

Differential Gene Expression 3 – quantifying differences with Bioconductor

Biol4230 Tues, April 4, 2017

Bill Pearson wrp@virginia.edu 4-2818 Jordan 6-057

- Bioconductor: a comprehensive 'R' package for expression and genome analysis
 - Obtaining/installing
 - Datasets
 - Vignettes
 - Major packages (affy, edgeR)
- Using Bioconductor/EdgeR for RNAseq
 - reading in data (what to look for)
 - removing genes with low/no signal
 - normalization
 - finding differentially expressed genes

fasta.bioch.virginia.edu/biol4230

1

To learn more:

1. Pevsner, Chapter 8 pp. 331-373
2. Draghici, Soren (2012) "Statistics and data analysis for microarrays using R and Bioconductor" Chapman and Hall
3. Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature protocols* **8**, 1765–1786 (2013).
4. http://www.bioconductor.org/help/workflows/rnaseq_Gene/

fasta.bioch.virginia.edu/biol4230

2

bioconductor.org

- More than 600 packages of functions for genome and expression analysis
 - expression analysis: *affy*
 - RNA-seq: *edgeR*, *DESeq2*
 - ChIP-seq (interaction of protein with DNA in chromatin)
 - extracting genomic features
- "Vignettes" that come with data for research problems
- Must be installed (often individually)
- Work with 'R' objects typically much more abstract than `data.frames()`
- Use the common 'R' logic for selecting rows and columns from data

fasta.bioch.virginia.edu/biol4230

3

bioconductor.org



Home Install Help Developers About

Search:

About Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, [934 software packages](#), and an active user community. Bioconductor is also available as an [AMI](#) (Amazon Machine Image) and a series of [Docker](#) images.

News

- [Nature Methods Orchestrating high-throughput genomic analysis with Bioconductor \(abstract\)](#); full-text free with registration and other recent [literature citations](#).
- Read our [latest newsletter](#).
- Updated [course material](#) and [videos](#).
- Use the [support site](#) to get help installing, learning and using Bioconductor.

Install »

Get started with Bioconductor

- [Install Bioconductor](#)
- [Explore packages](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

Learn »

Master Bioconductor tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

Use »

Create bioinformatic solutions with Bioconductor

- [Software, Annotation, and Experiment packages](#)
- [Amazon Machine Image](#)
- [Latest release announcement](#)
- [Support site](#)

Develop »

Contribute to Bioconductor

- [Use BioC 'devel'](#)
- ['Devel' Software, Annotation and Experiment packages](#)
- [Package guidelines](#)
- [New package submission](#)
- [Developer resources](#)
- [Build reports](#)

fasta.bioch.virginia.edu/biol4230

4

bioconductor installation

```
>R ...
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> source("http://bioconductor.org/biocLite.R")
trying URL
'http://www.bioconductor.org/packages/3.0/bioc/bin/macosx/mavericks/contrib/3.1/Bi
ocInstaller_1.16.2.tgz'
Content type 'application/x-gzip' length 49063 bytes (47 KB)
opened URL
=====
downloaded 47 KB
The downloaded binary packages are in
  /var/folders/cd/56p6y3_xllq_pr5lmldqgngc0000jc/T//Rtmp8pZlmn/downloaded_
packages
Bioconductor version 3.0 (BiocInstaller 1.16.2), ?biocLite for help
> biocLite()
BioC_mirror: http://bioconductor.org
Using Bioconductor version 3.0 (BiocInstaller 1.16.2), R version 3.1.3.
Installing package(s) 'Biobase' 'IRanges' 'AnnotationDbi'
also installing the dependencies 'BiocGenerics', 'S4Vectors', 'GenomeInfoDb',
'DBI', 'RSQLite'
```

Bioconductor installs packages incrementally

fastabioch.virginia.edu/biol4230

5

Bioconductor installs packages incrementally

```
> library(affy)
Error in library(affy) : there is no package called 'affy'
> biocLite('affy')
BioC_mirror: http://bioconductor.org
Using Bioconductor version 3.0 (BiocInstaller 1.16.2), R version 3.1.3.
Installing package(s) 'affy'
also installing the dependencies 'affyio', 'preprocessCore', 'zlibbioc'
trying URL
'http://bioconductor.org/packages/3.0/bioc/bin/macosx/mavericks/contrib/3.1/affyio
_1.34.0.tgz'
Content type 'application/x-gzip' length 89679 bytes (87 KB)
opened URL
=====
downloaded 87 KB
trying URL
'http://bioconductor.org/packages/3.0/bioc/bin/macosx/mavericks/contrib/3.1/prepro
cessCore_1.28.0.tgz'
Content type 'application/x-gzip' length 137216 bytes (134 KB)
opened URL
=====
The downloaded binary packages are in
  /var/folders/cd/56p6y3_xllq_pr5lmldqgngc0000jc/T//Rtmp8pZlmn/downloaded_packages
```

