- Single-end or paired-end
  - Because genes are often shorter than the full length of paired-end, data is analyzed one end at a time
- Generate billions of reads
- “fastq” format



# fastq Format

- 1 Sequence header - information regarding the machine

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNTAGTTTC
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcffffcfeeffcfffffdddf`feed]`}_Ba_^__[YBBBBBBBBBRT\}][
```

# fastq Format

- 1 Sequence header - information regarding the machine
- 2 Sequence content - bases

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNTAGTTTC
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcffffcfeeffcfffffdddf`feed]`}_Ba_^__[YBBBBBBBBBRT\]]
```

# fastq Format

- 1 Sequence header - information regarding the machine
- 2 Sequence content - bases
- 3 Sequence header - repeat of above (+ instead of @)

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTNNNNNNNNNTAGTTTC
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcffffffcfeeffcffffffdddf`feed]`}_Ba_^__[YBBBBBBBBBRT\]]
```



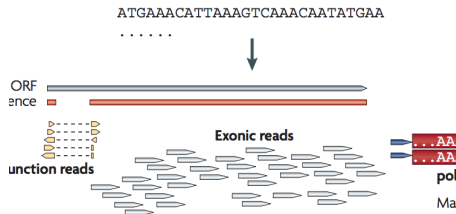
# fastq Format

- 1 Sequence header - information regarding the machine
- 2 Sequence content - bases
- 3 Sequence header - repeat of above (+ instead of @)
- 4 Sequence quality - ASCII coded

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTNNNNNNNNNTAGTTTC
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcffffffcfeffffffffffddfd`feed]`}_Ba_^__[YBBBBBBBBBRT\]]
```

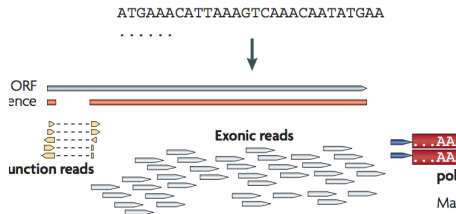
# Now what do we do?

- Map reads to a genome/annotation
  - (or not! - *de novo*)



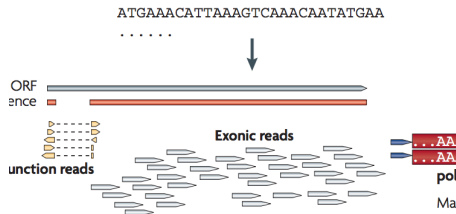
# Now what do we do?

- Map reads to a genome/annotation
  - (or not! - *de novo*)
- Count depth of reads at each gene



# Now what do we do?

- Map reads to a genome/annotation
  - (or not! - *de novo*)
- Count depth of reads at each gene
- Infer amount of RNA (and thus protein) in the cells



- Map to genes? or exons?
  - Exon mapping can help determine splice variants

# Statistics

- Map to genes? or exons?
  - Exon mapping can help determine splice variants
- Standardize by amount of data

# Statistics

- Map to genes? or exons?
  - Exon mapping can help determine splice variants
- Standardize by amount of data
  - FPM - Fragments (mapped) Per Million bases (of NGS data)

# Statistics

- Map to genes? or exons?
  - Exon mapping can help determine splice variants
- Standardize by amount of data
  - FPM - Fragments (mapped) Per Million bases (of NGS data)
- Standardize by length of gene as well



# Statistics

- Map to genes? or exons?
  - Exon mapping can help determine splice variants
- Standardize by amount of data
  - FPM - Fragments (mapped) Per Million bases (of NGS data)
- Standardize by length of gene as well
  - FPKM - Fragments Per Kilobase of transcript per Million mapped reads

# Statistics

- Map to genes? or exons?
  - Exon mapping can help determine splice variants
- Standardize by amount of data
  - FPM - Fragments (mapped) Per Million bases (of NGS data)
- Standardize by length of gene as well
  - FPKM - Fragments Per Kilobase of transcript per Million mapped reads
- Sampling error at low numbers

# Statistics

- Map to genes? or exons?
  - Exon mapping can help determine splice variants
- Standardize by amount of data
  - FPM - Fragments (mapped) Per Million bases (of NGS data)
- Standardize by length of gene as well
  - FPKM - Fragments Per Kilobase of transcript per Million mapped reads
- Sampling error at low numbers
  - Not always normal variance (can't sample below 0)

# Statistics

- Map to genes? or exons?
  - Exon mapping can help determine splice variants
- Standardize by amount of data
  - FPM - Fragments (mapped) Per Million bases (of NGS data)
- Standardize by length of gene as well
  - FPKM - Fragments Per Kilobase of transcript per Million mapped reads
- Sampling error at low numbers
  - Not always normal variance (can't sample below 0)
  - Poisson distribution

