Exercise 1 Review

Make a shell script

tophat -o A -G testgenome.gff3 --no-novel-juncs testgenome a.fastq.gz

tophat -o B -G testgenome.gff3 --no-novel-juncs testgenome b.fastq.gz

mv A/accepted_hits.bam ./a.bam

mv B/accepted_hits.bam ./b.bam

samtools index a.bam

samtools index b.bam

Run a shell script

nohup sh /home/my_user_ID/runtophat.sh >& mylog &
PATH in Linux

Absolute PATH

/workdir/mydir/myDataFile

Relative PATH

myDataFile
my_Directory/ myDataFile
./myDataFile
../ myDataFile
PATH in Linux

```
$ pwd
/workdir/ff111
```

Use “pwd” to get current directory

tophat -o mydir testgenome a.fastq.gz
mv mydir/accepted_hits.bam ./a.bam

nohup sh /home/my_user_ID/runtophat.sh >& mylog &
Genome Databases for TOPHAT

• On /local_data directory:
  human, mouse, Drosophila, C. elegans, yeast, Arabidopsis, maize.

• On /shared_data/genome_db/:
  rice, grape, apple, older versions of databases.
Create aliases for files

ln -s /local_data/Homo_sapiens_UCSC_hg19/Bowtie2Index/* ./

tophat /local_data/Homo_sapiens_UCSC_hg19/Bowtie2Index/genome a.fastq.gz

tophat genome a.fastq.gz
How to prepare TOPHAT genome database

bowtie2-build rice7.fa rice7

Genome fasta file
Give the database a name

* Keep a copy of the indexed genome in home directory so that the files can be reused next time
RNA-seq Data Analysis
Lecture 2

1. Quantification (count reads per gene)

2. Normalization (normalize counts between samples)

3. Differentially expressed genes
Different summarization strategies will result in the inclusion or exclusion of different sets of reads in the table of counts.
Complications in quantification

1. Multi-mapped reads

Cufflinks/Cuffdiff

- uniformly divide each read to all mapped positions
- multi-mapped read correction (default off, can be enabled with --multi-read-correct option)

HTSeq

- Count unique and multi-mapped reads separately
Complications in quantification

2. Assign reads to isoforms

Cufflinks/Cuffdiff

- Use its own model to estimate isoform abundance;

HTSeq

- A set of arbitrary rules specified by mode option, including (a) skip or (b) counted towards each feature.

* Gene level read counts is more reliable than isoform level read counts
2. Normalization

MA Plots

Before normalization

After normalization

- Y axis: log ratio of expression level between two conditions;
- With the assumption that most genes are expressed equally, the log ratio should mostly be close to 0
A simple normalization

**FPKM (CUFFLINKS)**
Fragments Per Kilobase Of Exon Per Million Fragments
Normalization factor:
- compatible-hits-norm: reads compatible with reference transcripts
- total-hits-norm: all reads

**CPM (EdgeR)**
Count Per Million Reads
Normalization factor:
- reads compatible with reference transcripts
- Normalized with TMM
Default in EdgeR: TMM Normalization

Normalization methods

- **Total-count normalization**
  - By total mapped reads

- **Upper-quantile normalization**
  - By read count of the gene at upper-quantile

- **Normalization by housekeeping genes**

- **Trimmed mean (TMM) normalization**
Normalization methods

- **Total-count normalization (FPKM, RPKM)**
  - By total mapped reads

- **Upper-quantile normalization**
  - By read count of the gene at upper-quantile

- **Normalization by housekeeping genes**

- **Trimmed mean (TMM) normalization**

  - Default: cuffdiff

  - EdgeR & DESeq
3. Differentially expressed genes

If we could do 100 biological replicates,

Distribution of Expression Level of A Gene

- Condition 1
- Condition 2
The reality is, we could only do 3 replicates,
Statistical modeling of gene expression and test for differentially expressed genes

1. **Estimate of variance.**
   Eg. EdgeR uses a combination of
   1) a common dispersion effect from all genes;
   2) a gene-specific dispersion effect.