fastabioch.virginia.edu/biol4230

6

Bioconductor installs packages incrementally

```
> library(affy)
Loading required package: BiocGenerics
Loading required package: parallel
Attaching package: 'BiocGenerics'
The following objects are masked from 'package:parallel':
...
  parLapplyLB, parRapply, parSapply, parSapplyLB
The following object is masked from 'package:stats':
  xtabs
The following objects are masked from 'package:base':
  Filter, Find, Map, Position, Reduce, anyDuplicated, append,
...
  unique, unlist, unsplit

Loading required package: Biobase
Welcome to Bioconductor

  Vignettes contain introductory material; view with
  'browseVignettes()'. To cite Bioconductor, see
  'citation("Biobase")', and for packages 'citation("pkgname")'.
```

fastabioch.virginia.edu/biol4230

7

Bioconductor: getting help

The screenshot shows a web-based documentation interface for the Bioconductor project. At the top, there is a navigation bar with tabs for 'Files', 'Plots', 'Packages', 'Help', and 'Viewer'. The 'Packages' tab is currently selected. Below the navigation bar, there is a search bar with the text 'edgeR' and a 'Find in Topic' button. The main content area displays information about the 'edgeR-package {edgeR}'. The title is 'Empirical analysis of digital gene expression data in R'. Under the 'Description' section, it is stated that 'edgeR is a package for the analysis of digital gene expression data arising from RNA sequencing technologies such as SAGE, CAGE, Tag-seq or RNA-seq, with emphasis on testing for differential expression.' It also mentions that 'Particular strengths of the package include the ability to estimate biological variation between replicate libraries, and to conduct exact tests of significance which are suitable for small counts. The package is able to make use of even minimal numbers of replicates.' At the bottom of the page, there is a link '[Package edgeR version 3.8.6 Index]'

fastabioch.virginia.edu/biol4230

8

Bioconductor: getting help

R: Empirical analysis of digital gene expression data in R ▾

Empirical analysis of digital gene expression data in



Documentation for package 'edgeR' version 3.8.6

- [DESCRIPTION file.](#)
- [User guides, package vignettes and other documentation.](#)
- [Package NEWS.](#)

Help Pages

[A](#) [B](#) [C](#) [D](#) [E](#) [G](#) [L](#) [M](#) [N](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [V](#) [W](#) [Z](#) [misc](#)

[edgeR-package](#) Empirical analysis of digital gene expression data in R

-- A --

adjustedProfileLik	Adjusted Profile Likelihood for the Negative Binomial Dispersion Parameter
as.data.frame.TopTags	Turn a TopTags Object into a Dataframe
as.matrix.DGEList	Turn a DGEList Object into a Matrix
aveLogCPM	Average Log Counts Per Million
aveLogCPM.default	Average Log Counts Per Million
aveLogCPM.DGEGLM	Average Log Counts Per Million
aveLogCPM.DGEList	Average Log Counts Per Million

fasta.bioch.virginia.edu/biol4230

9

Bioconductor: getting help

R: Empirical analysis of digital gene expression data in R ▾

decideTestsDGE	Multiple Testing Across Genes and Contrasts
designAsFactor	Fit Negative Binomial Generalized Linear Model to Multiple Response Vectors: Low Level Functions
DGEExact-class	differential expression of Digital Gene Expression data - class
DGEGLM-class	Digital Gene Expression Generalized Linear Model results - class
DGEList	DGEList Constructor
DGEList-class	Digital Gene Expression data - class
DGELRT-class	Digital Gene Expression Likelihood Ratio Test data and results - class
dgimStdResid	Visualize the mean-variance relationship in DGE data using standardized residuals
diffSpliceDGE	Test for Differential Exon Usage
dim.DGEExact	Retrieve the Dimensions of a DGEList, DGEExact, DGEGLM, DGELRT or TopTags Object
dim.DGEGLM	Retrieve the Dimensions of a DGEList, DGEExact, DGEGLM, DGELRT or TopTags Object
dim.DGEList	Retrieve the Dimensions of a DGEList, DGEExact, DGEGLM, DGELRT or TopTags Object
dim.DGELRT	Retrieve the Dimensions of a DGEList, DGEExact, DGEGLM, DGELRT or TopTags Object
dim.TopTags	Retrieve the Dimensions of a DGEList, DGEExact, DGEGLM, DGELRT or TopTags Object
dimnames.DGEExact	Retrieve the Dimension Names of a DGE Object
dimnames.DGEGLM	Retrieve the Dimension Names of a DGE Object
dimnames.DGEList	Retrieve the Dimension Names of a DGE Object
dimnames.DGELRT	Retrieve the Dimension Names of a DGE Object
dimnames.TopTags	Retrieve the Dimension Names of a DGE Object
dimnames-<-DGEGLM	Retrieve the Dimension Names of a DGE Object
dimnames-<-DGEList	Retrieve the Dimension Names of a DGE Object
dispBinTrend	Estimate Dispersion Trend by Binning for NB GLMs
dispCoxReid	Estimate Common Dispersion for Negative Binomial GLMs
dispCoxReidInterpolateTagwise	Estimate Tagwise Dispersion for Negative Binomial GLMs by Cox-Reid Adjusted Profile Likelihood

