Exercise 1 Review

Setting parameters

```
STAR --quantMode GeneCounts --genomeDir genomedb --runThreadN 2 --outFilterMismatchNmax 2 --readFilesIn WTa.fastq.gz --readFilesCommand zcat --outFileNamePrefix WTa --outFilterMultimapNmax 1 --outSAMtype BAM SortedByCoordinate
```

Some other parameters:

--outFilterMismatchNmax : max number of mismatch (Default 10)

--outReadsUnmapped fastx: output unmapped reads

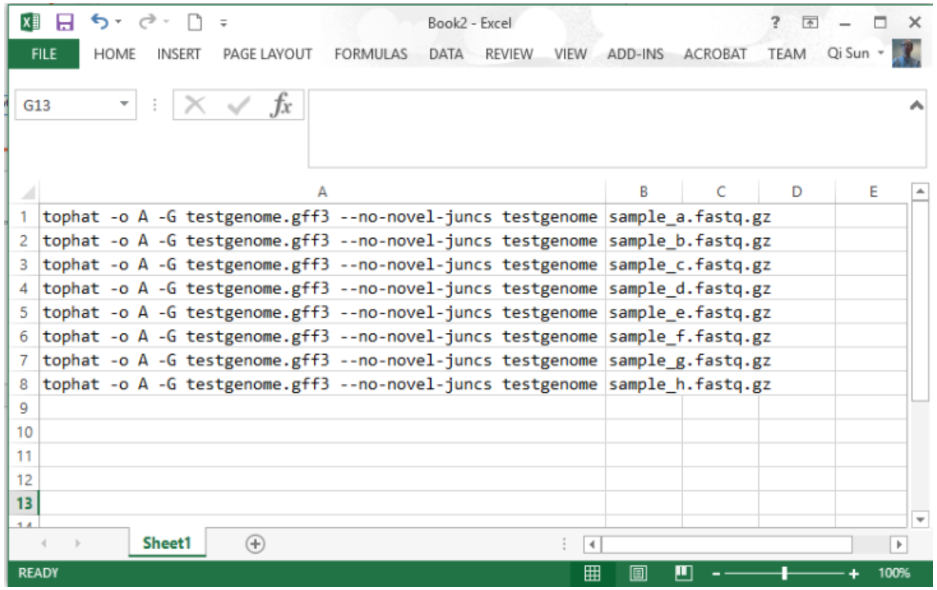
Manual:https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

Making Shell Script

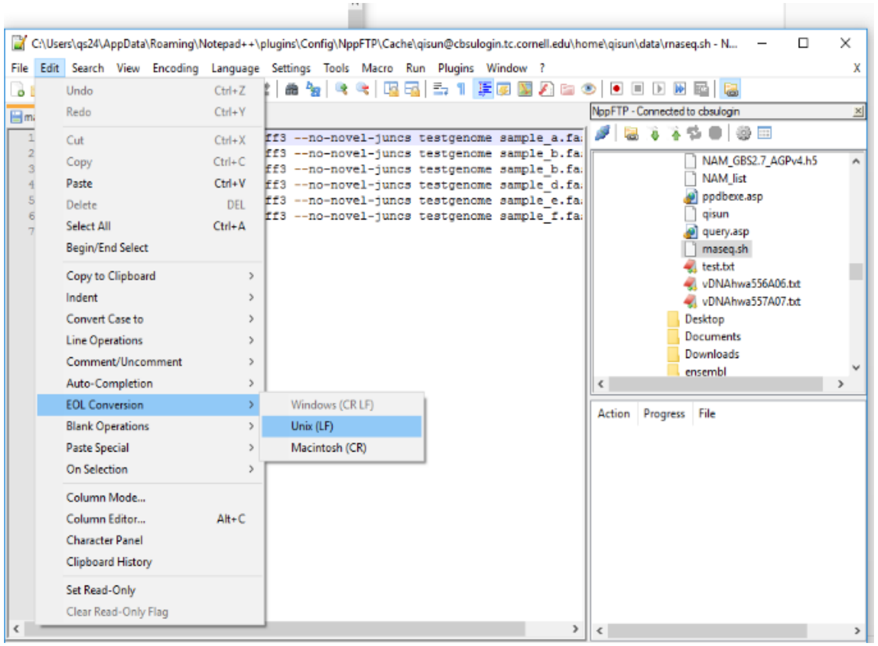
1. You can use Excel to make a shell script, and copy to the Notepad++/Text Wrangler.

2. Mac Excel user:
Make sure to use “mac2unix myfile” command to convert it to Linux file.

3. Windows user
Make sure to save as UNIX file in NotePad++. Or use the “dos2unix myfile” command to convert it to Linux file.



	A	B
1	tophat -o A -G testgenome.gff3 --no-novel-juncs testgenome	sample_a.fastq.gz
2	tophat -o A -G testgenome.gff3 --no-novel-juncs testgenome	sample_b.fastq.gz
3	tophat -o A -G testgenome.gff3 --no-novel-juncs testgenome	sample_c.fastq.gz
4	tophat -o A -G testgenome.gff3 --no-novel-juncs testgenome	sample_d.fastq.gz
5	tophat -o A -G testgenome.gff3 --no-novel-juncs testgenome	sample_e.fastq.gz
6	tophat -o A -G testgenome.gff3 --no-novel-juncs testgenome	sample_f.fastq.gz
7	tophat -o A -G testgenome.gff3 --no-novel-juncs testgenome	sample_g.fastq.gz
8	tophat -o A -G testgenome.gff3 --no-novel-juncs testgenome	sample_h.fastq.gz


