edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Robinson MD, McCarthy DJ and Smyth GK (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” *Bioinformatics*, 26, pp. -1.

http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

Factors Affecting Differential Expression Analysis

- Sequencing depth
- Replication number

  - Biological variation

  -Technical variation (library preparation, sequencing)

  - At least 3 biological replicates (still quite small for statistical testing)

Software tools for differential expression analysis

- DeSeq, DeSeq2, edgeR, tweeDESeq, Limma, SAMSeq, NOISeq, CuffDiff, BitSeq, efSeq

- The right software choice depends on objectives of your experiment and data set.

  (type of experimental design, number of replicates, genes or isoforms)

- edgeR can identify more true positive

General pipeline for differential expression analysis

(edgeR uses negative binomial distribution)

- Filtering
- Normalization
- Dispersion estimation
- Hypothesis testing

Datasets - immature fruit (2 biological replicates)
SRR404331
SRR404333

- breaker fruit (2 biological replicates)
SRR404334
SRR404336

edgeR

- **Input**
  A table of integer read counts:
    Rows: genes
    Columns: independent libraries.

```
bioinfo@biodebian:~$ cd Desktop/Slch04_demo/
bioinfo@biodebian:~/Desktop/Slch04_demo$ ls
                                - gene_count_matrix.csv
```

bioinfo@biodebian:~/Desktop/Slch04_demo$ less gene_count_matrix.csv

,SRR404331_ch4.sort,SRR404333_ch4.sort,SRR404334_ch4.sort,SRR404336_ch4.sort

gene:Solyc02g032840.1,0,0,0,0

gene:Solyc04g050480.3,59,57,38,69

gene:Solyc04g080270.3,1050,1388,869,1138

gene:Solyc04g071590.3,0,27,0,7

gene:Solyc04g079110.1,1,2,1,1

gene:Solyc04g081870.3,12,10,0,211

gene:Solyc04g074080.3,0,107,29,17

gene:Solyc04g071260.3,0,0,8,14

Rstudio

- edgeR_bioinfo.R
- Loading edgeR : library(“edgeR”)
- Setting working directory:
  setwd("/home/bioinfo/Desktop/Slch04_demo")
- Importing data into R:
  x <- read.csv("gene_count_matrix.csv", row.names=1, header=TRUE)
- Grouping:
  group <- (c(1,1,2,2))

edgeR stores RNA-seq data as a DGEList object
(data structure = list)

- **Putting data into a DGEList object:**
  y <- DGEList(counts=x, group=group)

- **Checking the DGEList**
  y
  $counts: gene names, gene counts for each library

  $samples: library names, groups, lib.size, norm.factors

> y

An object of class "DGEList"

$counts

	SRR404331_ch4.sort	SRR404333_ch4.sort	SRR404334_ch4.sort	SRR404336_ch4.sort
gene:Solyc10g054820.2	0	0	0	0
gene:Solyc12g098195.1	0	0	0	0
gene:Solyc10g046810.1	0	0	0	0
gene:Solyc02g062000.3	0	0	0	0
gene:Solyc07g019650.3	0	0	0	0

34874 more rows ...

$samples

	group	lib.size	norm.factors
SRR404331_ch4.sort	1	468307	1
SRR404333_ch4.sort	1	398150	1
SRR404334_ch4.sort	2	304826	1
SRR404336_ch4.sort	2	494570	1

- **head(y$counts)**

	SRR404331_ch4.sort	SRR404333_ch4.sort	SRR404334_ch4.sort	SRR404336_ch4.sort
gene:Solyc10g054820.2	0	0	0	0
gene:Solyc12g098195.1	0	0	0	0
gene:Solyc10g046810.1	0	0	0	0
gene:Solyc02g062000.3	0	0	0	0
gene:Solyc07g019650.3	0	0	0	0
gene:Solyc01g013760.1	0	0	0	0

- **dim(y$counts)**

34879 4

- **y$samples**

	group	lib.size	norm.factors
SRR404331_ch4.sort	1	468307	1
SRR404333_ch4.sort	1	398150	1
SRR404334_ch4.sort	2	304826	1
SRR404336_ch4.sort	2	494570	1

Data processing

- **Saving a copy of raw data before data processing**

  y.rawdata <-y

- **Filtering to remove very low counts:**

  If you want to keep genes with more than 1 CPM, in at least 2 samples

  keep <- rowSums(cpm(y)>1) >= 2

  (TRUE=1, FALSE=0)

  (Question: Is 1 count per million suitable for our data?)