fasta.bioch.virginia.edu/biol4230

10

Bioconductor: the installation loop

- Initial intall:
`source("http://www.bioconductor.org/biocLite.R")
biocLite()`
- When you need something:
`> library(simpleaffy)`

```
Error in library(simpleaffy) : there is no package  
called 'simpleaffy'  
> biocLite("simpleaffy")  
BioC_mirror: http://bioconductor.org  
Using Bioconductor version 3.0 (BiocInstaller 1.16.2),  
Installing package(s) 'simpleaffy'  
also installing the dependencies 'Biostrings', 'gcrma'  
...  
> library(simpleaffy)  
Loading required package: genefilter  
Attaching package: 'genefilter'
```

fastabioch.virginia.edu/biol4230

11

Differential Gene Expression

- Large quantity of data (>20,000 genes)
 - Affychip data has ~20 replicates per gene
 - RNAseq has counts (FPKM: Fragments per Kilobase per Million mapped reads)
 - but a small number of biological replicates
- Ideally, identify modest change (1.5x or larger) for modest levels of transcription
 - 10 or fewer transcripts may account for 90% of reads, so 5,000 – 10,000 transcripts for < 10% of reads
 - If technical replicates vary more than 2x, how do you measure 1.5x change?
- Large numbers of tests: how to correct?
 - Family-wide-error-rate (FWER) Bonferroni correction (used for similarity search E()-values)
 - False-discovery-rate (FDR, qvalue)

fastabioch.virginia.edu/biol4230

12

Identifying differentially expressed genes

1. convert to FPKM (probably not done properly in my example) (cpm)
2. With RNA-seq data, make sure counts > 1
3. Normalize, adjust medians, quantile normalization
4. Look at bulk properties:
 - PCA analysis should group replicates
 - variance should be relatively linear
5. Calculate pair-wise differential expression with t-tests
6. Use topTags to do FDR correction, identify largest changes
 - go back and compare topTags results to actual counts
7. Log₂(FC) vs Log₁₀(abundance)
8. Volcano plots show fold-change, q-value tradeoff

fasta.bioch.virginia.edu/biol4230

13

Measuring differences – sources of variation

Technical

- RNA isolation
- cDNA synthesis
- hybridization (AffyChip)
- PCR amplification
- G+C content
- sequencing depth
- location on AffyChip/sequencing "lane"

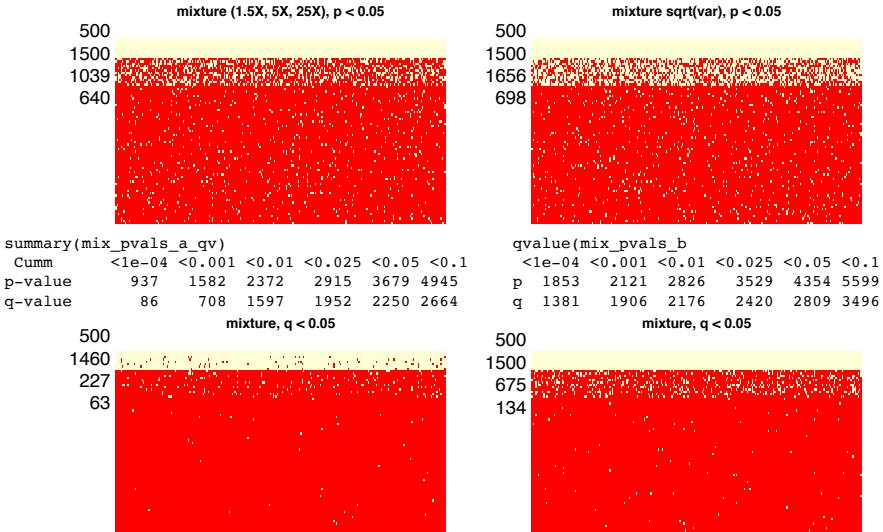
Biological

- genetic background
- sex
- last meal/sleep/exercise
- dividing/quiescent
- cell type within tissue type
- ...

fasta.bioch.virginia.edu/biol4230

14

Reducing variance improves detection



fasta.bioch.virginia.edu/biol4230

15

Identifying differentially expressed genes

- Differences in expression in the presence of noise:
 - are the differences significant? statistical model
 - t-test (normally distributed, array data)
 - negative binomial (variance increases with mean)
 - are the differences biological?
 - batch effects from experiment
 - gene effects (length, G+C)
- Analysis packages (edgeR, deSeq2) visualize batch effects, normalize data, apply statistical model

fasta.bioch.virginia.edu/biol4230

16

Differential Gene Expression with edgeR

- Starting data: HTS counts (not FPKM)

```
> GSE_HTSeq <- read.table("GSE_ENCODE_HTSseq.txt",
+                           row.names=1,sep='\t',header=T)
# row.names=1 uses gene names
> summary(GSE_HTSeq)
   GM12892_Rep1      GM12892_Rep2      GM12892_Rep3      H1.hESC_Rep1      H1.hESC_Rep2
Min.   : 0   Min.   : 0.0   Min.   : 0.0   Min.   : 0   Min.   : 0
1st Qu.: 1   1st Qu.: 1.0   1st Qu.: 1.0   1st Qu.: 4   1st Qu.: 4
Median : 103  Median : 56.0  Median : 47.0  Median : 200  Median : 208
Mean   : 1830  Mean   : 814.2  Mean   : 765.1  Mean   : 1418  Mean   : 1460
3rd Qu.: 1246  3rd Qu.: 630.0  3rd Qu.: 567.5  3rd Qu.: 1159  3rd Qu.: 1164
Max.   :1045434 Max.   :482679.0 Max.   :426204.0 Max.   :646940 Max.   :628301
```