C:\Users\qs24\AppData\Roaming\Notepad++\plugins\Config\NppFTP\Cache\qisun@cbsulogin.tc.cornell.edu\home\qisun\data\rnaseq.sh - N...

```
ff3 --no-novel-juncs testgenome sample_a.fa
ff3 --no-novel-juncs testgenome sample_b.fa
ff3 --no-novel-juncs testgenome sample_b.fa
ff3 --no-novel-juncs testgenome sample_d.fa
ff3 --no-novel-juncs testgenome sample_e.fa
ff3 --no-novel-juncs testgenome sample_f.fa
```

End of line in Text file

Linux: /n
Win: /r/n
Mac (9): /r
Mac (x): /n (Excel still use OS 9 style)

Running Shell Script

```
nohup sh /home/my_user_ID/runtophat.sh >& mylog &
```

Monitoring a job

```
top
top -o %MEM
ps -fu myUserID
ps –fu myUserID | grep STAR
```

Kill a job:

```
kill PID   ## you need to kill both shell script and STAR alignment that is still running
kill -9 PID
killall userID
```

Run multiple jobs:

```
nohup perl_fork_univ.pl script.sh 5 >& runlog &
```

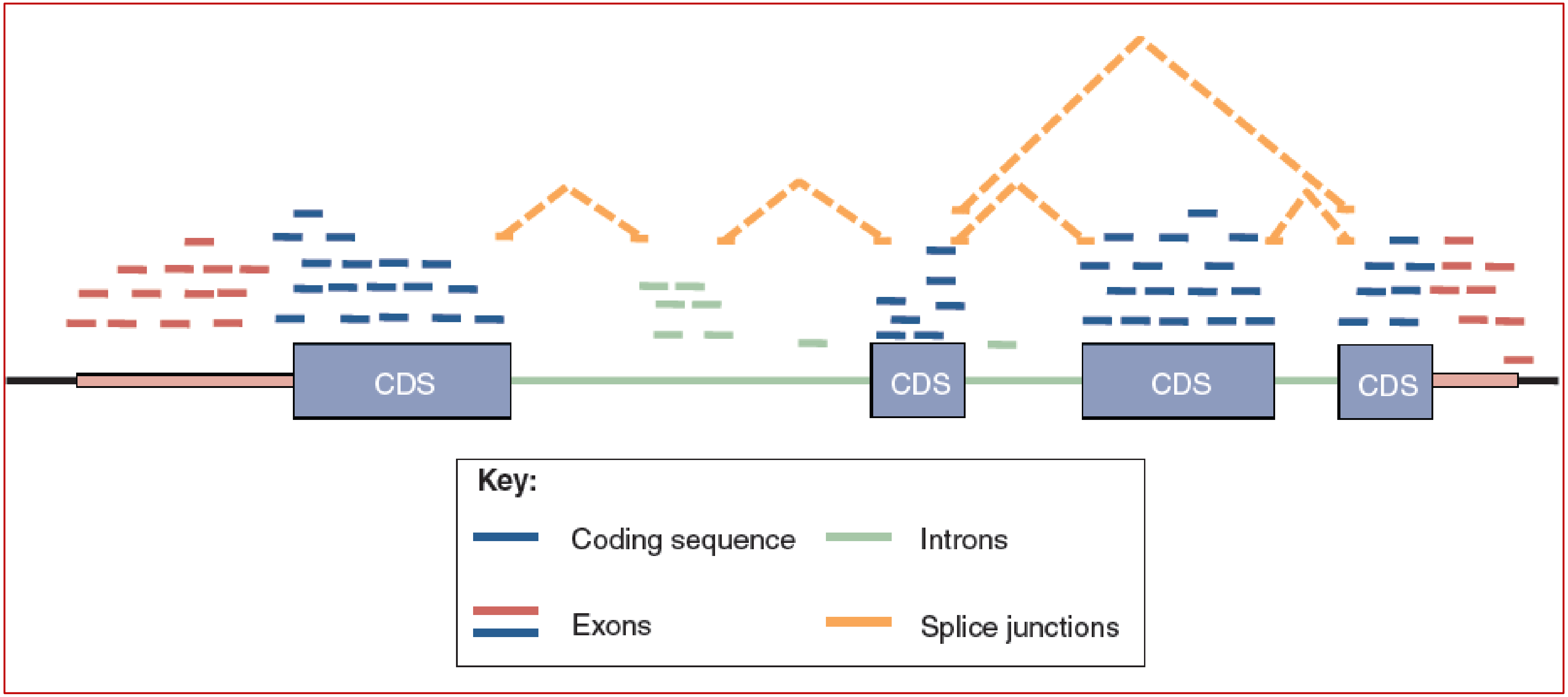
RNA-seq Data Analysis
Lecture 2

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

Multi-mapped reads

STAR and HTSeq

- Discard multi-mapped reads

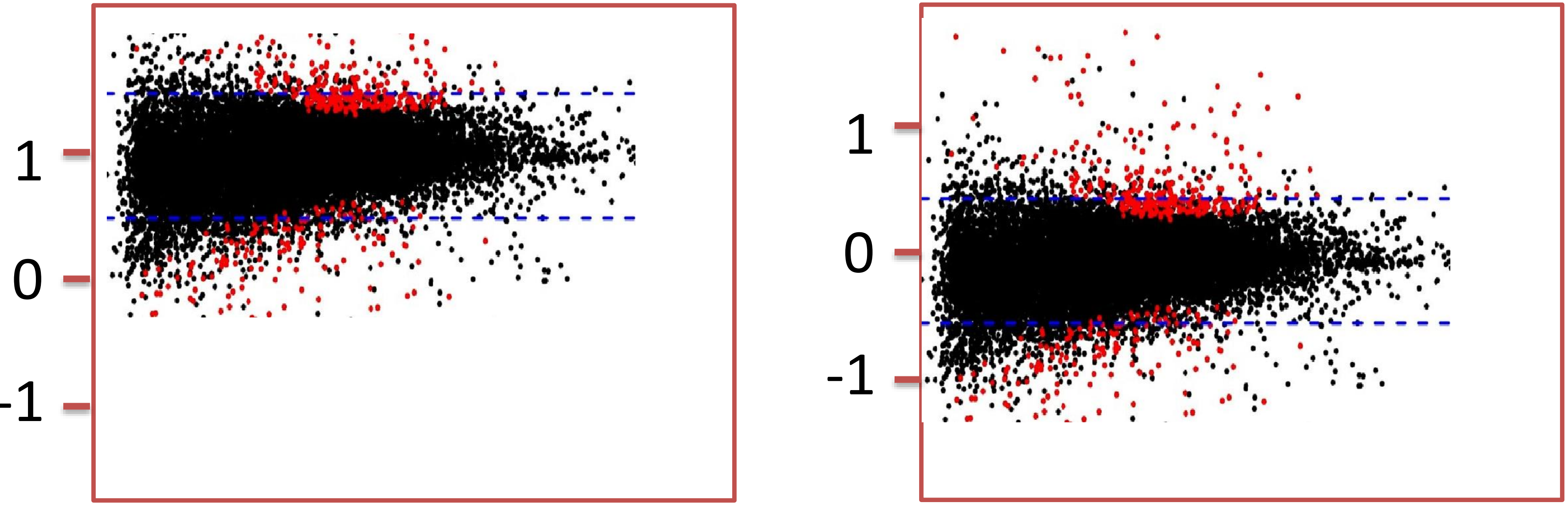
Cufflinks/Cuffdiff

- uniformly divide each read to all mapped positions

2. Normalization

Is necessary in RNAseq as total read counts are different in different samples

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:

Default: total reads from genes defined in GFF

-total-hits-norm: all aligned reads

CPM (EdgeR)

Count Per Million Reads

Normalization factor:

- total reads from genes defined in GFF
- Correction with TMM

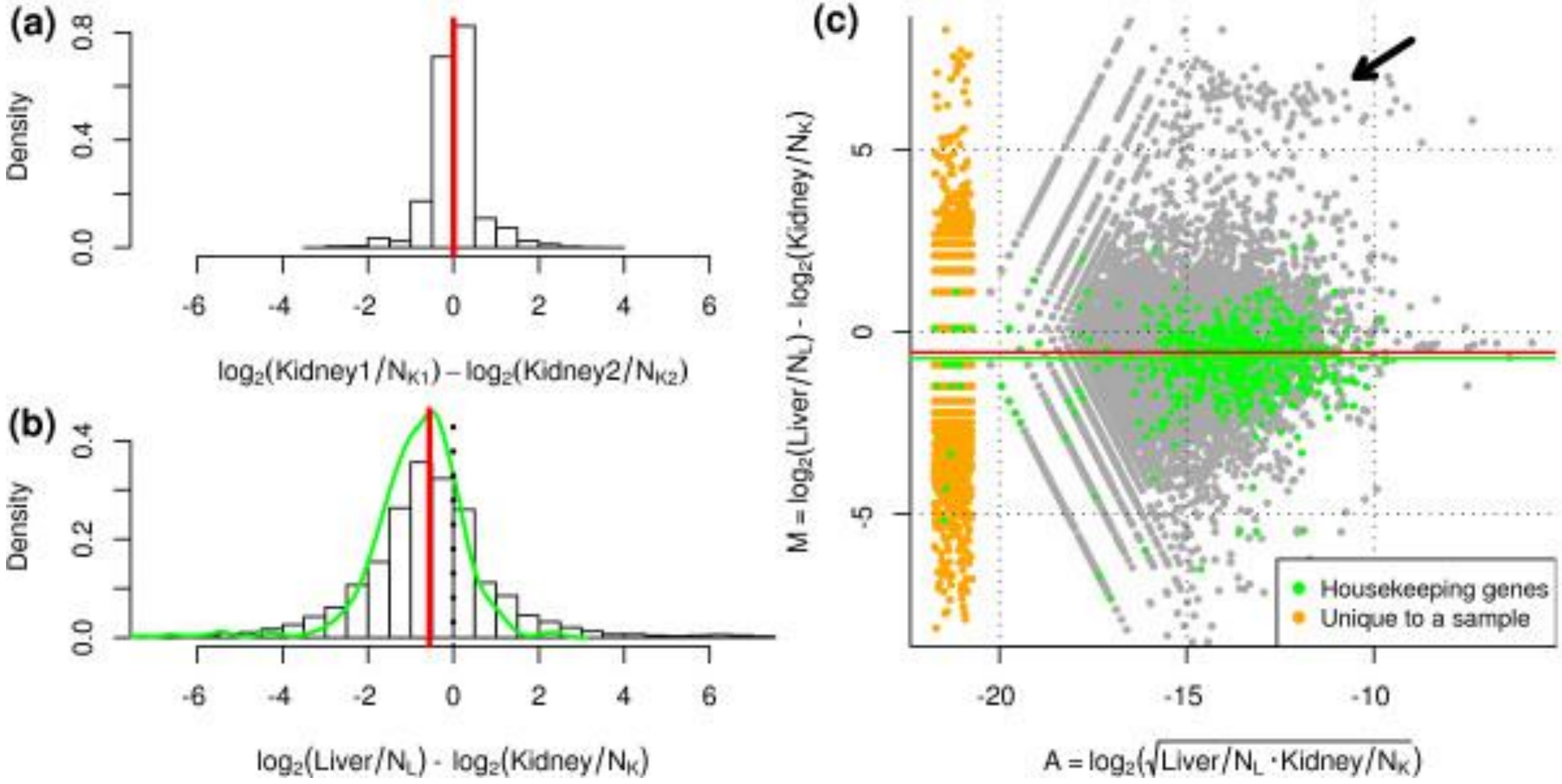
Reads that are not mapped to gene region (e.g. rRNA, pseudo-genes would not affect normalization