2. **Model the expression level with negative binomial distribution.**
   DESeq and EdgeR

3. **Multiple test correction**
   Default in EdgeR: Benjamini-Hochberg
For each gene:

- Read count (raw & normalized)
- Fold change (Log2 fold)
- P-value
- Q(FDR) value.

Output from RNA-seq pipeline

Using both fold change and FDR value to filter:

E.g. Log2(fold) > 1 or < -1 & FDR < 0.05
<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Cuffdiff</th>
<th>DESeq</th>
<th>edgeR</th>
<th>limmaVoom</th>
<th>PoissonSeq</th>
<th>baySeq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalization and clustering</td>
<td></td>
<td></td>
<td></td>
<td>All methods performed equally well</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE detection accuracy measured by AUC at increasing qRT-PCR cutoff</td>
<td>Decreasing</td>
<td>Consistent</td>
<td>Consistent</td>
<td>Decreasing</td>
<td>Increases up to log expression change ≤ 2.0</td>
<td>Consistent</td>
</tr>
<tr>
<td>Null model type I error</td>
<td>High number of FPs</td>
<td>Low number of FPs</td>
<td>Low number of FPs</td>
<td>Low Number of FPs</td>
<td>Low number of FPs</td>
<td>Low number of FPs</td>
</tr>
<tr>
<td>Signal-to-noise vs P value correlation for genes detected in one condition</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Good</td>
<td>Moderate</td>
<td>Good</td>
</tr>
<tr>
<td>Support for multi-factored experiments</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Support DE detection without replicated samples</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Detection of differential isoforms</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Runtime for experiments with three to five replicates on a 12 dual-core 3.33 GHz, 100 G RAM server</td>
<td>Hours</td>
<td>Minutes</td>
<td>Minutes</td>
<td>Minutes</td>
<td>Seconds</td>
<td>Hours</td>
</tr>
</tbody>
</table>

AUC, area under curve; DE, differential expression; FP, false positive.
RNA-seq Workflow at Bioinformatics Facility

Tophat -> BAM files

Cuffdiff -> raw read counts
(File: genes.read_group_tracking)

EdgeR -> Normalization & DE Genes

http://cbsu.tc.cornell.edu/lab/doc/rna_seq_draft_v8.pdf
Using Cuffdiff for Quantification

• **Cufflinks**
  – Input: one single BAM from TOPHAT;
  – Reference guide transcript assembly;
  – Output: GTF

• **Cuffdiff**
  – Input: multiple BAM files from TOPHAT;
  – Quantification & DE gene detection
  – Output: Read count; DE gene list
CUFFDIFF command

cuffdiff -p 2 -o outDir rice7.gff3  \
A_r1.bam,A_r2.bam  B_r1.bam,B_r2.bam

A_r1 :  timepoint 1; repeat 1
A_r2 :  timepoint 1; repeat 2

B_r1 :  timepoint 1; repeat 1
B_r2 :  timepoint 2; repeat 2
Connection between CUFFDIFF and EdgeR