- Do the replicates look similar?
- Approx how many genes have <= 1 count?
- Why is the Max 1000X the 3rd quartile?
- how much data?

```
> dim(GSE_HTSeq)
[1] 21711      10
```

Rapaport, F. *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* **14**, R95 (2013).

fasta.bioch.virginia.edu/biol4230

17

Differential Gene Expression with edgeR

- Wide distribution of abundance

```
> summary(GSE_HTS)
   GM12892_Rep1      GM12892_Rep2      GM12892_Rep3      H1.hESC_Rep1      H1.hESC_Rep2
Min.   : 0   Min.   : 0.0   Min.   : 0.0   Min.   : 0   Min.   : 0
1st Qu.: 1   1st Qu.: 1.0   1st Qu.: 1.0   1st Qu.: 4   1st Qu.: 4
Median : 103  Median : 56.0  Median : 47.0  Median : 200  Median : 208
Mean   : 1830  Mean   : 814.2  Mean   : 765.1  Mean   : 1418  Mean   : 1460
3rd Qu.: 1246  3rd Qu.: 630.0  3rd Qu.: 567.5  3rd Qu.: 1159  3rd Qu.: 1164
Max.   :1045434 Max.   :482679.0 Max.   :426204.0 Max.   :646940 Max.   :628301

   H1.hESC_Rep3      H1.hESC_Rep4      MCF.7_Rep1      MCF.7_Rep2      MCF.7_Rep3
Min.   : 0   Min.   : 0.0   Min.   : 0   Min.   : 0   Min.   : 0
1st Qu.: 4   1st Qu.: 2.0   1st Qu.: 2   1st Qu.: 1   1st Qu.: 2
Median : 206  Median : 120.0  Median : 215  Median : 173  Median : 187
Mean   : 1385  Mean   : 867.5  Mean   : 2160  Mean   : 2450  Mean   : 2400
3rd Qu.: 1130  3rd Qu.: 713.0  3rd Qu.: 1796  3rd Qu.: 1733  3rd Qu.: 1628
Max.   :597077  Max.   :406388.0 Max.   :639987  Max.   :816273  Max.   :885833
```

fasta.bioch.virginia.edu/biol4230

18

Differential Gene Expression with edgeR

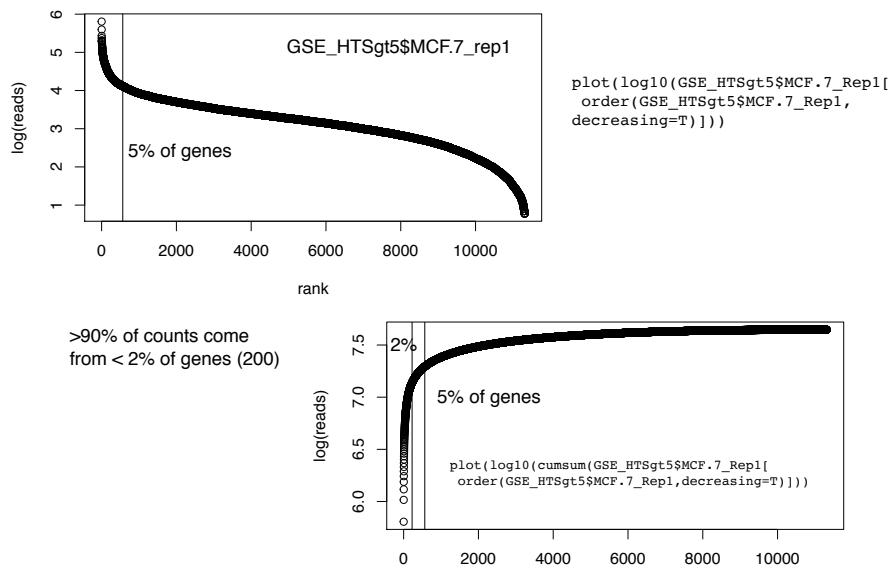
```
> min_cnt<-GSE_HTS$MCF.7_Rep1 > 5 & GSE_HTS$MCF.7_Rep2 > 5 & GSE_HTS$MCF.7_Rep3>5 &
GSE_HTS$H1.hESC_Rep1 > 5 & GSE_HTS$H1.hESC_Rep2 > 5 & GSE_HTS$H1.hESC_Rep3 > 5 &
GSE_HTS$H1.hESC_Rep4 > 5 & GSE_HTS$GM12892_Rep1 > 5 & GSE_HTS$GM12892_Rep2 > 5 &
GSE_HTS$GM12892_Rep3 > 5
> GSE_HTSgt5 <- GSE_HTS[min_cnt,]
> summary(GSE_HTSgt5)
GM12892_Rep1      GM12892_Rep2      GM12892_Rep3      H1.hESC_Rep1      H1.hESC_Rep2
Min. :       6   Min. :       6   Min. :       6   Min. :       6   Min. :       6
1st Qu.:    347   1st Qu.:    182   1st Qu.:    152   1st Qu.:    341   1st Qu.:    345
Median :  1083   Median :  537   Median :  487   Median :  936   Median :  942
Mean   :  3294   Mean   : 1458   Mean   : 1380   Mean   : 2486   Mean   : 2559
3rd Qu.:  2613   3rd Qu.: 1235   3rd Qu.: 1133   3rd Qu.: 2257   3rd Qu.: 2256
Max.  :1045434   Max.  :482679   Max.  :426204   Max.  :646940   Max.  :628301

H1.hESC_Rep3      H1.hESC_Rep4      MCF.7_Rep1      MCF.7_Rep2      MCF.7_Rep3
Min. :       6   Min. :       6   Min. :       6   Min. :  6.0   Min. :       6
1st Qu.:    342   1st Qu.:    213   1st Qu.:    530   1st Qu.: 470.8   1st Qu.:    417
Median :  922   Median :  585   Median : 1542   Median : 1412.0  Median : 1352
Mean   : 2420   Mean   : 1528   Mean   : 3918   Mean   : 4072.6  Mean   : 4294
3rd Qu.: 2195   3rd Qu.: 1393   3rd Qu.: 3642   3rd Qu.: 3441.2  3rd Qu.: 3656
Max.  :597077   Max.  :406388   Max.  :639987   Max.  :808982.0  Max.  :885833
> colSums(GSE_HTSgt5)
GM12892_Rep1  GM12892_Rep2  GM12892_Rep3  H1.hESC_Rep1  H1.hESC_Rep2  H1.hESC_Rep3  H1.hESC_Rep4
37278141     16498913     15621942     28137254     28956711     27384301     17293143
MCF.7_Rep1    MCF.7_Rep2    MCF.7_Rep3
44331343     46086020     48589601
```

