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

C. Ryan Campbell

Duke University

c.ryan.campbell@duke.edu

31 Oct 2017

Overview

1. Workflow
2. Review Steps
   - Indexing
3. Alignment
   - tophat
4. Counting
5. Tutorial Files
   - Commands
   - DESeq2

Today's Goals

- Logon to interactive SLURM node
- Run tophat2
- Run htseq-count

Workflow

1. Fill in the missing blanks on my diagram
   - Fill in either a file status or a software/process
   - Should take about 1 minute

slogin OR sbatch script

- What is the difference?
- Make sure you're making a conscious choice between the two
- Today we'll be working on slogin with SMALL files
- (why does this matter?)
- When you analyze your files, make sure to use an sbatch script

SLURM Interactive Node

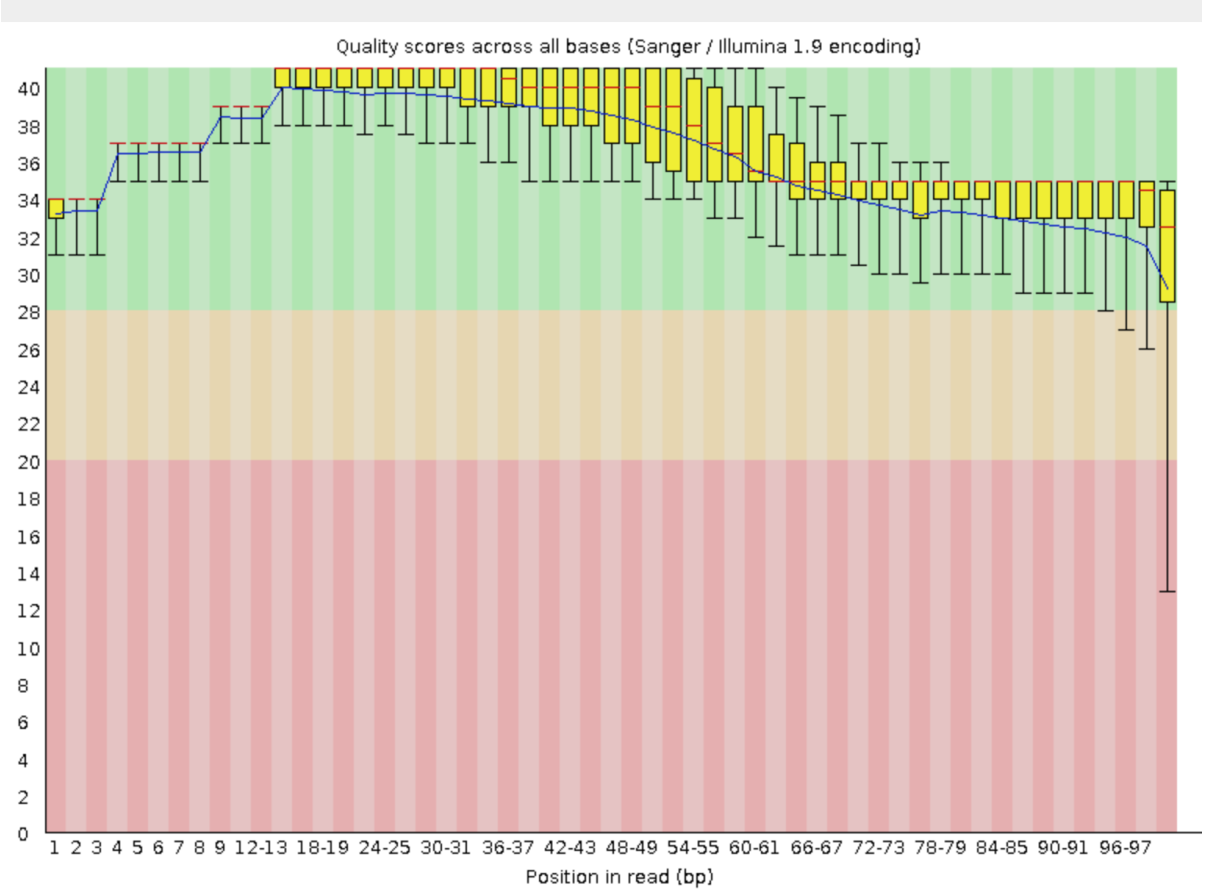
- Later we'll be troubleshooting `htseq-count`
- For that you should use an "interactive node"
- This runs like a sbatch job, but it appears as a terminal that you can interact with

```
srun -mem-per-cpu=4000MB -pty bash -i
```

