

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

Use “pwd” to get current directory

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

```
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**

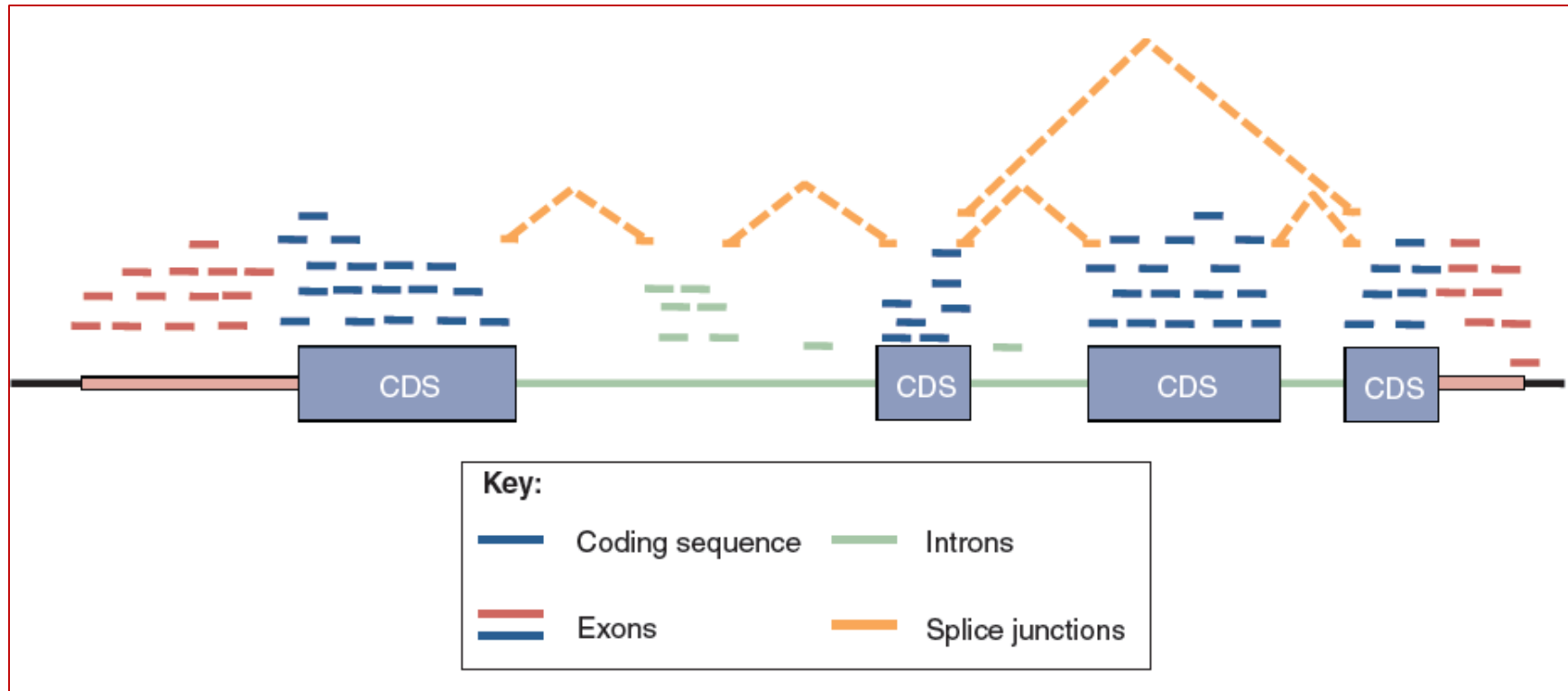
* Keep a copy of the indexed genome in home directory so that the files can be reused next time

RNA-seq Data Analysis

Lecture 2

- 1. Quantification** (count reads per gene)
- 2. Normalization** (normalize counts between samples)