fasta.bioch.virginia.edu/biol423

19

Abundance differences in MCF.7 mRNA



fasta.bioch.virginia.edu/biol423

20

Differential Gene Expression with edgeR

- Convert to a dge (edgeR) structure:

```
>GSE_dge<-DGEList(counts=GSE_HTSeq,lib.size=colSums(GSE_HTSeq),  
+ group=c(rep("GM128",3),rep("H1",4),rep("MCF7",3)))
```

- Select genes with at least n counts

```
>GSE_cpmns<-cpm(GSE_dge) # cpm, counts per kb per million (FPKM),  
# needs gene lengths  
>keep2 <- rowSums(GSE_cpmns[,1:3])>5 & rowSums(GSE_cpmns[,4:7]) > 5 &  
rowSums(GSE_cpmns[,8:10])>5  
> length(GSE_cpmns[keep2,1])  
[1] 10147
```

- Set up groups of factors so replicates can be combined:

```
> GSE_groups<-c(rep("GM128",3),rep("H1",4),rep("MCF7",3))  
> GSE_groups  
[1] "GM128" "GM128" "GM128" "H1" "H1" "H1" "H1" "MCF7"  
"MCF7"
```

fasta.bioch.virginia.edu/biol4230

21

Differential Gene Expression with edgeR

- build new dge (edgeR) structure for genes with counts:

```
>GSE_d2<-DGEList(counts=GSE_counts2,lib.size=colSums(GSE_counts2),  
+ group=GSE_groups)  
> summary(GSE_counts2[,c(1,2,4,5,8)])  
GM12892_Rep1 GM12892_Rep2 H1.hESC_Rep1 H1.hESC_Rep2 MCF.7_Rep1  
Min. : 0 Min. : 0.0 Min. : 18 Min. : 31.0 Min. : 0  
1st Qu.: 522 1st Qu.: 266.5 1st Qu.: 472 1st Qu.: 475.5 1st Qu.: 741  
Median : 1274 Median : 635.0 Median : 1097 Median : 1103.0 Median : 1798  
Mean : 3678 Mean : 1624.4 Mean : 2736 Mean : 2816.0 Mean : 4259  
3rd Qu.: 2884 3rd Qu.: 1353.5 3rd Qu.: 2496 3rd Qu.: 2472.5 3rd Qu.: 3998  
Max. :1045434 Max. :482679.0 Max. :646940 Max. :628301.0 Max. :639987
```

– still see differences in bulk properties, but what is 1st quartile now?

– notice that mean is > 3rd quartile. Why?

```
> summary(GSE_HTSeq)  
GM12892_Rep1 GM12892_Rep2 H1.hESC_Rep1 H1.hESC_Rep2  
Min. : 0 Min. : 0.0 Min. : 0 Min. : 0  
1st Qu.: 1 1st Qu.: 1.0 1st Qu.: 4 1st Qu.: 4  
Median : 103 Median : 56.0 Median : 200 Median : 208  
Mean : 1830 Mean : 814.2 Mean : 1418 Mean : 1460  
3rd Qu.: 1246 3rd Qu.: 630.0 3rd Qu.: 1159 3rd Qu.: 1164  
Max. :1045434 Max. :482679.0 Max. :646940 Max. :628301
```