Factors that influence normalization:

Check these if you get weird results (e.g. poor correlation between replicates):

1. Make sure that rRNA are not annotated as genes in the GFF3/GTF file;
2. Manually check the top 10 genes in each sample, remove those highly expressed stress response genes, e.g. heatshock proteins;

Evaluate normalization with M-A plot



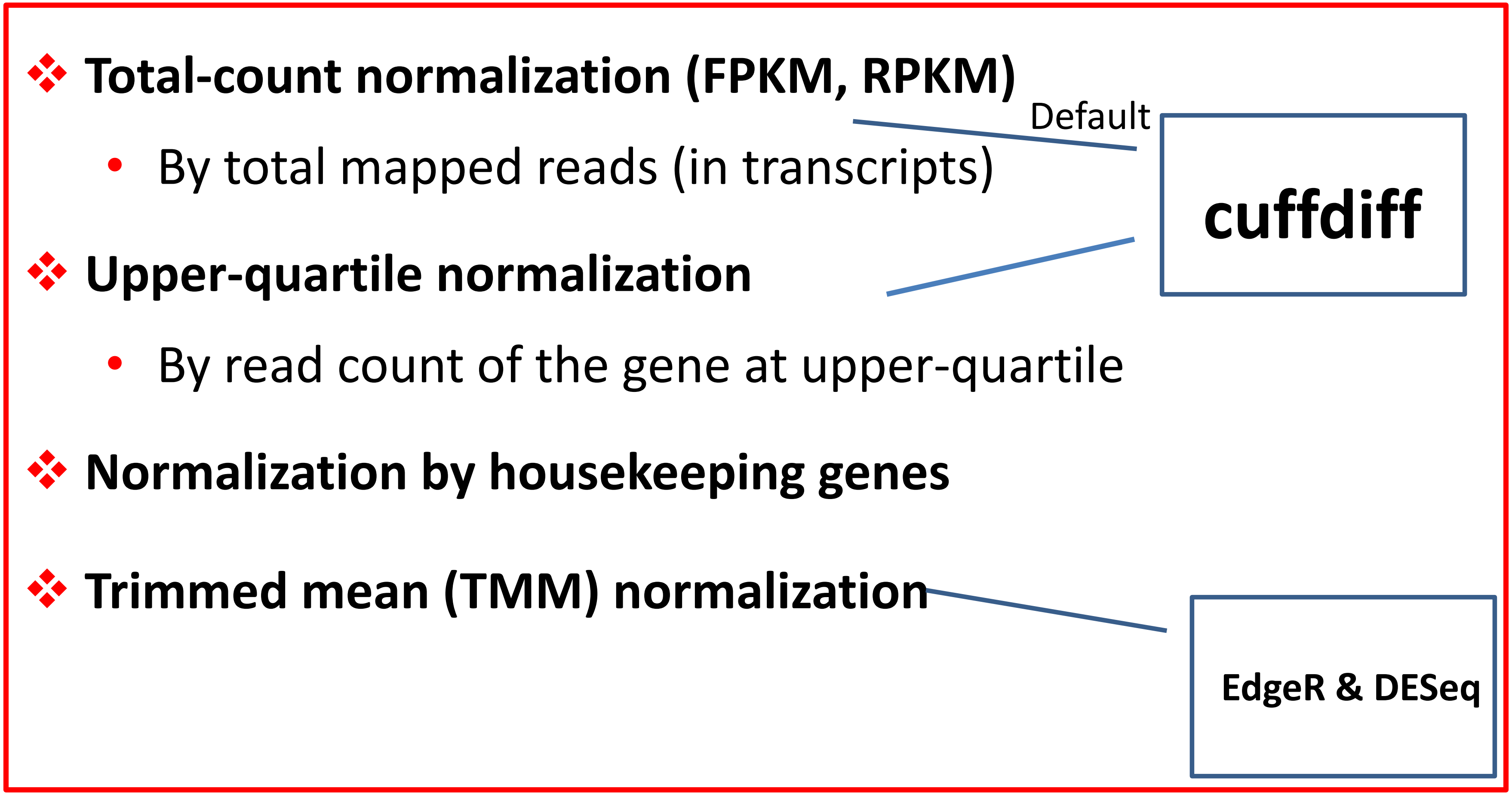
Default normalization in EdgeR: TMM

Robinson & Oshlack 2010 Genome Biology 2010, 11:R25.

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 (in transcripts)
- **Upper-quartile normalization**
  - By read count of the gene at upper-quartile
- **Normalization by housekeeping genes**
- **Trimmed mean (TMM) normalization**

3. Differentially expressed genes

Given a gene:

Read counts in control samples:

