RNAseq Walkthrough (now with Genes & Geography)

C. Ryan Campbell

Duke University

c.ryan.campbell@duke.edu

2 Nov 2017

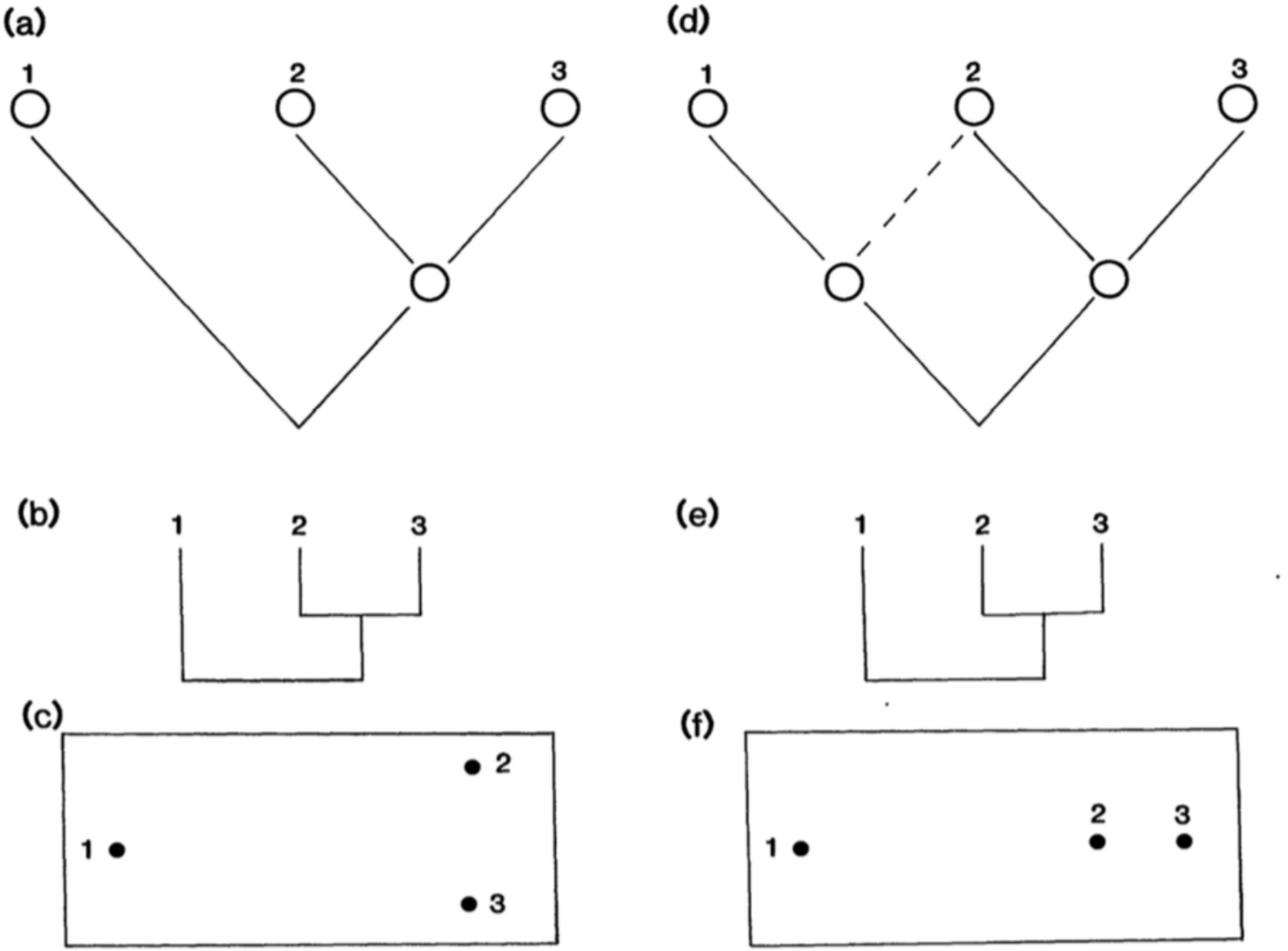
Overview

1. Genes & Geography
2. Workflow
3. Alignment
   - tophat
4. Counting
5. Tutorial Files
   - Commands
   - DESeq2

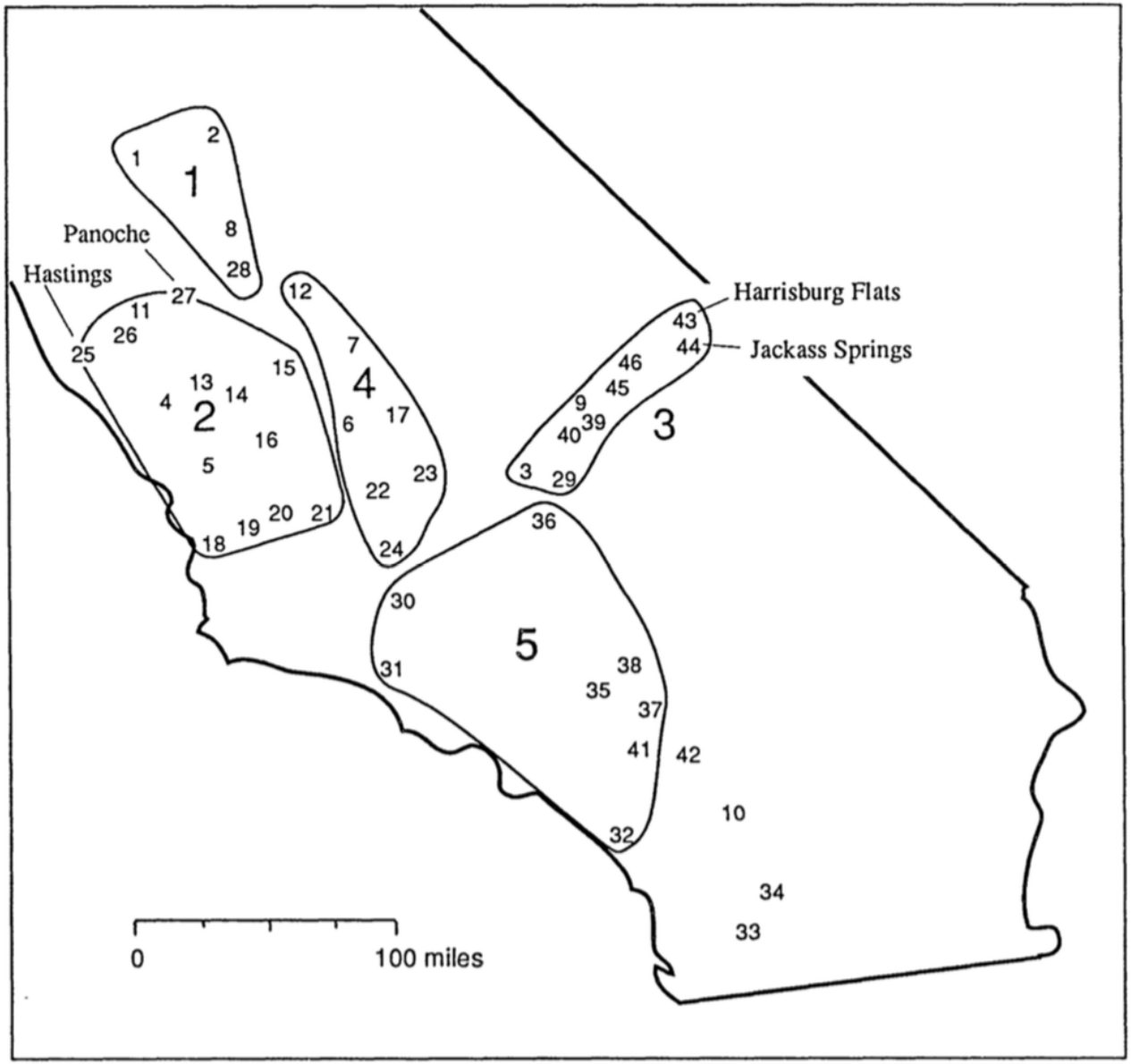
Today's Goals

- Run tophat2
- Run htseq-count
- Understand htseq-count output