# Many choices produce many results

- Different analyses software packages

# Many choices produce many results

- Different analyses software packages
- Different choices (laid out on prior slide)

# Many choices produce many results

- Different analyses software packages
- Different choices (laid out on prior slide)
- Expectation is noisy results

# Many choices produce many results

- Different analyses software packages
- Different choices (laid out on prior slide)
- Expectation is noisy results
- Recent feud between kallisto and salmon authors



# What can RNAseq tell us?

- Which genes are being transcribed

# What can RNAseq tell us?

- Which genes are being transcribed
  - What genes/proteins are involved in specific functions

# What can RNAseq tell us?

- Which genes are being transcribed
  - What genes/proteins are involved in specific functions
  - How those genes change with conditions

# What can RNAseq tell us?

- Which genes are being transcribed
  - What genes/proteins are involved in specific functions
  - How those genes change with conditions
- Allele-Specific Expression (ASE)

# What can RNAseq tell us?

- Which genes are being transcribed
  - What genes/proteins are involved in specific functions
  - How those genes change with conditions
- Allele-Specific Expression (ASE)
  - Which version of a gene is being expressed more

# What can RNAseq tell us?

- Which genes are being transcribed
  - What genes/proteins are involved in specific functions
  - How those genes change with conditions
- Allele-Specific Expression (ASE)
  - Which version of a gene is being expressed more
  - How smaller changes in a gene alter expression

# What can RNAseq tell us?

- Which genes are being transcribed
  - What genes/proteins are involved in specific functions
  - How those genes change with conditions
- Allele-Specific Expression (ASE)
  - Which version of a gene is being expressed more
  - How smaller changes in a gene alter expression
- Where the genes are in a genome?

# What can RNAseq tell us?

- Which genes are being transcribed
  - What genes/proteins are involved in specific functions
  - How those genes change with conditions
- Allele-Specific Expression (ASE)
  - Which version of a gene is being expressed more
  - How smaller changes in a gene alter expression
- Where the genes are in a genome?
  - Help annotation (listing exons and introns) of new genomes



# What can RNAseq tell us?

- Which genes are being transcribed
  - What genes/proteins are involved in specific functions
  - How those genes change with conditions
- Allele-Specific Expression (ASE)
  - Which version of a gene is being expressed more
  - How smaller changes in a gene alter expression
- Where the genes are in a genome?
  - Help annotation (listing exons and introns) of new genomes
  - Quick way to analyze functional portion of the genome, target of selection

# Strengths of RNAseq

- No *a priori* info needed

# Strengths of RNAseq

- No *a priori* info needed
- No limit to discrimination

# Strengths of RNAseq

- No *a priori* info needed
- No limit to discrimination
- Systematically reproducible

# Strengths of RNAseq

- No *a priori* info needed
- No limit to discrimination
- Systematically reproducible
  - qPCR to verify results

# Strengths of RNAseq

- No *a priori* info needed
- No limit to discrimination
- Systematically reproducible
  - qPCR to verify results
  - Controls of known concentration

# Weaknesses of RNAseq

- Library generation has biases

# Weaknesses of RNAseq

- Library generation has biases
  - Fragmentation of long RNA depletes the ends of gene transcripts



# Weaknesses of RNAseq

- Library generation has biases
  - Fragmentation of long RNA depletes the ends of gene transcripts
  - cDNA based towards 3-prime end

# Weaknesses of RNAseq

- Library generation has biases
  - Fragmentation of long RNA depletes the ends of gene transcripts
  - cDNA based towards 3-prime end
- Abundant RNA read or PCR-duplicate artifact?

# Weaknesses of RNAseq

- Library generation has biases
  - Fragmentation of long RNA depletes the ends of gene transcripts
  - cDNA based towards 3-prime end
- Abundant RNA read or PCR-duplicate artifact?
  - Compare across biological replicates

# Weaknesses of RNAseq

- Library generation has biases
  - Fragmentation of long RNA depletes the ends of gene transcripts
  - cDNA based towards 3-prime end
- Abundant RNA read or PCR-duplicate artifact?
  - Compare across biological replicates
- Messy results

# Weaknesses of RNAseq

- Library generation has biases
  - Fragmentation of long RNA depletes the ends of gene transcripts
  - cDNA based towards 3-prime end
- Abundant RNA read or PCR-duplicate artifact?
  - Compare across biological replicates
- Messy results
- Bioinformatic methods (largely improved)

# Weaknesses of RNAseq

- Library generation has biases
  - Fragmentation of long RNA depletes the ends of gene transcripts
  - cDNA based towards 3-prime end
- Abundant RNA read or PCR-duplicate artifact?
  - Compare across biological replicates
- Messy results
- Bioinformatic methods (largely improved)
- Strand specificity (also fixable)