Repeat 1	**24**
Repeat 2	**25**
Repeat 3	**27**

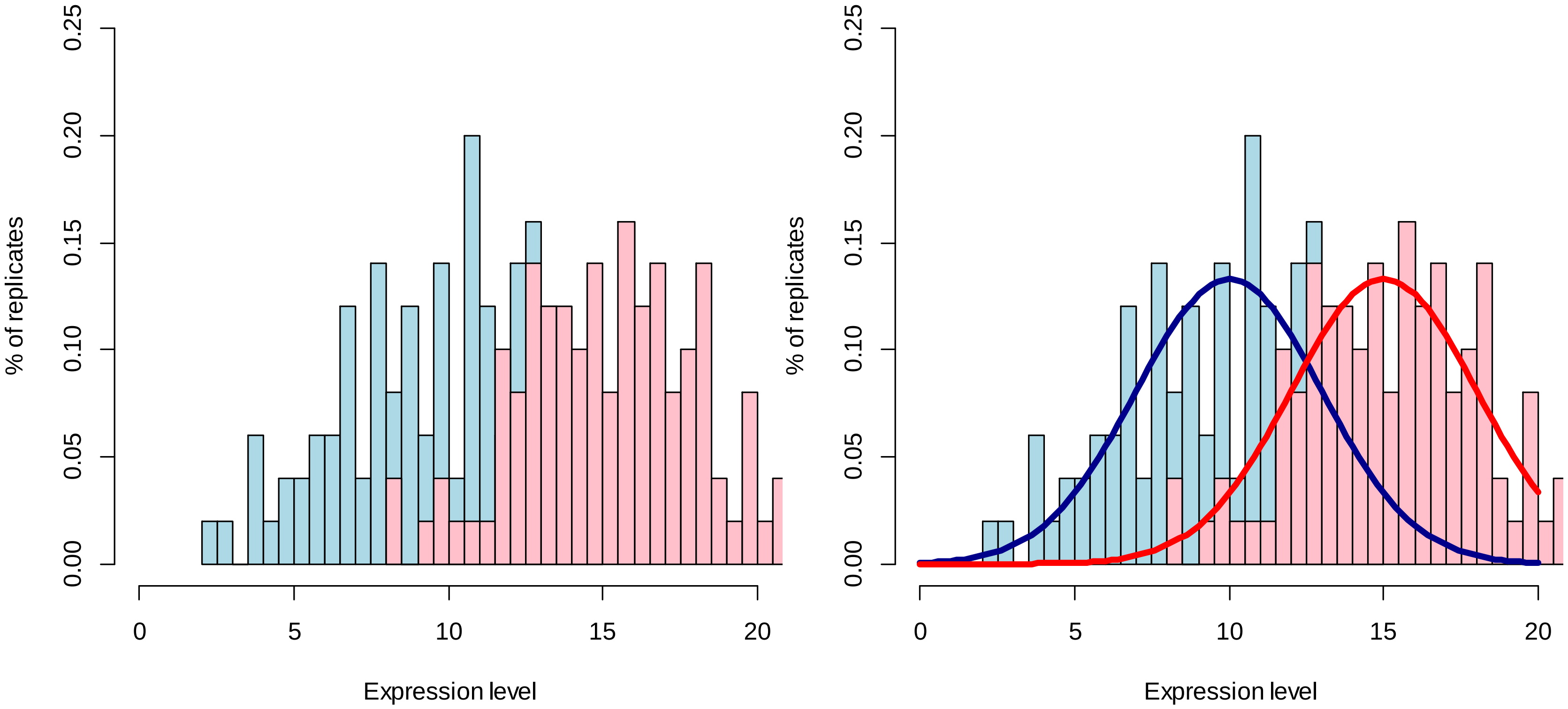
Read counts in treated samples:

Repeat 1	**23**
Repeat 2	**47**
Repeat 3	**29**

Different statistics model might give you different P or Q values.