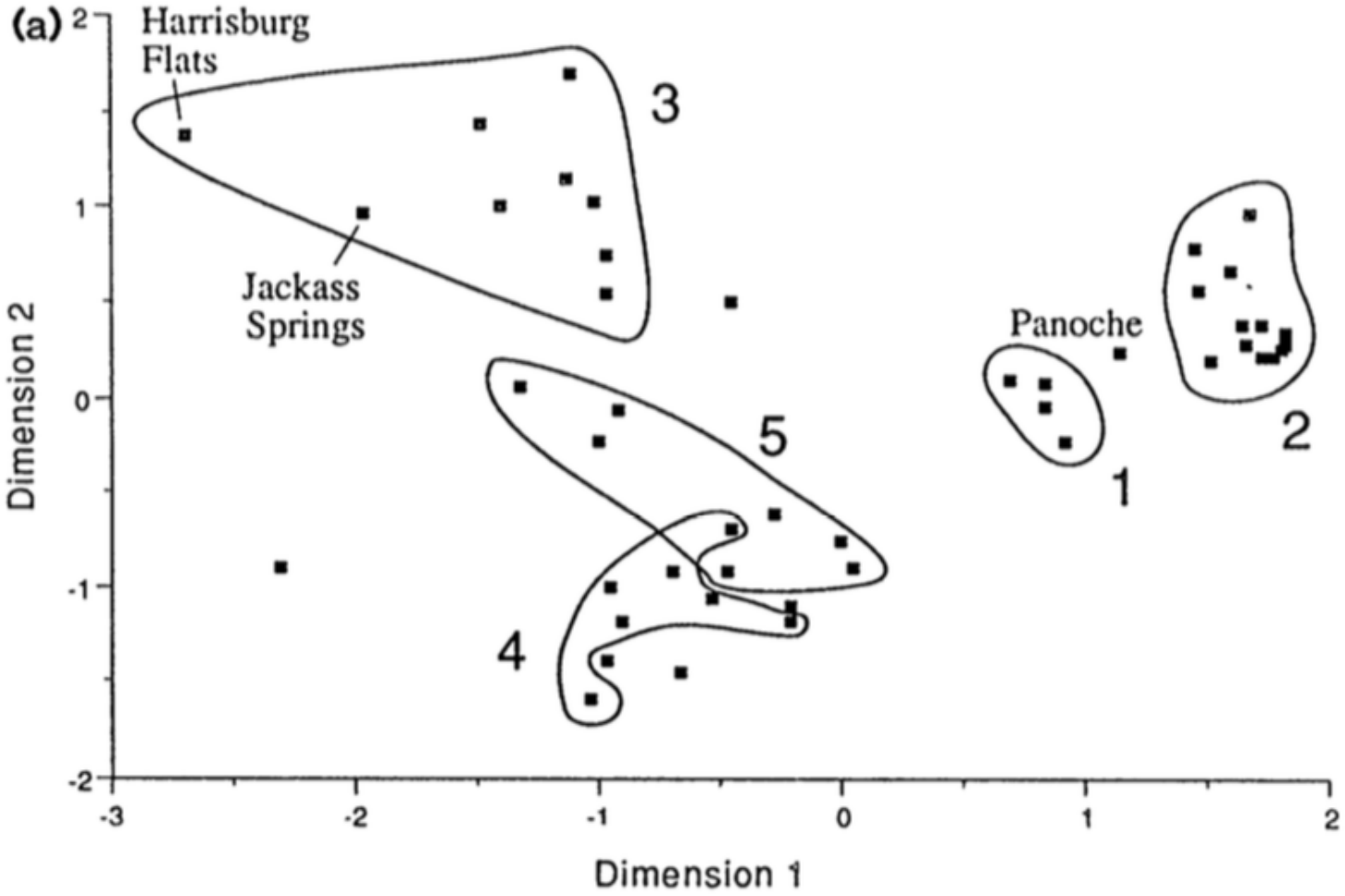
Lessa 1990



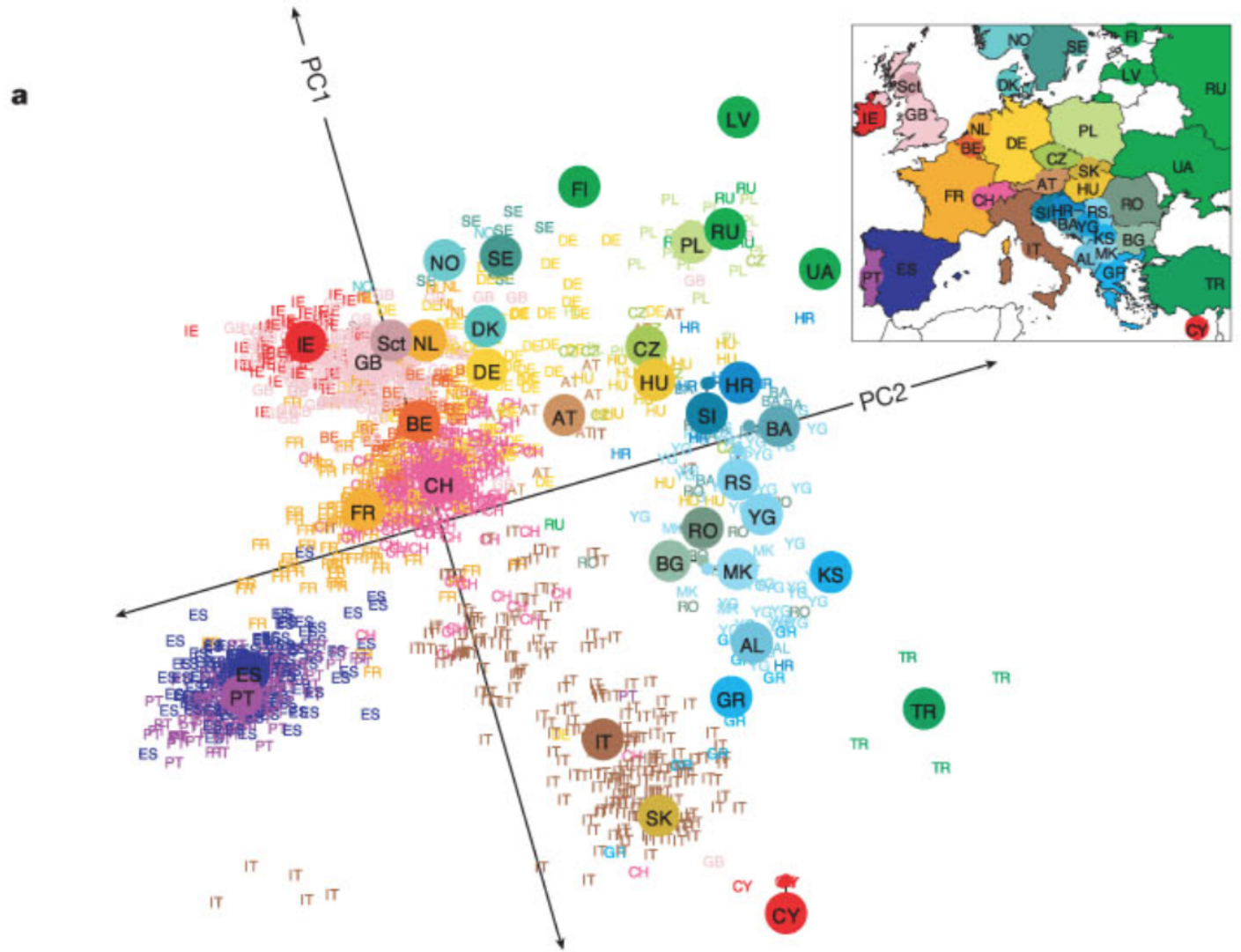
Lessa 1990


1
2
3
4
5
Panoche
Hastings
Harrisburg Flats
Jackass Springs
0
100 miles

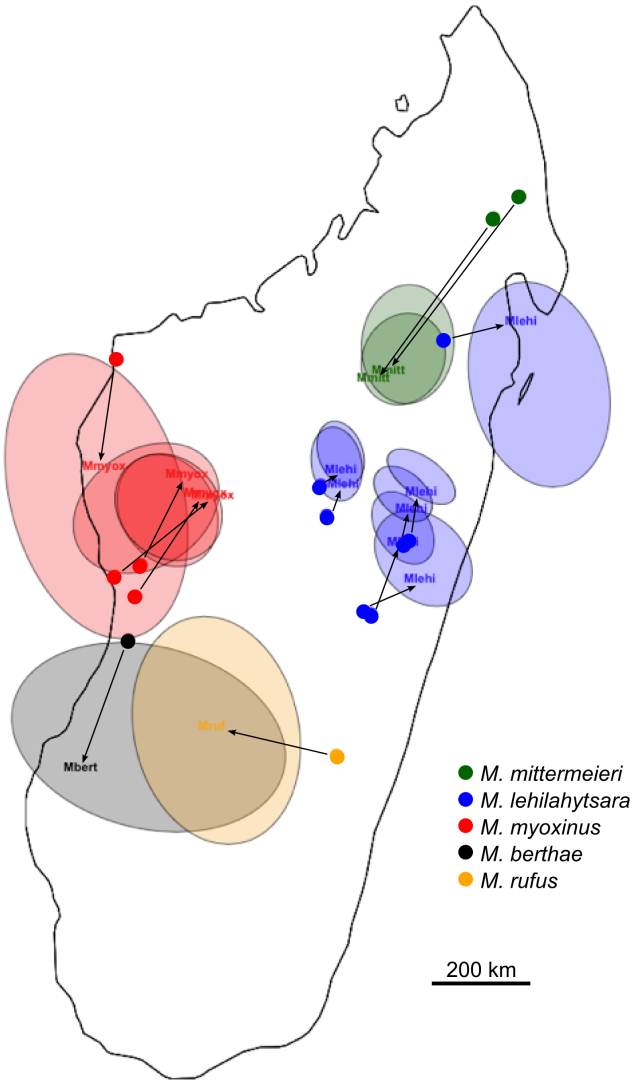
Lessa 1990


(a)
Harrisburg Flats
Jackass Springs
Panoche
1
2
3
4
5
Dimension 2
Dimension 1
2
1
0
-1
-2
-3
-2
-1
0
1
2

Novembre 2008



Mouse Lemurs



Novembre Talk

UPGG Distinguished Lecturer Series

John Novembre, University of Chicago

"New lenses on human genetic variation: Tools for interpreting geographic structure in genetic data"

4:00pm, 103 Bryan Research Auditorium

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 4GB (powerful laptop) size node on SLURM

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
```

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
```

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

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

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 Link
- Focus on "Quick Start" and more specifically:
- Setting the R objects `cts` and `coldata` correctly
- Using `paste` (a unix command) to format your data into `cts`

  http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

The End