CUFFDIFF output file with raw read count: genes.read_group_tracking

<table>
<thead>
<tr>
<th>tracking_id</th>
<th>condition</th>
<th>replicate</th>
<th>raw_frags</th>
<th>internal_scaled_fra gs</th>
<th>external_scaled_fra gs</th>
<th>FPKM</th>
<th>effective_length</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>q1</td>
<td>0</td>
<td>16</td>
<td>11.3905</td>
<td>11.3905</td>
<td>0.305545</td>
<td>- OK</td>
<td></td>
</tr>
<tr>
<td>gene1</td>
<td>q1</td>
<td>1</td>
<td>12</td>
<td>8.08334</td>
<td>8.08334</td>
<td>0.216832</td>
<td>- OK</td>
<td></td>
</tr>
<tr>
<td>gene1</td>
<td>q2</td>
<td>0</td>
<td>15</td>
<td>26.084</td>
<td>26.084</td>
<td>0.699692</td>
<td>- OK</td>
<td></td>
</tr>
<tr>
<td>gene1</td>
<td>q2</td>
<td>1</td>
<td>19</td>
<td>21.9805</td>
<td>21.9805</td>
<td>0.589617</td>
<td>- OK</td>
<td></td>
</tr>
<tr>
<td>gene2</td>
<td>q1</td>
<td>0</td>
<td>61</td>
<td>43.4262</td>
<td>43.4262</td>
<td>4.50677</td>
<td>- OK</td>
<td></td>
</tr>
<tr>
<td>gene2</td>
<td>q1</td>
<td>1</td>
<td>53</td>
<td>35.7014</td>
<td>35.7014</td>
<td>3.69312</td>
<td>- OK</td>
<td></td>
</tr>
<tr>
<td>gene2</td>
<td>q2</td>
<td>0</td>
<td>35</td>
<td>60.8627</td>
<td>60.8627</td>
<td>6.35236</td>
<td>- OK</td>
<td></td>
</tr>
<tr>
<td>gene2</td>
<td>q2</td>
<td>1</td>
<td>30</td>
<td>34.7061</td>
<td>34.7061</td>
<td>3.59016</td>
<td>- OK</td>
<td></td>
</tr>
</tbody>
</table>

EdgeR input file:

<table>
<thead>
<tr>
<th>Gene</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>16</td>
<td>12</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>gene2</td>
<td>61</td>
<td>53</td>
<td>35</td>
<td>30</td>
</tr>
</tbody>
</table>

File conversion PERL script:

```
parse_cuffdiff_readgroup.pl
```

- The script would produce a raw read count table (edgeR_count.xls) and a FPKM table (edgeR_FPKM.xls).
- If you want to get this script, you can use FileZilla to download it, it is located at /programs/bin/perlscripts/parse_cuffdiff_readgroup.pl
Using EdgeR to make MDS plot of the samples

- Check reproducibility from replicates, remove outliers
- Check batch effects;
Use EdgeR to identify DE genes

<table>
<thead>
<tr>
<th>Treat</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1-3</td>
<td>Drug</td>
</tr>
<tr>
<td>Sample 4-6</td>
<td>Drug</td>
</tr>
<tr>
<td>Sample 7-9</td>
<td>Drug</td>
</tr>
</tbody>
</table>

group <- factor(c(1,1,1,2,2,2,3,3,3))
design <- model.matrix(~0+group)
fit <- glmFit(myData, design)

lrt12 <- glmLRT(fit, contrast=c(1,-1,0))  # compare 0 vs 1h
lrt13 <- glmLRT(fit, contrast=c(1,0,-1))  # compare 0 vs 2h
lrt23 <- glmLRT(fit, contrast=c(0,1,-1))  # compare 1 vs 2h
### Multiple-factor Analysis in EdgeR

<table>
<thead>
<tr>
<th></th>
<th>Treat</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1-3</td>
<td>Placebo</td>
<td>0 hr</td>
</tr>
<tr>
<td>Sample 4-6</td>
<td>Placebo</td>
<td>1 hr</td>
</tr>
<tr>
<td>Sample 7-9</td>
<td>Placebo</td>
<td>2 hr</td>
</tr>
<tr>
<td>Sample 10-12</td>
<td>Drug</td>
<td>0 hr</td>
</tr>
<tr>
<td>Sample 13-15</td>
<td>Drug</td>
<td>1 hr</td>
</tr>
<tr>
<td>Sample 16-18</td>
<td>Drug</td>
<td>2 hr</td>
</tr>
</tbody>
</table>

```r
group <- factor(c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6))
design <- model.matrix(~0+group)
fit <- glmFit(mydata, design)

lrt <- glmLRT(fit, contrast=c(-1,0,1,1,0,-1))
### equivalent to (Placebo.2hr – Placebo.0hr) – (Drug.2hr-Drug.1hr)
```
Exercise

• Using cuffdiff for quantification and identifying differentially expressed genes of two different biological conditions A and B. There are two replicates for each condition.

• Using EdgeR package to make MDS plot of the 4 libraries, and identify differentially expressed genes