3. Differentially expressed genes

If we could do 100 biological replicates,

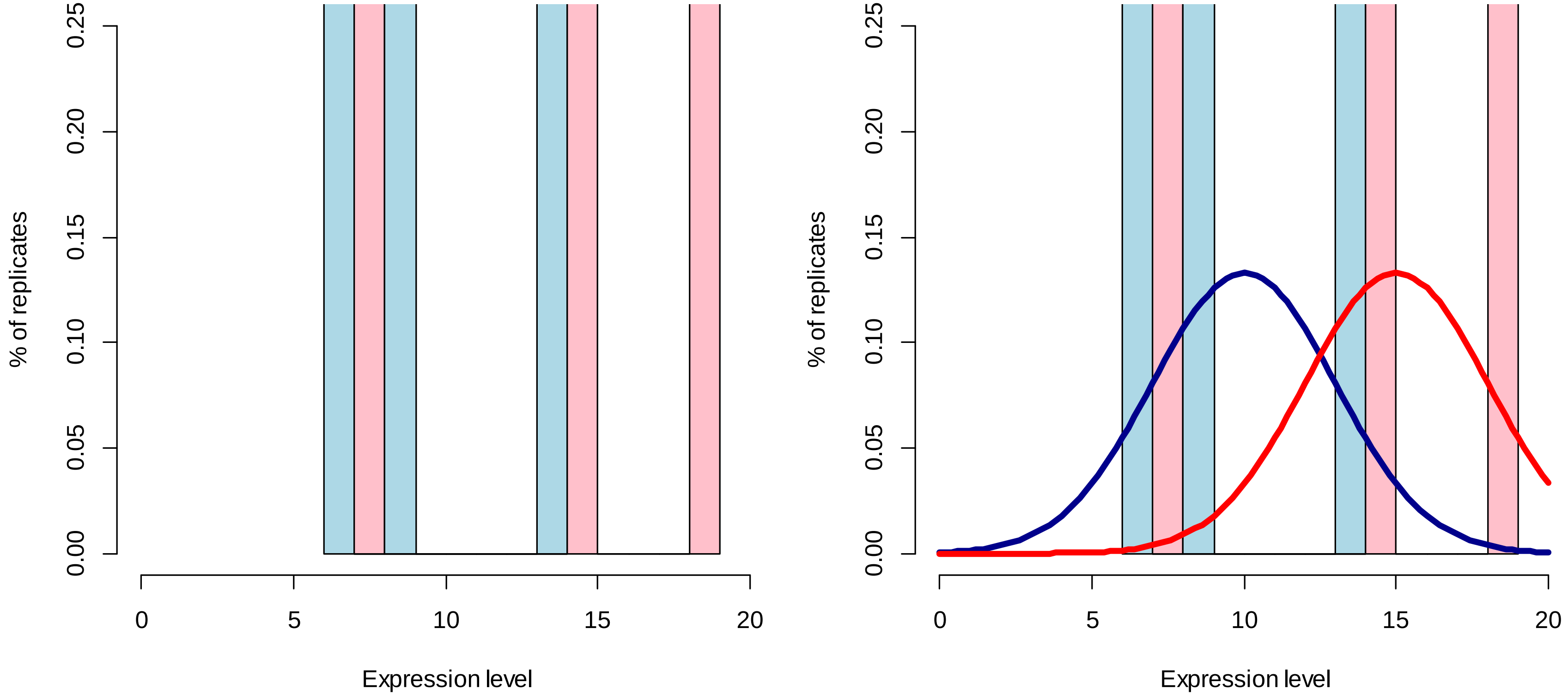


Distribution of Expression Level of A Gene

🟦 Condition 1

🟥 Condition 2

The reality is, we could only do 3 replicates,



Distribution of Expression Level of A Gene

🟦 Condition 1

🟥 Condition 2

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 bionomial distribution.

DESeq and EdgeR

3. Multiple test correction

Default in EdgeR: Benjamini-Hochberg

Output table from RNA-seq pipeline

Values for each gene:

- Read count (raw & normalized)
- Fold change (Log2 fold) between the two conditions
- P-value
- Q(FDR) value after multiple test.

Filter by:

a. fold change;

b. FDR value to filter;

c. Expression level.

E.g. Log2(fold) >1 or <-1

FDR < 0.05

Comparison of Methods

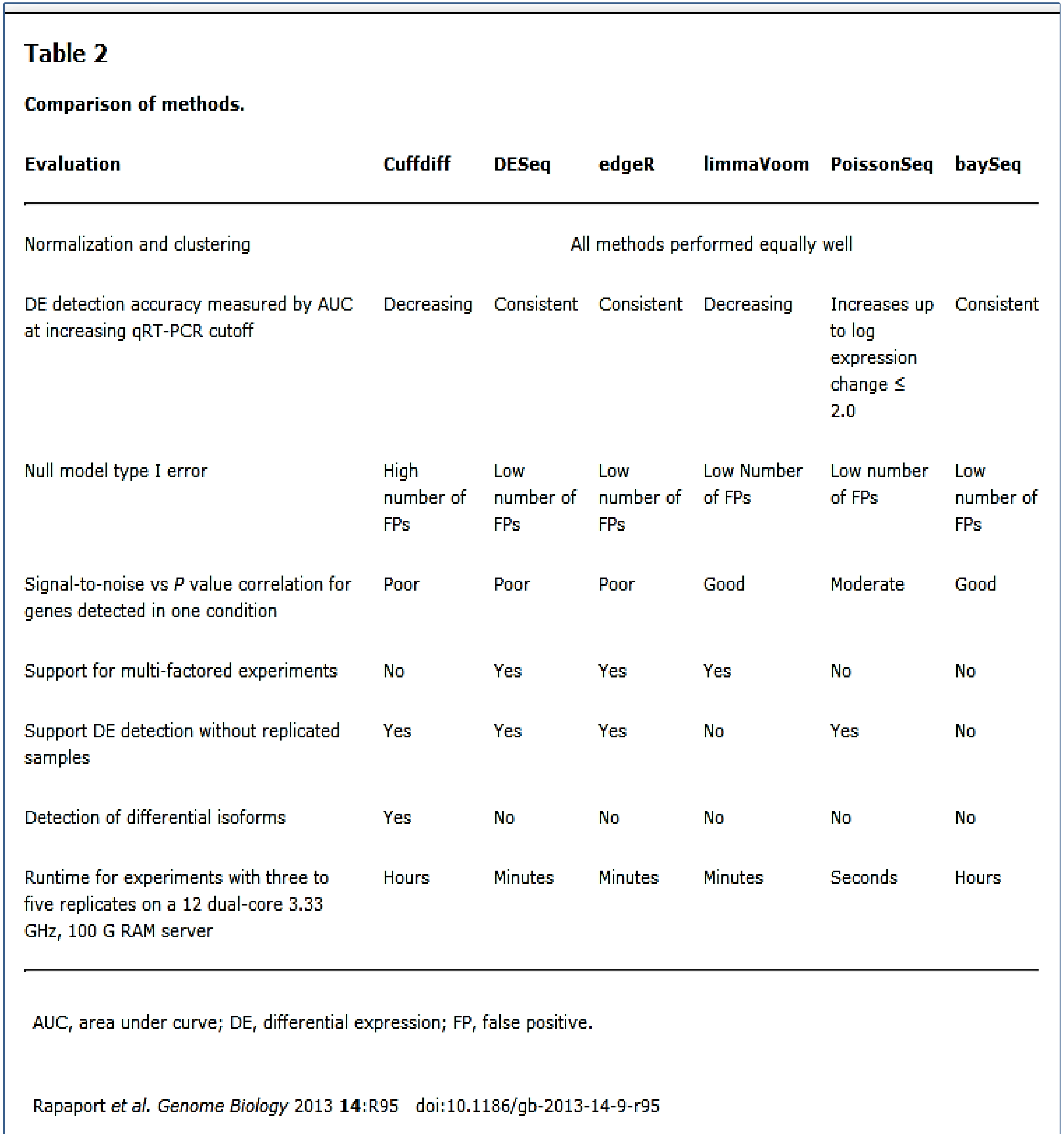


Table 2

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 *P* 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.