# “New” Insights

(from the 2009 paper)

- Mapping genes and exons

# “New” Insights

(from the 2009 paper)

- Mapping genes and exons
- Catalog transcript complexity



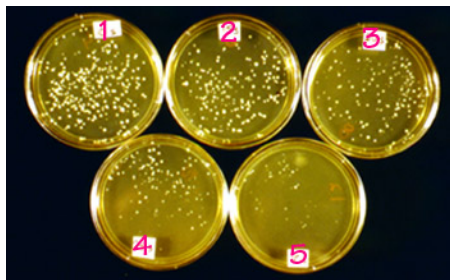
# “New” Insights

(from the 2009 paper)

- Mapping genes and exons
- Catalog transcript complexity
- Transcribed and not translated regions

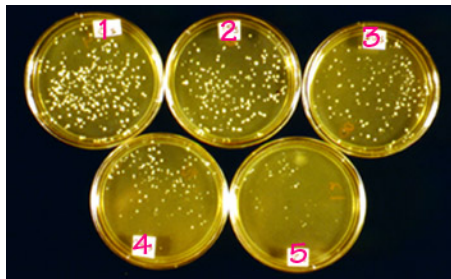
# A Parable

- Steve and Susan wanted to know how food source affects yeast



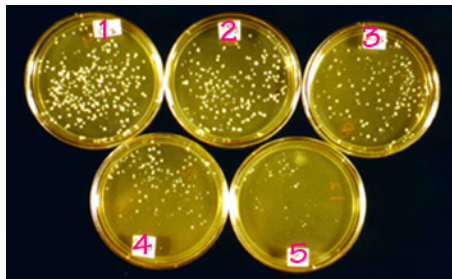
# A Parable

- Steve and Susan wanted to know how food source affects yeast
- They decide to grow the same yeast in 5 different plates in their classroom



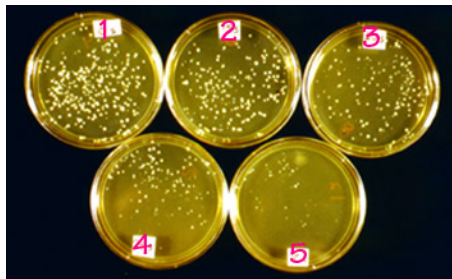
# A Parable

- Each plate contains .8% lactose agar



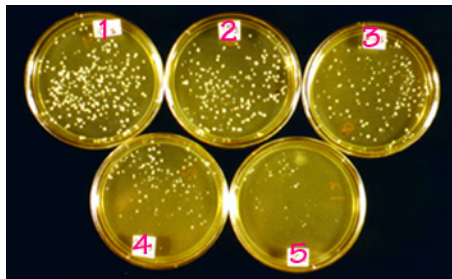
# A Parable

- Each plate contains .8% lactose agar
- At the end of a week the measure the gene expression of each



# A Parable

- Each plate contains .8% lactose agar
- At the end of a week the measure the gene expression of each
- What is wrong with this “experiment”?



# Dotstorming Check-In

- Did I cover the top questions?
- Dotstorming Board

# Learning to Code (and other skills)



# Learning to Code (and other skills)

**Jake VanderPlas**  
@jakevdp

Follow

My advice on learning Python:

Don't set out to "learn Python". Choose a problem you're interested in and learn to solve it with Python.

11:24 AM - 10 Sep 2017

2,414 Retweets 7,493 Likes



 179  2.4K  7.5K 



**Jake VanderPlas** @jakevdp · Sep 10

Replying to @jakevdp

(This applies to most things in tech, and in life as well)

 7  24  200 

# Meet your Group!

- Jennifer, Nayib, Chris

# Meet your Group!

- Jennifer, Nayib, Chris
- Austin, Alan, Raymond

# Meet your Group!

- Jennifer, Nayib, Chris
- Austin, Alan, Raymond
- Othmane, Kevin, Alvin

# Meet your Group!

- Jennifer, Nayib, Chris
- Austin, Alan, Raymond
- Othmane, Kevin, Alvin
- Helena, Jake, Rahul

# Meet your Group!

- Jennifer, Nayib, Chris
- Austin, Alan, Raymond
- Othmane, Kevin, Alvin
- Helena, Jake, Rahul
- Hank, Sisi, Joy

- Today's Tasks:

- Today's Tasks:
- Meet with your group



- Today's Tasks:
- Meet with your group
- Discuss common interests

- Today's Tasks:
- Meet with your group
- Discuss common interests
- Exchange and ideas or datasets you had hoped to use

- Today's Tasks:
- Meet with your group
- Discuss common interests
- Exchange and ideas or datasets you had hoped to use
- Leave with a plan to find data by next Tuesday

# The End