fasta.bioch.virginia.edu/biol4230

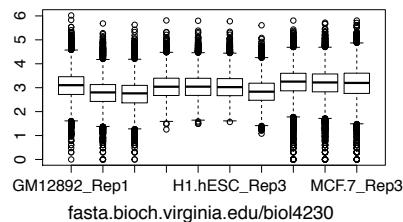
22

Differential Gene Expression with edgeR

- build new dge (edgeR) structure for genes with counts:

```
>GSE_d2<-DGEList(counts=GSE_counts2,lib.size=colSums(GSE_counts2),
+  group=GSE_groups)
> summary(GSE_counts2[,c(1,2,4,5,8)])
GM12892_Rep1      GM12892_Rep2      H1.hESC_Rep1      H1.hESC_Rep2      MCF.7_Rep1
Min. : 0           Min. : 0.0        Min. : 18          Min. : 31.0       Min. : 0
1st Qu.: 522       1st Qu.: 266.5     1st Qu.: 472         1st Qu.: 475.5     1st Qu.: 741
Median : 1274      Median : 635.0      Median : 1097        Median : 1103.0     Median : 1798
Mean   : 3678      Mean   : 1624.4     Mean   : 2736        Mean   : 2816.0      Mean   : 4259
3rd Qu.: 2884      3rd Qu.: 1353.5     3rd Qu.: 2496        3rd Qu.: 2472.5     3rd Qu.: 3998
Max.  :1045434     Max.  :482679.0     Max.  :646940       Max.  :628301.0     Max.  :639987
```

- still see differences in bulk properties, but what is 1st quartile now?
- notice that mean is > 3rd quartile. Why?
- Are the bulk properties similar?



23

How to compare relative mRNA expression?

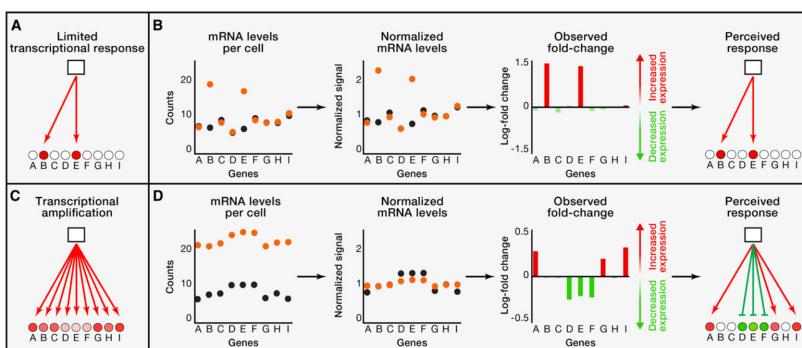


Figure 1. Normalization and Interpretation of Expression Data

Lovén, J. et al. *Cell* **151**, 476–482 (2012).

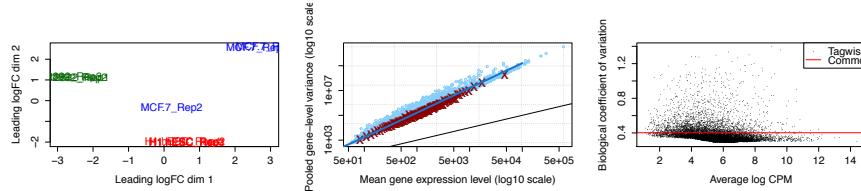
fasta.bioch.virginia.edu/biol4230

24

Differential Gene Expression with edgeR

- do some simple normalization, evaluate data quality:

```
> GSE_d2<-calcNormFactors(GSE_d2)
# plot Principal Components Analysis (PCA) of fold-changes
> plotMDS(GSE_d2,labels=colnames(GSE_counts2),
+ col=c("darkgreen", "red", "blue")[factor(GSE_groups)])
> GSE_d2<-estimateCommonDisp(GSE_d2)
> GSE_d2<-estimateTagwiseDisp(GSE_d2)
> plotMeanVar(GSE_d2,show.tagwise.vars=TRUE,NBline=TRUE)
> plotBCV(GSE_d2) # BCV = Biological Coefficient of Variation
```



fasta.bioch.virginia.edu/biol4230

25

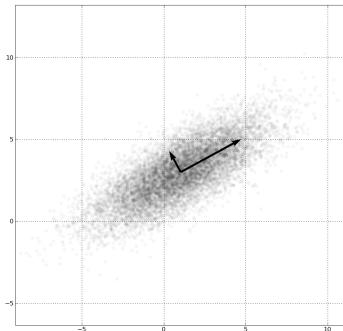
Diversion: Principal Components Analysis (PCA)

- We are interested in the differences, and similarities, between biological samples (replicate treated and controls), from the perspective of expression levels of 10,000 – 20,000 genes.
 - imagine that in the treated sample (e.g. a BHA-treated liver vs normal), only ONE gene has increased expression: (GSTM1), all the rest are the same.
 - to find this gene, we might plot our 6 samples (3 treated, 3 controls) in n=20,000 dimensional space (one axis for every gene), and look so see which point has moved between treated and controls.
 - but if only ONE gene has changed expression level, then all the other genes will be highly correlated, so we do not need 20,000 dimensions, we only need ONE (or possibly two, the second for random noise)
- Principal Components Analysis (PCA) examines the correlation between the datasets, and reduces the dimensionality to the minimum number of "axes" (Principal Components) to explain the variation in the data.
 - first component has most variance – shows a weighting of a gene set that with internal expression correlation, but different from genes not in the set
 - replicate samples should be similar; different samples should be different
 - check for outliers

fasta.bioch.virginia.edu/biol4230

26

Diversion: Principal Components Analysis (PCA)