- 3. Differentially expressed genes**

Quantification: Count reads per gene



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

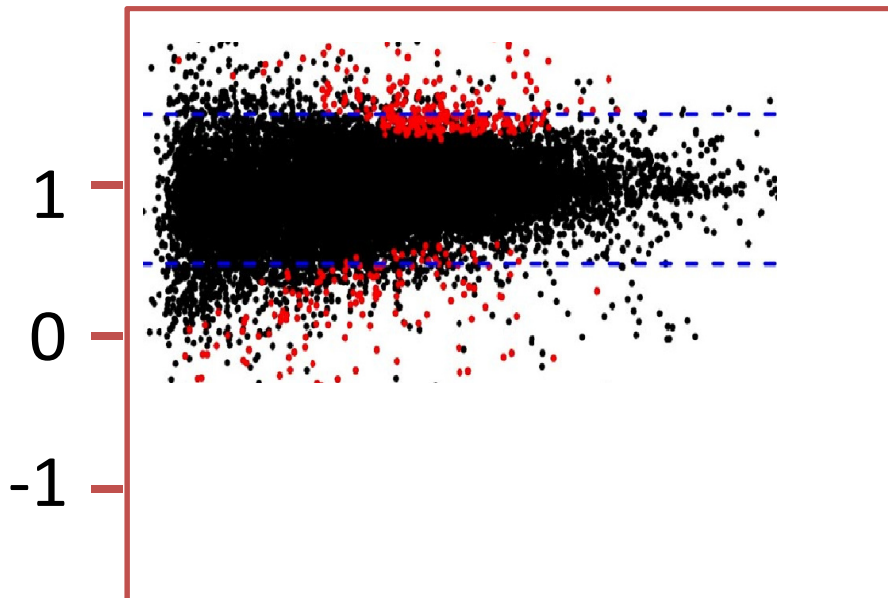
- A set of arbitrary rules specified by mode option, including (a)skip or (b)counted towards each feature.

* Gene level read counts is more reliable than isoform level read counts

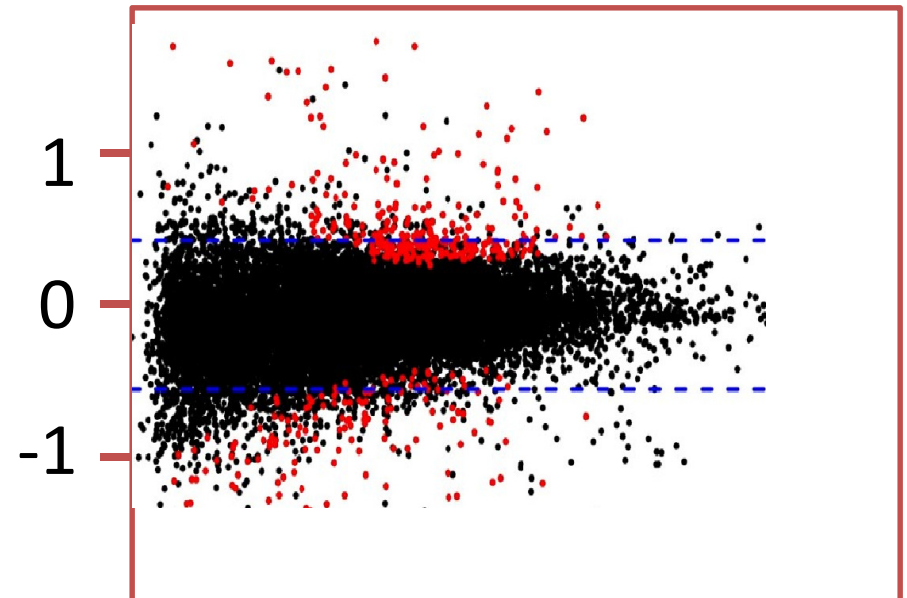
2. Normalization

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 **P**er **K**ilobase **O**f Exon **P**er **M**illion Fragments

