RNAseq Analysis: A Practical Walkthrough (part 2/3)

C. Ryan Campbell

Duke University

c.ryan.campbell@duke.edu

26 Oct 2017

Overview

Today's Goals

- Recap the workflow
- Discuss analyses
- Take the next analysis step

RNAseq Workflow

1. Get in your groups
2. Recreate the diagram from two weeks ago
   - Pick a step
   - Fill in a description and software
   - Should take about 10 minutes

Group Guessing Game - No Phones!

- How tall, in feet, is Duke Chapel?
- How many people does it hold?
- What year was it completed?
- Your answer = 10 * height + capacity + year completed

Group Guessing Game - No Phones!

- Your answer = 10 * height + capacity + year completed
- Work with your group to guess the number and write it and your name on a piece of paper
- The most accurate guess will get first pick of which section to describe
- Make your guesses now, you have 2 minutes

The Answer...

- is 210 feet, 1800 occupants, completed in 1932
- (210 * 10) + 1800 + 1932 = 5832
- Pick which section you'd like to recap
  1. Library Prep
  2. Sequencing
  3. Quality Check Data
  4. Trim Data
  5. Align
- Recap - input, output, software

RNAseq Workflow

1. Get in your groups
2. Recreate the diagram from two weeks ago
   - Pick a step
   - Fill in a description and software
   - Should take about 10 minutes

slogin OR sbatch script

- Refresher: What is the difference?
- Make sure you're making a conscious choice between the two

- When you analyze your files, make sure to use an sbatch script

slogin OR laptop

- What is the difference?
- Which processes will go where?

- What are the strengths and weaknesses of each?

SLURM Interactive Node

- Later we'll be troubleshooting `htseq-count`
- For that you should use an "interactive node"
- This runs like a sbatch job, but it appears as a terminal that you can interact with

```
srun -mem-per-cpu=16000MB -pty bash -i
```

- You've just requested a 16GB (powerful laptop) size node on SLURM

tophat Output

- tophat - alignment software
- What (in english) is that doing?
- What is the output file type?

tophat bam files

- tophat output = bam file
- fastq data aligned to your genome
- How many bam files should you have per treament/condition?
- IGV can be used to “check” the bam

Counting & Analysis

- Now that we have bam files the next step is to count the reads
- And using those counts compare gene expression levels
- We'll be using HTseq and DeSeq2

Software Options

- There are many options for counting and analysis
- cufflinks (cuffdiff/cuffcompare/etc) is popular
- HTSeq and DESeq2 are more straightforward
- DESeq2 gets us into R quicker
- (thus my decision to use them for the tutorial)

HTSeq

- python-based program to count reads
- Input:
- .bam file and .gtf/.gff
- Output:
- A table of counts by gene

DESeq2

- R-based program to analyze expression
- Input:
- A table of counts by gene
- Output:
- Graphs and (hopefully) Answers!!!

Files to Use

- I've set up some example files to use for this tutorial
- They're human RNAseq files from a hypoxia experiment:

```
ls -lthr /work/cc216/490S/cc216/RNAseq_pt2
```

- What do you see? Which will you use?

Files to Use

```
ls -lthr /work/cc216/490S/cc216/RNAseq_pt2
```

```
cc216@dcc-slogin-01  /work/cc216/490S/cc216/RNAseq_pt2 $ ls -lthr
total 686M
lrwxrwxrwx. 1 cc216 root   59 Oct 25 23:02 hsap_annotations.gff -> /work/keh65/genomes/GCF_000001405.36_GRCh38.p10_genomic.gff
-rw-r--r--. 1 cc216 root 269M Oct 25 23:57 hsap_norm_accepted_hits_0.10.bam
-rw-r--r--. 1 cc216 root 259M Oct 26 00:01 hsap_hypox_accepted_hits_0.10.bam
-rw-r--r--. 1 cc216 root  21M Oct 26 01:01 hsap_hypox.counts
-rw-r--r--. 1 cc216 root    0 Oct 26 07:36 hsap_hypox_0.10.counts
drwxr-xr-x. 2 cc216 root  190 Oct 26 07:48 unused
```

htseq-count

- We'll be using htseq-count
- This will count the number of reads mapped to each gene
- That data will be taken into DESeq2

```
/opt/apps/rhel7/Python-2.7.11/bin/htseq-count
```

(go ahead and put it in your path)

SLURM Interactive Node

- Now that we're about to troubleshoot `htseq-count` hopefully your SLURM node is open

```
srun -mem-per-cpu=16000MB -pty bash -i
```

htseq-count

- What does HTSeq do?
- What are its flags and options?

```
htseq-count <options> <alignment bam> <gff file> > <count output>
```

- Probably important: -f, -s, -t

htseq-count

- Here are the flags that I got it to work with:

```
htseq-count -s no -r pos -t exon -i gene -f bam
hsap_hypox_accepted_hits_0.10.bam hsap_annotations.gff >
hsap_hypox_0.10.counts
```

DESeq2

- What does DESeq2 do?
- Compares the count matrices from many samples
- Where do you run it?
- In R on your laptop

DESeq2 guides

- We'll get into DESeq2 next week
- If you want to get started here are some guides:
- Walkthrough
- Bioconductor Manual
- Bioconductor Walkthrough

The End