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

fastqc output

- Run Quality
- Base Content
- Run Length



Trimming Data

- Remove data that is low quality
  - You have TONS of data, taking out 5% is OK
- We will set a couple of parameters:
  - Minimum quality at the end of the read
  - Average quality along a sliding window
  - Overall read quality
- If the read doesn't meet some, or all, of these the whole read is tossed
- We'll be using trimmomatic

trimmomatic example

- To run trimmomatic:

```
EXAMPLE:

java -jar
/work/cc216/490S/software/Trimmomatic-0.36/trimmomatic-0.36.jar
<PE or SE> -phred33 -trimlog <output log> <Read 1.fq> <Read
2.fq> <Read 1 output> <Read 1 output unpaired> <Read 2 output>
<Read 2 output unpaired> LEADING:3 TRAILING:3 SLIDINGWINDOW:5:20
MINLEN:50
```

- Remove (leading/trailing) low quality or N bases (below quality 3)
- Scan the read with a 5-base wide sliding window, cutting when the average quality per base drops below 20
- Drop reads which are less than 50 bases long after these steps

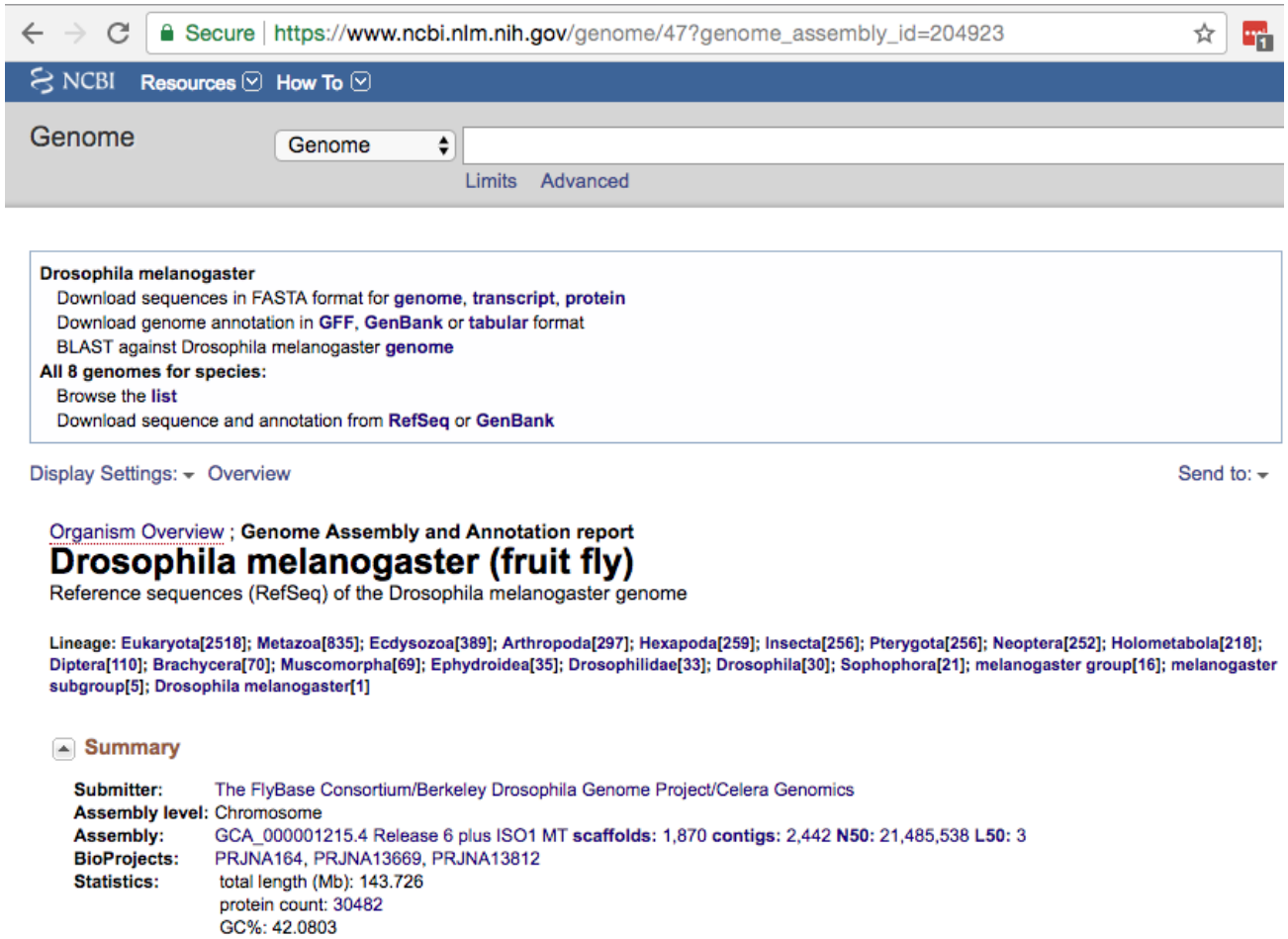
trimmomatic example

- Output:

```
Input Read Pairs:  1000000 Both Surviving:  955447 (95.54%)
Forward Only Surviving:  29029 (2.90%) Reverse Only Surviving:
9577 (0.96%) Dropped:  5947 (0.59%)

>head RNAseq.log

SRR848963.63 ILLUMINA:322:D0UFKACXX:3:1101:11445:2184 length=101
101 0 101 0

SRR848963.64 ILLUMINA:322:D0UFKACXX:3:1101:11909:2032 length=101
97 1 98 3

SRR848963.64 ILLUMINA:322:D0UFKACXX:3:1101:11909:2032 length=101
98 0 98 3
```

Downloading genomes

- Model species have a page like this
- Download from the “FASTA format for genome” and “annotation in GFF” links



Indexing

- Most aligners require their genome to be indexed
- What do you think this means?
- You'll need to index using bowtie2 (the aligner tophat2 uses)

tophat

- tophat2 - alignment software
- In: Sequence data
- Out: .bam file - where that data aligns to/fits in the genome

tophat

- tophat2 is the command/software that aligns the reads to the genome