Normalization factor:

- compatible-hits-norm: reads compatible with reference transcripts
- total-hits-norm: all reads

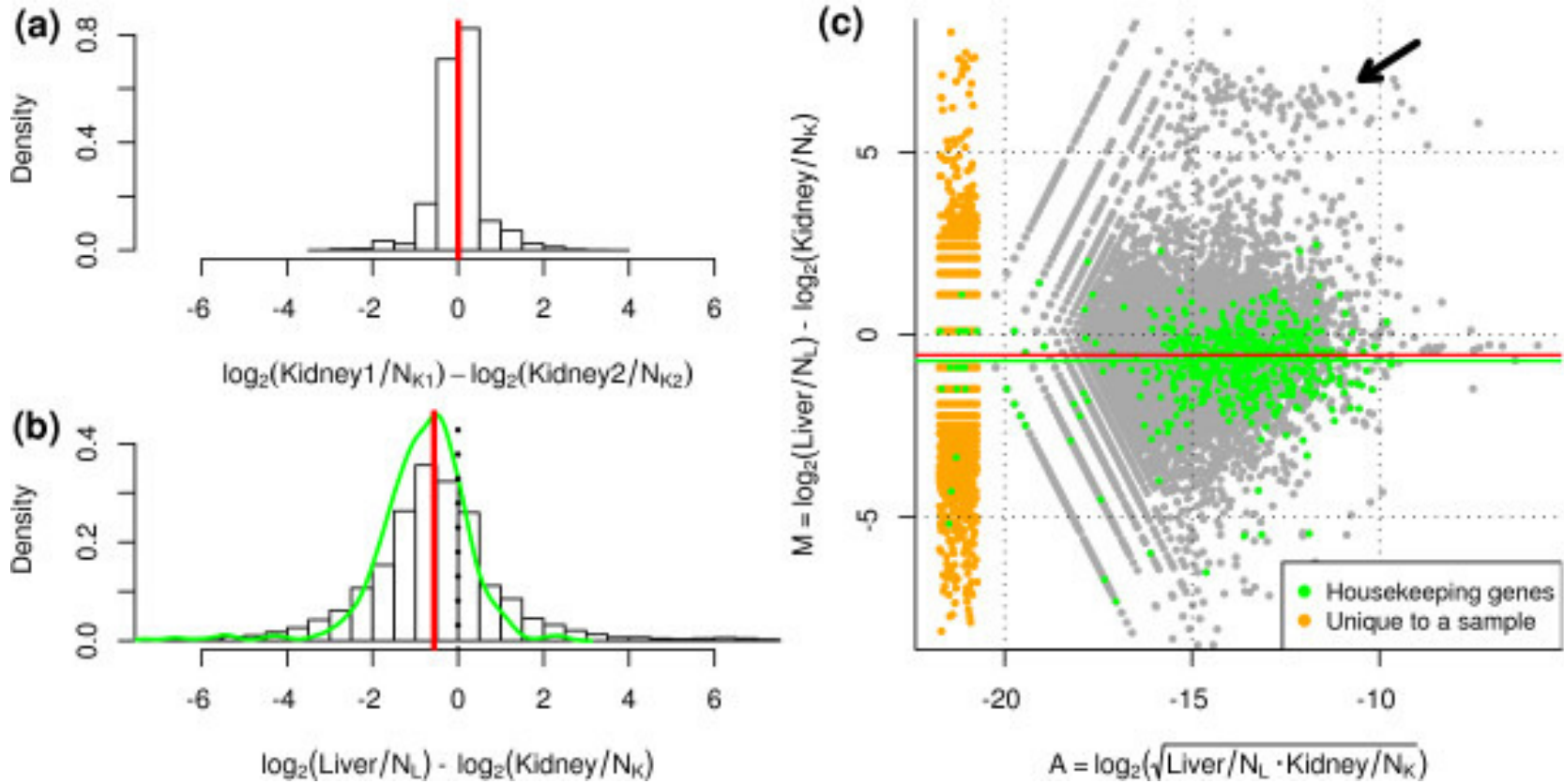
CPM (EdgeR)

Count **P**er **M**illion 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

Default

cuffdiff

❖ Upper-quantile normalization

- By read count of the gene at upper-quantile

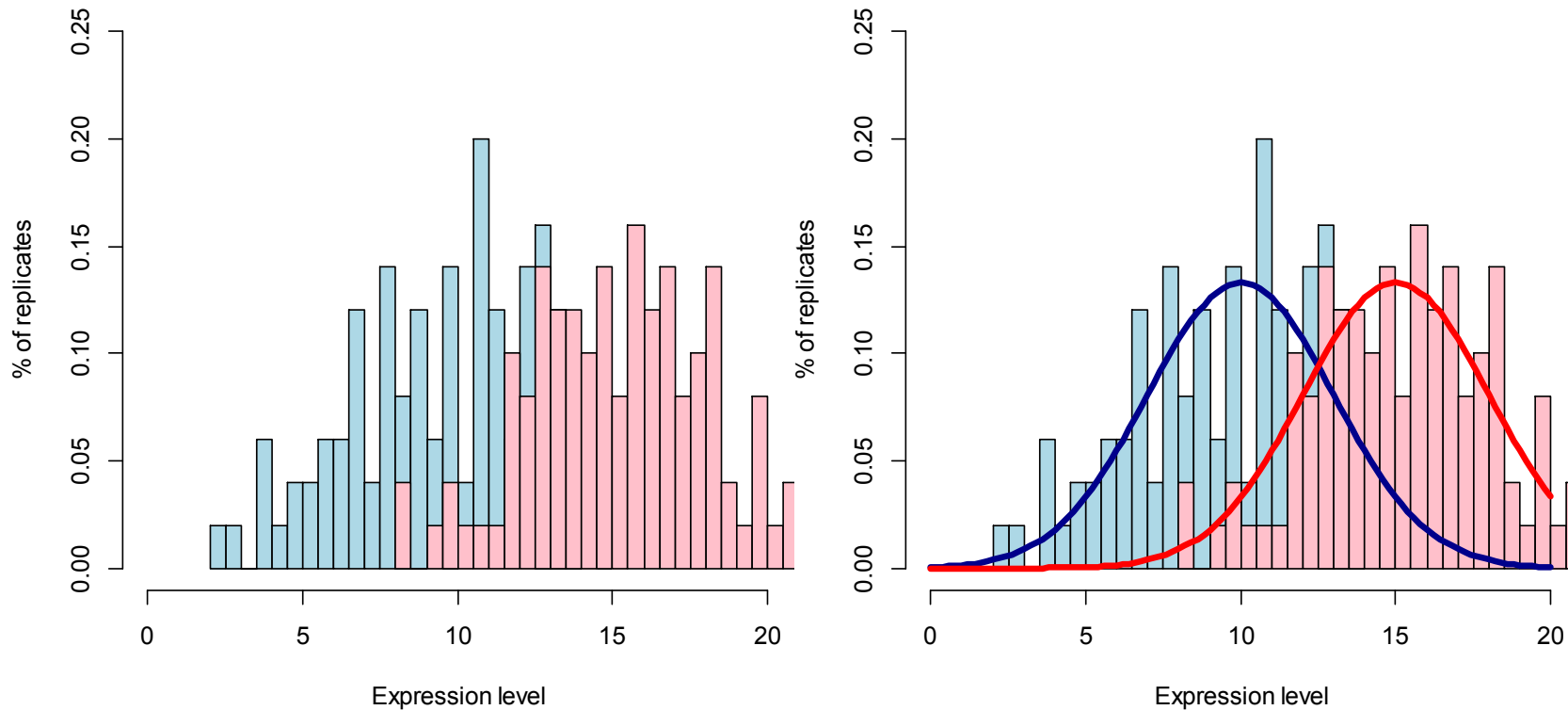
❖ Normalization by housekeeping genes