- **Checking filtering step**

  table(keep)

  FALSE TRUE

  33286 1593

- **Modifying DGEList**

  y <- y[keep, , keep.lib.sizes=FALSE]

- **Checking filtered data set**

  dim(y)

  1593 4

  head(y$counts)

	SRR404331_ch4.sort	SRR404333_ch4.sort	SRR404334_ch4.sort	SRR404336_ch4.sort
gene:Solyc04g050480.3	59	57	38	69
gene:Solyc04g080270.3	1050	1388	869	1138
gene:Solyc04g008310.2	480	256	111	179
gene:Solyc04g074240.3	657	526	373	570
gene:Solyc04g071590.3	0	27	0	7
gene:Solyc04g054840.1	3	0	1	1

- **y$samples**

	group	lib.size	norm.factors	
SRR404331_ch4.sort	1	468168	1	(468307)
SRR404333_ch4.sort	1	398023	1	(398150)
SRR404334_ch4.sort	2	304799	1	(304826)
SRR404336_ch4.sort	2	494176	1	(494570)

Normalization

- edgeR

  - uses TMM (Trimmed mean of M-value) method to eliminate RNA composition effect

  - automatically adjusts for difference in library size caused by sequencing depth

  - doesn't adjust for gene length

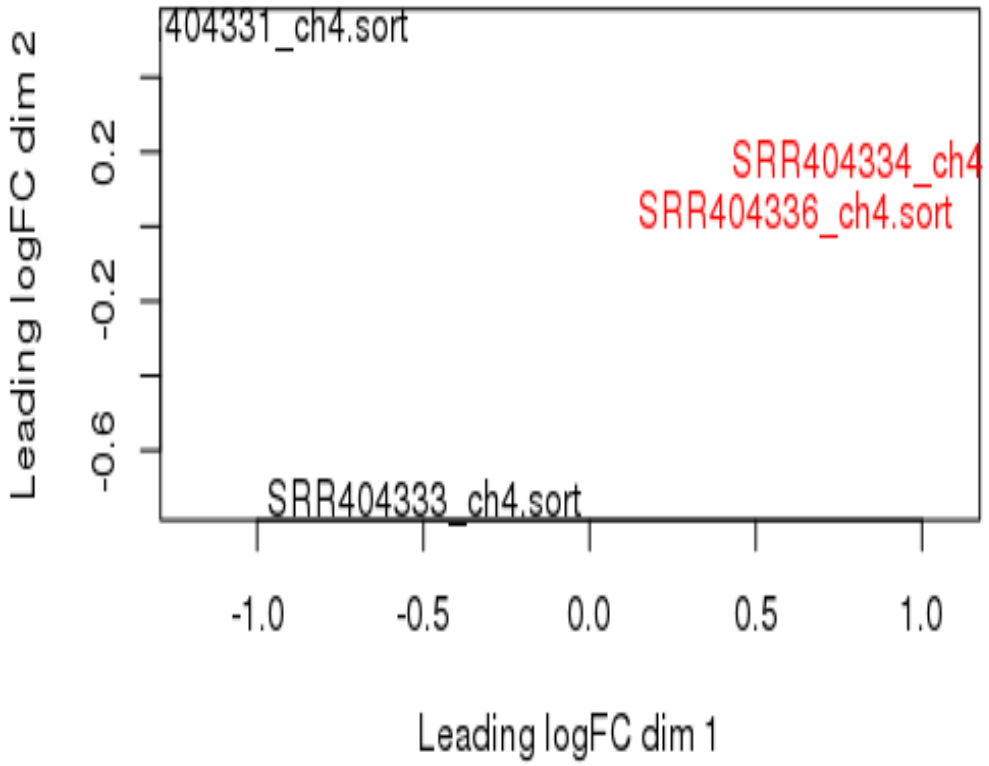
Normalization

- y <- calcNormFactors(y)
- y$samples

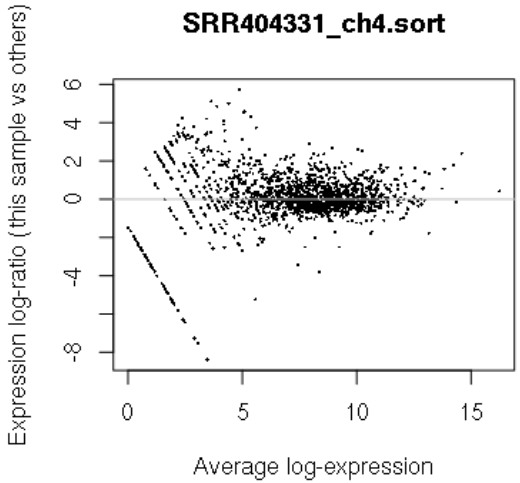
	group	lib.size	norm.factors
SRR404331_ch4.sort	1	468168	0.8929376
SRR404333_ch4.sort	1	398023	1.0082953
SRR404334_ch4.sort	2	304799	1.0647144
SRR404336_ch4.sort	2	494176	1.0431769

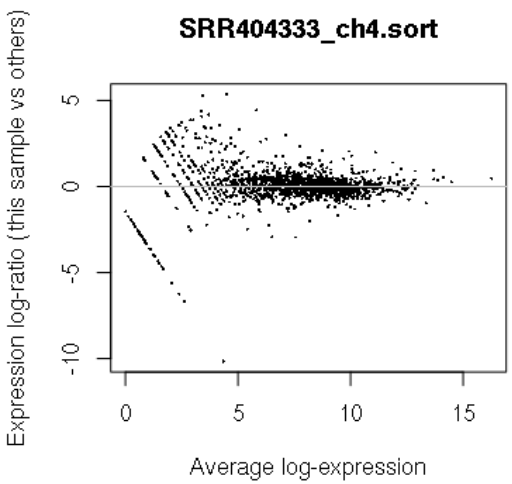
Data Exploration

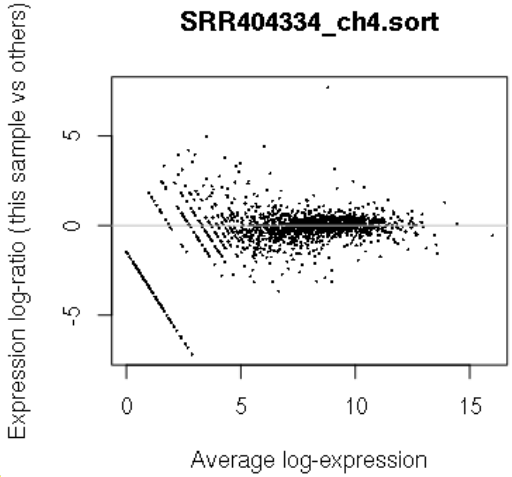
Multi-dimensional (MDS) plot: plotMDS (y)

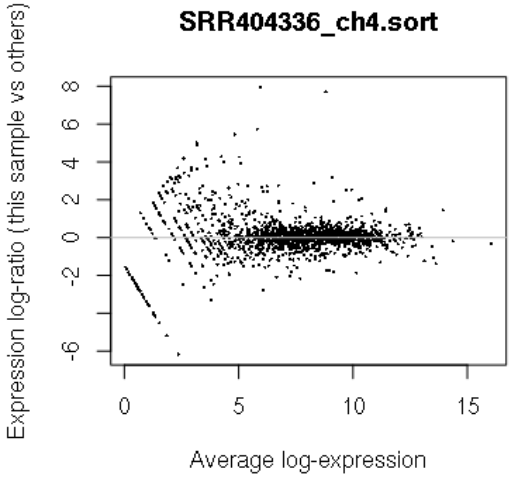


Data Exploration
Mean-different (MD) plot: plotMD (y)