A scatter plot of samples that are distributed according a multivariate (bivariate) Gaussian distribution centered at (1,3) with a standard deviation of 3 in roughly the (0.878, 0.478) direction and of 1 in the orthogonal direction. The directions represent the Principal Components (PC) associated with the sample

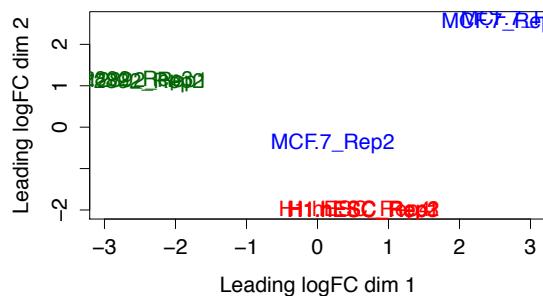
en.wikipedia.org/wiki/Principal_component_analysis#/media/File:GaussianScatterPCA.png
"GaussianScatterPCA" by — Ben FrantzDale (talk)

fasta.bioch.virginia.edu/biol4230

27

Diversion: Principal Components Analysis (PCA)

```
> GSE_c2_prin<-princomp(GSE_counts2,cor=T)
> summary(GSE_c2_prin)
Importance of components:
          Comp.1        Comp.2        Comp.3        Comp.4        Comp.5        Comp.6 
Standard deviation   2.9115294  0.78123223  0.66206994  0.55342497  0.301356842 0.200813419 
Proportion of Variance 0.8477003  0.06103238  0.04383366  0.03062792  0.009081595 0.004032603 
Cumulative Proportion  0.8477003  0.90873271  0.95256637  0.98319429  0.992275889 0.996308492
```



fasta.bioch.virginia.edu/biol4230

28

Differential Gene Expression with edgeR

- Normalization done: do t-tests on gene groups

```
> GSE_de2 <- exactTest(GSE_d,pair=c("GM128","H1")) # two-way compare
> GSE_tt2<-topTags(GSE_de2,n=nrow(GSE_de2))
> head(GSE_tt2$table,n=10)
   logFC      logCPM      PValue        FDR
LCP1    -10.077901 10.155089 8.290374e-57 8.412243e-53
RASSF5     -6.159025 5.542985 5.657588e-51 2.870377e-47
TIFA     -5.809657 5.629944 1.409506e-44 4.767419e-41
SMARCA2    -5.407642 7.584283 3.756685e-43 9.529771e-40
DOCK10    -5.948203 5.386774 3.129221e-41 6.350442e-38
CD58     -5.876585 5.406012 1.194831e-40 2.020658e-37
SMAP2     -5.036339 6.867746 3.039079e-40 4.405362e-37
NEAT1     -5.451577 9.251840 1.435513e-38 1.820769e-35
EPHB4      5.930016 7.329484 5.557970e-38 6.266302e-35
BCL2     -5.237690 5.635830 1.829338e-37 1.856230e-34
```

- what direction are the fold changes? all the same?
- why are PValues < FDR?

fasta.bioch.virginia.edu/biol4230

29

Differential Gene Expression with edgeR

- look at "top tags"

```
> GSE_counts2[row.names(head(GSE_tt2$table,n=10)),c(1:3,4:5)]
   GM12892_Rep1 GM12892_Rep2 GM12892_Rep3 H1.hESC_Rep1 H1.hESC_Rep2
LCP1         153736      50863      58206       107        89
RASSF5        5356       2659       1983        63        66
TIFA          3788       2800       2543        91        63
SMARCA2       21764      8381       8004       433       323
DOCK10        4040       2295       2024        77        39
CD58          3145       2201       2219        56        49
SMAP2          12461      5928       4385       317       288
NEAT1          57519      21160      26498      1056       896
EPHB4           149        64         47       7104      7266
BCL2          4073       3012       2634       121       111
```

- Most significant changes have high counts in GM128, low counts in H1.hESC, or vice-versa (EPHB4).
- Go back and look at the data. Does it make sense?

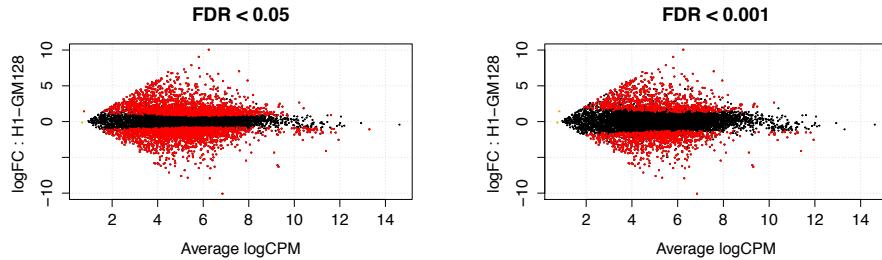
fasta.bioch.virginia.edu/biol4230

30

Differential Gene Expression with edgeR

- look at differential expression 1:

```
> GSE_deg2<-GSE_rn[GSE_tt2$table$FDR < 0.05] # 3879 genes  
> GSE_deg2_001<-GSE_rn[GSE_tt2$table$FDR < 0.001] # 1840 genes  
> plotSmean(GSE_d2,de.tags=GSE_deg2,main="FDR < 0.05")  
> plotSmean(GSE_d2,de.tags=GSE_deg2_001,main="FDR < 0.001")
```