- This (also) is going to be a computationally intensive process
- So write a submission script to do it:

```
export PATH=/opt/apps/tophat-bowtie/:$PATH


tophat2
```

tophat

- tophat2 example (all one line):

```
tophat2 -p 4 -o s01 -G
/work/cc216/490S/cc216/RNAseq_pt2/hsap_annotations.gff
/work/cc216/490S/cc216/RNAseqpt2/hsap
/work/cc216/490S/cc216/RNAseq_pt2/s01_hsap_norm_R1.fastq
/work/cc216/490S/cc216/RNAseq_pt2/s01_hsap_norm_R2.fastq
```

translated:

```
tophat2 -p <number of threads> -o <output dir> -G <gff file,
annotations> <bowtie2 index> <R1 fastq> <R2 fastq>
```

- Help can be found by running “tophat2”
- Or in the tophat2 manual online
- http://ccb.jhu.edu/software/tophat/manual.shtml

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

HTSeq

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

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)

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

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
```

SLURM Interactive Node

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

```
srun -mem-per-cpu=4000MB -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

tophat

- tophat2 example (all one line):

```
tophat2 -p 4 -o s01 -G
/work/cc216/490S/cc216/RNAseq_pt2/hsap_annotations.gff
/work/cc216/490S/cc216/RNAseqpt2/hsap
/work/cc216/490S/cc216/RNAseq_pt2/s01_hsap_norm_R1.fastq
/work/cc216/490S/cc216/RNAseq_pt2/s01_hsap_norm_R2.fastq
```

translated:

```
tophat2 -p <number of threads> -o <output dir> -G <gff file,
annotations> <bowtie2 index> <R1 fastq> <R2 fastq>
```

- Help can be found by running “tophat2”
- Or in the tophat2 manual online
- http://ccb.jhu.edu/software/tophat/manual.shtml

Today's Goals

1. Logon to interactive SLURM node
2. Run tophat2
3. Run htseq-count

DESeq2

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

DESeq2

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

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