❖ Trimmed mean (TMM) normalization

EdgeR & DESeq

3. Differentially expressed genes

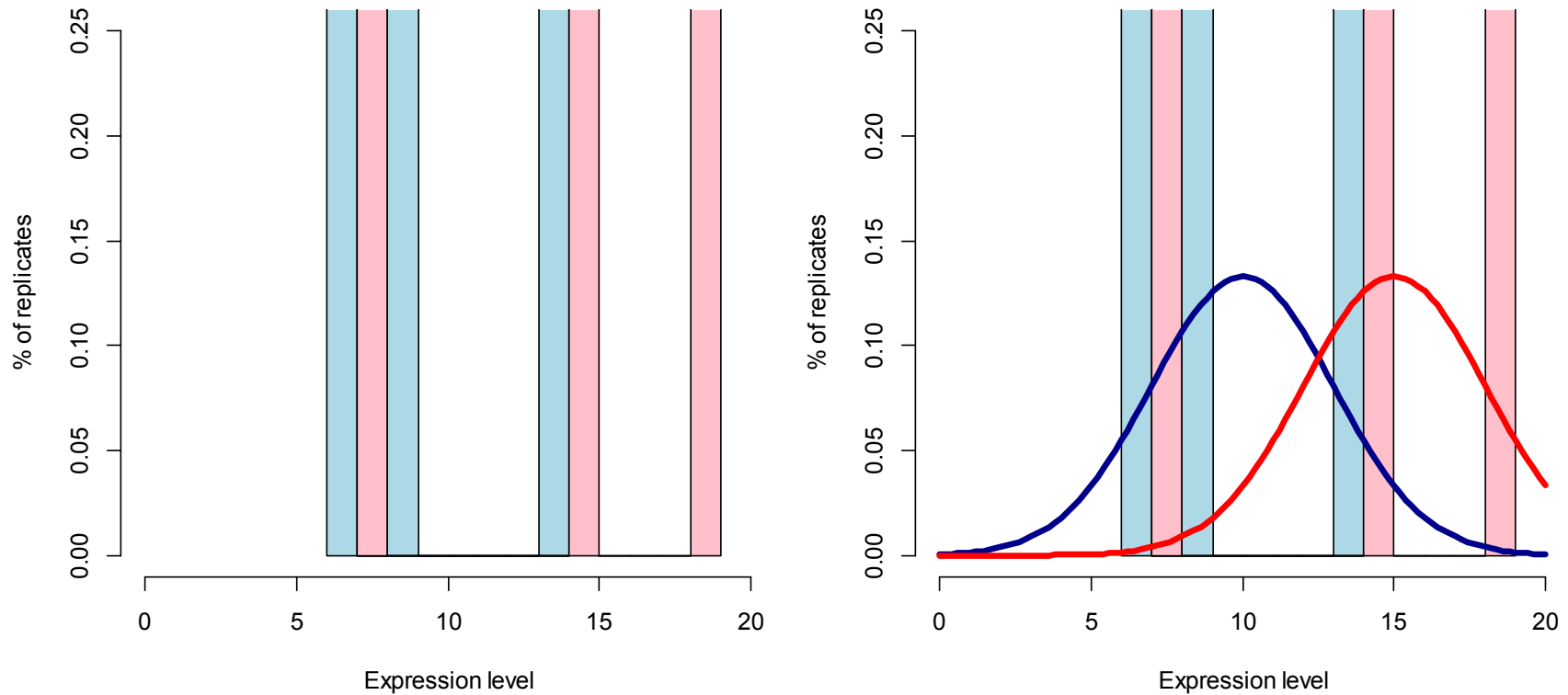
If we could do 100 biological replicates,



Distribution of Expression Level of A Gene



The reality is, we could only do 3 replicates,



Distribution of Expression Level of A Gene



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


Output from RNA-seq pipeline

For each gene:

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



**Using both fold change
and FDR value to filter:**



E.g. $\text{Log}_2(\text{fold}) > 1$ or < -1
&
 $\text{FDR} < 0.05$

Comparison of Methods

Table 2

Comparison of methods.

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

AUC, area under curve; DE, differential expression; FP, false positive.

Rapaport *et al. Genome Biology* 2013 **14**:R95 doi:10.1186/gb-2013-14-9-r95

Rapaport F et al.
Genome Biology,
2013 14:R95

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
```