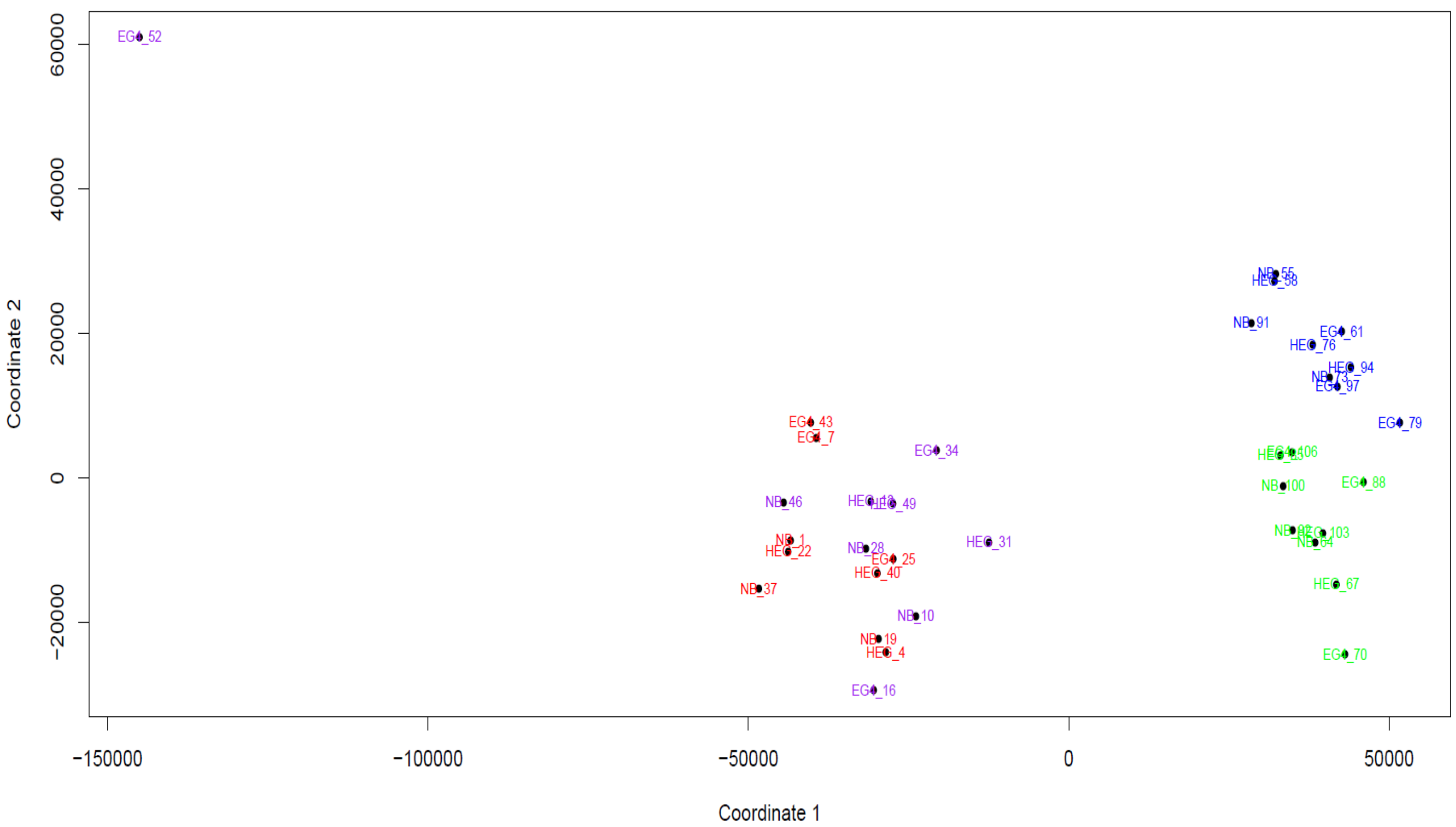
Rapaport *et al. Genome Biology* 2013 **14**:R95 doi:10.1186/gb-2013-14-9-r95

Can I trust P-value?
Can I trust Adjusted P-value?

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

Using EdgeR to make MDS plot of the samples

Metric MDS for Cold−treated vs Controlled Rice Samples



Cold−treated: hour1 in blue, hour3 in green; Controlled: hour1 in red, hour3 in purple

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

RNA-seq Pipelines at Bioinformatics Facility

Pipeline 1

STAR
-> DESeq or EdgeR

Pipeline 2

Tophat (alignment)
-> HTSeq or Cuffdiff (read count)
->DESeq or EdgeR

http://cbsu.tc.cornell.edu/lab/doc/rna_seq_draft_v8.pdf

Output files from STAR

*Log.final.out

```
                 Number of input reads |       13547152
             Average input read length |       49
                           UNIQUE READS:
          Uniquely mapped reads number |       12970876
               Uniquely mapped reads % |       95.75%
                 Average mapped length |       49.32
              Number of splices: Total |       1891468
   Number of splices: Annotated (sjdb) |       1882547
              Number of splices: GT/AG |       1873713
              Number of splices: GC/AG |       15843
              Number of splices: AT/AC |       943
      Number of splices: Non-canonical |       969
```