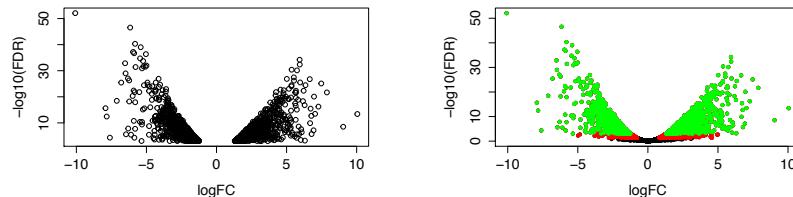
fasta.bioch.virginia.edu/biol4230

31

Differential Gene Expression with edgeR

- look at differential expression 2: volcano plots

```
> plot(GSE_tt2$table$logFC, -log10(GSE_tt2$table$FDR),  
+ xlab="logFC",ylab="-log10(FDR)")
```



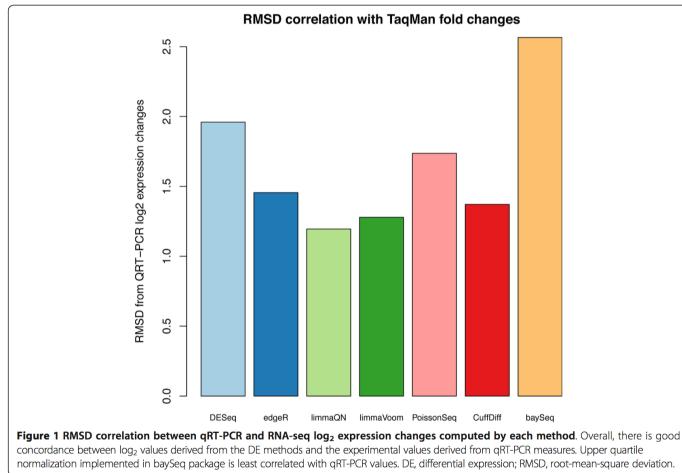
look at differential expression 2: volcano plots

```
> plot(GSE_tt2$table[,1], -log10(GSE_tt2$table[,4]),  
+ xlab="logFC",ylab="-log10(FDR)",pch=20)  
> points(GSE_tt2$table[GSE_deg2,1], -log10(GSE_tt2$table[GSE_deg2,4]),  
+ pch=20,col='red')  
> points(GSE_tt2$table[GSE_deg2_001,1], -log10(GSE_tt2$table[GSE_deg2_001,4]),  
+ pch=20,col='green')
```

fasta.bioch.virginia.edu/biol4230

32

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.

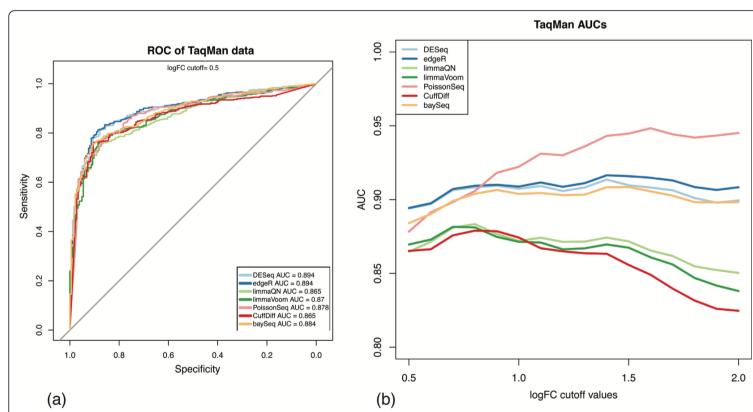


Rapaport, F. et al.. *Genome Biol* **14**, R95 (2013).

fasta.bioch.virginia.edu/biol4230

33

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.



Rapaport, F. et al.. *Genome Biol* **14**, R95 (2013).

fasta.bioch.virginia.edu/biol4230

34

Identifying differentially expressed genes

1. convert to FPKM (probably not done properly in my example) (cpm)
2. With RNA-seq data, make sure counts > 1
3. Normalize, adjust medians, quantile normalization
4. Look at bulk properties:
 - PCA analysis should group replicates
 - variance should be relatively linear
5. Calculate pair-wise differential expression with t-tests
6. Use topTags to do FDR correction, identify largest changes
 - go back and compare topTags results to actual counts
7. Log₂(FC) vs Log₁₀(abundance)
8. Volcano plots show fold-change, q-value tradeoff

fasta.bioch.virginia.edu/biol4230

35

Bioconductor: summary

- Bioconductor: a comprehensive 'R' package for expression and genome analysis
 - Obtaining/installing
 - Datasets
 - Vignettes
 - Major packages (affy, edgeR2)
- Using Bioconductor/EdgeR for RNAseq
 - reading in data (what to look for)
 - removing genes with low/no signal
 - normalization
 - finding differentially expressed genes

Always look at the RAW data
that produced the list of genes

fasta.bioch.virginia.edu/biol4230

36