Space

comma

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

tracking_id	condition	replicate	raw_frags	internal_s caled_frags	external_s caled_frags	FPKM	effective_ length	status
gene1	q1	0	16	11.3905	11.3905	0.305545	-	OK
gene1	q1	1	12	8.08334	8.08334	0.216832	-	OK
gene1	q2	0	15	26.084	26.084	0.699692	-	OK
gene1	q2	1	19	21.9805	21.9805	0.589617	-	OK
gene2	q1	0	61	43.4262	43.4262	4.50677	-	OK
gene2	q1	1	53	35.7014	35.7014	3.69312	-	OK
gene2	q2	0	35	60.8627	60.8627	6.35236	-	OK
gene2	q2	1	30	34.7061	34.7061	3.59016	-	OK

EdgeR input file:

Gene	A1	A2	B1	B2
gene1	16	12	15	19
gene2	61	53	35	30

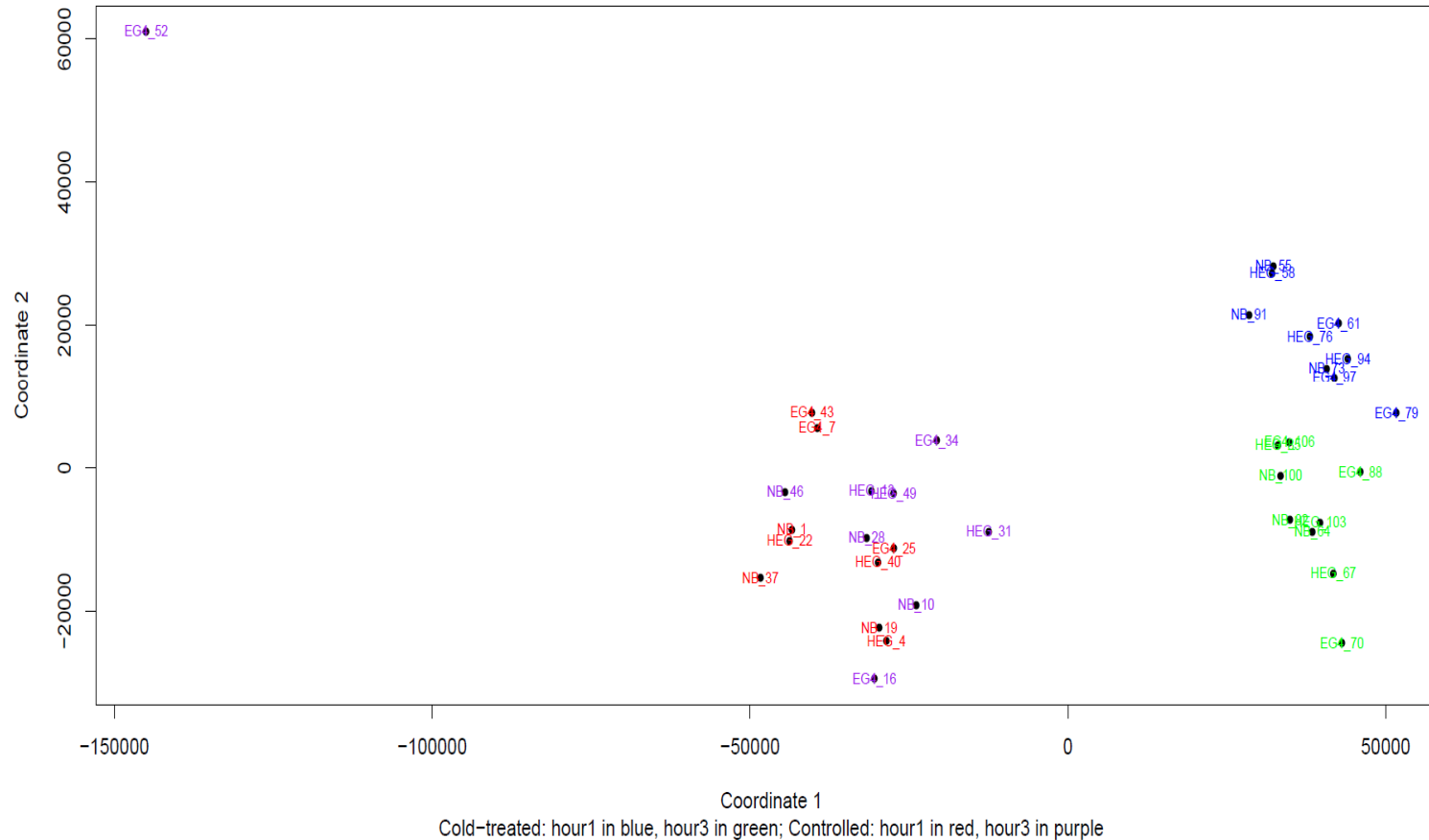
File conversion PERL script:

```
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

Metric MDS for Cold-treated vs Controlled Rice Samples



- Check reproducibility from replicates, remove outliers
- Check batch effects;

Use EdgeR to identify DE genes

	Treat	Time
Sample 1-3	Drug	0 hr
Sample 4-6	Drug	1 hr
Sample 7-9	Drug	2 hr

```
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

	Treat	Time
Sample 1-3	Placebo	0 hr
Sample 4-6	Placebo	1 hr
Sample 7-9	Placebo	2 hr
Sample 10-12	Drug	0 hr
Sample 13-15	Drug	1 hr
Sample 16-18	Drug	2 hr

```
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