*ReadsPerGene.out.tab

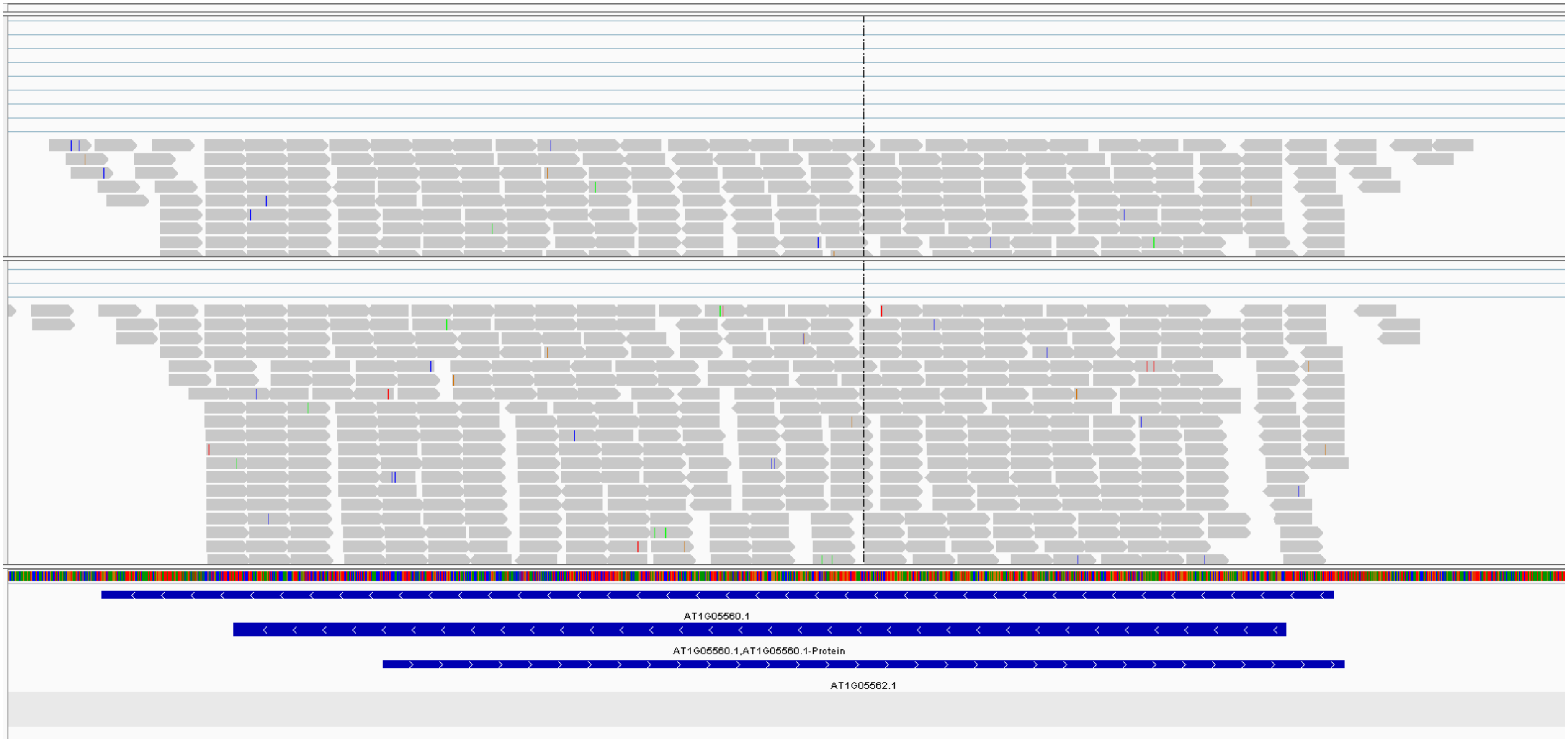
N_unmapped	1860780	1860780	1860780
N_multimapping	0	0	0
N_noFeature	258263	13241682	375703
N_ambiguous	461631	9210	17159
gene:AT1G01010	50	1	49
gene:AT1G01020	149	1	148
gene:AT1G03987	0	0	0
gene:AT1G01030	77	0	77
gene:AT1G01040	583	41	669
…			

*ReadsPerGene.out.tab

	Unstranded	Forward	Reverse
gene:AT1G05560	0	210	476
gene:AT1G05562	0	476	210

The two genes are on opposite strand (AT1G05562 is ncRNA)

*Aligned.sortedByCoord.out.bam



Connection between software

STAR output:

Sample1

N_unmapped	1860780	1860780	1860780
N_multimapping	0	0	0
N_noFeature	258263	13241682	375703
N_ambiguous	461631	9210	17159
gene:AT1G01010	50	1	49
gene:AT1G01020	149	1	148
gene:AT1G03987	0	0	0
gene:AT1G01030	77	0	77
gene:AT1G01040	583	41	669
…			

Sample2

N_unmapped	1637879	1637879	1637879
N_multimapping	0	0	0
N_noFeature	224759	11828019	354396
N_ambiguous	445882	8133	14924
gene:AT1G01010	57	0	57
gene:AT1G01020	174	2	172
gene:AT1G03987	1	1	0
gene:AT1G01030	91	3	88
gene:AT1G01040	516	27	594
gene:AT1G03993	0	81	2

EdgeR input:

gene	Sample1	Sample2	Sample3	Sample4
AT1G01010	57	49	36	40
AT1G01020	172	148	197	187
AT1G03987	0	0	0	0
AT1G01030	88	77	74	101
AT1G01040	594	669	504	633
AT1G03993	2	1	0	0
…	…	…	…	…

```
paste file1 file2 file3 file4 | \
cut -f1,4,8,12,16 | \
tail -n +5 \
> tmpfile

cat tmpfile | \
sed "s/^gene://" \
>gene_count.txt
```

Connection between software

Reading file into R

AT1G01010	57	49	36	40
AT1G01020	172	148	197	187
AT1G03987	0	0	0	0
AT1G01030	88	77	74	101
AT1G01040	594	669	504	633
AT1G03993	2	1	0	0
…	…	…	…	…

```
x <- read.delim("gene_count.txt", header=F, row.names=1)

colnames(x)<-c("WTa","WTb","MUa","MUb")
```

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

Normalization and Remove genes that are not expressed

```
library("edgeR")

group <- factor(c(1,1,2,2))

y <- DGEList(counts=x,group=group)

y <- calcNormFactors(y)

keep <-rowSums(cpm(y)>=1) >=2   # remove un-expressed genes

y<-y[keep,]
```

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

Fit the model:

```
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 – Placbo.0hr) – (Drug.2hr-
Drug.1hr)
```

Exercise

- Using STAR for read alignment and quantification and identifying differentially expressed genes of two different biological conditions WT and MU. There are two replicates (a, b) for each condition.

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