General pipeline

- Filtering
- Normalization
- Dispersion estimation
- Hypothesis testing

edgeR package

- **Classic edgeR: testing single factor**

  -Exact test

- **Generalized linear models (glms): testing multiple factors**

  -Likelihood ratio test
  -Quasi-likelihood method

Pairwise comparison (classic edgeR)

- **Dispersion estimation**

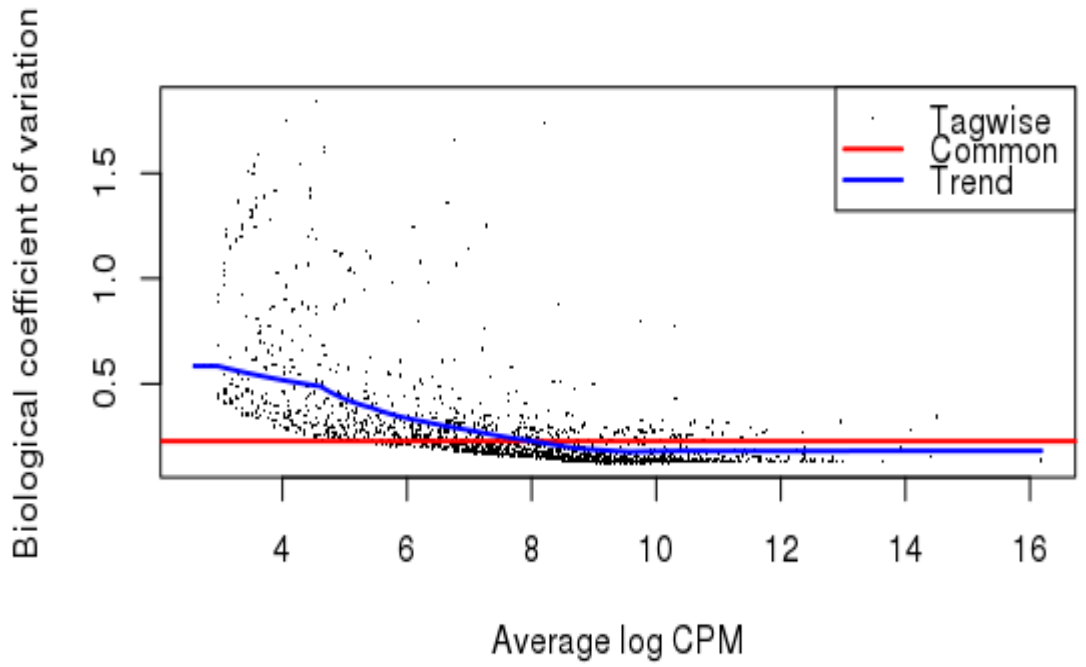
```
y <- estimateDisp(y)
```

```
y <- estimateCommonDisp(y)
y <- estimateTagwiseDisp(y)
```

y$tagwise.dispersion

[1] 0.05521917 0.03610428 0.04111724 0.02817601 1.00101862 0.57733575
[7] 0.33174515 0.02705387 0.02491446 0.04149004 0.68987488 0.02988139
.....**1593**

plotBCV(y)



Differential expression analysis

- **Hypothesis testing**
  de <- exactTest(y)

- **To display the most significant tags**
  topTags(de, n=10)

Comparison of groups: 2-1

	logFC	logCPM	PValue	FDR
gene:Solyc04g074840.3	10.674057	11.661710	5.008535e-101	7.978595e-98
gene:Solyc04g079960.1	4.780808	9.749903	1.003905e-52	7.996107e-50
gene:Solyc04g078460.3	3.010089	9.361145	4.374771e-27	2.323003e-24
gene:Solyc04g076780.3	-3.572084	9.164866	7.345144e-27	2.925203e-24
gene:Solyc04g009960.3	3.129447	10.605411	1.079804e-25	3.440257e-23
gene:Solyc04g071615.1	-2.843780	11.217141	9.526994e-25	2.529417e-22
gene:Solyc04g071650.3	-3.425306	10.313879	2.530218e-24	5.758053e-22
gene:Solyc04g081300.3	-3.222494	10.382392	4.571451e-24	9.102901e-22
gene:Solyc04g079900.3	-2.638129	9.831421	1.733762-23	3.068758e-21
gene:Solyc04g079560.3	-3.836033	8.269721	7.778344e-23	1.239090e-20

- **Selecting differentially expressed genes at a FDR of 5%**

```
de_05 <- decideTestsDGE(de)

de_05
   TestResults matrix
   0  0 -1  0  0
   1588 more rows ...


summary(de_05)
   -1   142      (down-regulated)
    0  1331
    1   120      (up-regulated)
```

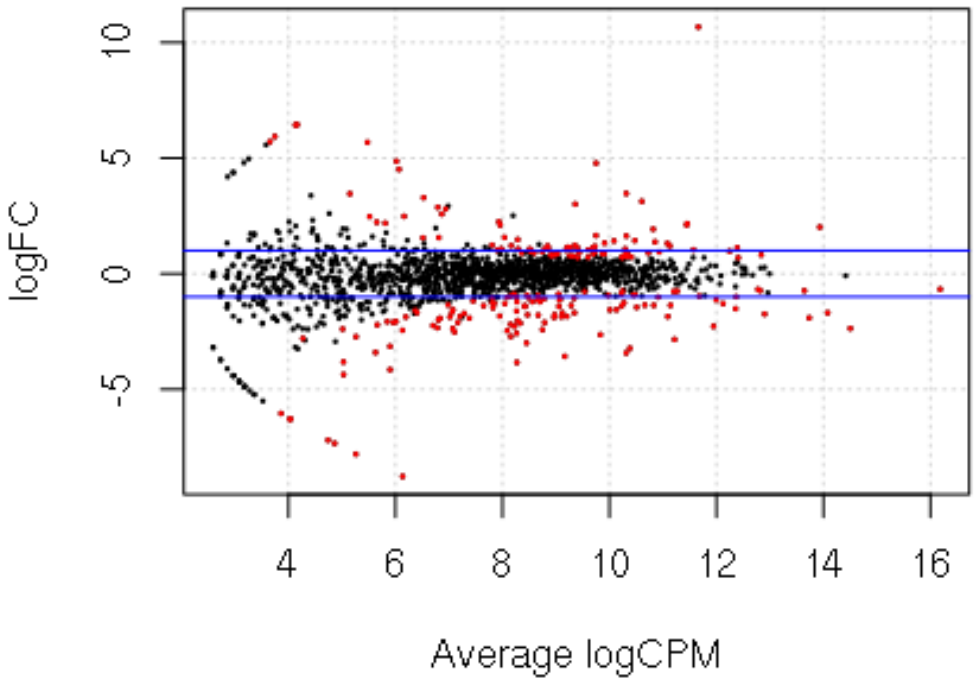
- **Generating a dataframe containing DE genes at a FDR at 5%**

```
isDE <- as.logical(de_05)

de_05name <-rownames(y)[isDE]

de_05.table <- de[de_05name, ]
```

- **Exporting data**

```
write.csv(de_05.table$table, file="de_05")

write.csv(de$table, file="de")
```

Smear Plot



The blue lines indicate 2 fold-changes

Exercise

- **How does parameter setting affect the number of differentially expressed genes**

  - Keeping genes with at least 20 CPM during filtering step.