

Intro to R/Bioconductor

HMS Research Computing

Spring 2016

Kristina M Holton



What can you do with R?

- R is a statistical language
- R is free!
- Bioconductor - thousands of packages for your workflow
- biomaRt - annotate samples
- Heavy duty statistics!
- Make plots!



Course Objectives

- Learn to run R on Orchestra
- Gain familiarity with R language
- Learn to import/export data
- Simple Statistics/Plotting
- Use case: RNA-seq



Notation

- mfk8@balcony:~\$

Orchestra bash (your console/terminal)

- >

In R (either personal computer or in Orchestra running R)

Bioconductor

- “Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data.”
- Fast, easy way to access packages.
- Source Bioconductor: tells R to look at Bioconductor repository for packages
- Top of script:

```
> source("http://bioconductor.org/biocLite.R")  
> biocLite()
```

Installing packages through Bioconductor

- Available packages:

http://www.bioconductor.org/packages/release/BiocViews.html#___Software

- Easy install of new packages:

```
> biocLite("nameofpackage")
```

At the top of every R session using the package, it must be called via:

```
> library("nameofpackage")
```

- Example:

```
> library("cummeRbund")
```

Installing packages through CRAN

- CRAN is the *Comprehensive R Archive Network*
- First time install of a package:

```
> install.packages("name-of-package", repos="http://cran.r-project.org")
```

(asks you to select a mirror: pick something from or near the country you're in)

- At the top of every R session using the package, the package must be called via:
> `library("nameofpackage")`

R documentation

- General R help on a function
 - > ?name_of_function
 - > help(name_of_function)
- Learn how to use your package (vignettes open in your web browser)
 - > browseVignettes(package = "name_of_package")

Setting your R “working directory”

- What is your current WD?
> getwd()
- Setting your WD:
> setwd(“path”)
- An Orchestra example:
> setwd(“/home/mfk8/MyDataDirectory”)
- A Mac example:
> setwd(“/Users/mfk8/MyDataDirectory”)
- A Windows example (note forward slashes):
> setwd(“C:/Users/mfk8/My Documents/MyDataDirectory”)

R Basics:

- To “print” in R
Just type a variable or object’s name, R will display as much as it can
- “Commenting” in R
means what appears afterwards is not passed as an argument
- Keep a file of your commands – anything from Sublime to Notepad++ to Google Drive notepad: have syntax highlighting
- You can copy-paste multiple times, this overwrites.
- Watch quotation marks: R gets fussy with certain smart quotes/fonts

How R Thinks: Variables

- Assign variables with a <- (traditional) or = (new way)
- A variable can be overwritten so be careful with naming
- Names can be upper/lowercase mixes

```
> myX <- 5
```

```
> myX
```

```
[1] 5
```

Data Type: Vectors

- Basic way to store data
- c stands for “concatenate”: put these together as a vector

```
> myvector <- c(3,5,7)
```

```
> myvector
```

```
[1] 3 5 7
```

Vectors Types

- numeric:

```
> mynumeric <- c(3,5,7)
```

- character:

```
> mycharacter <- c("bob", "nancy", "jose")
```

#note quotation marks “ ” (single or double fine, try to be consistent)

- logical:

```
mylogical <- c(TRUE, FALSE, TRUE)
```

Changing Your Vector Type:

- General workflow:
 - > variable <- as.type(variable)
 - > myvector <- as.character(myvector)
 - > myvector

[1] "3", "5", "7" #see how these are now in " " like a character vector
- Applicable to changing between other types
- Where this comes in handy: when R says you are trying to do an operation on your variable that is one type of vector, when it has to be another type. Use wisely.

Data Type: Lists

- Like vectors with mixed data types (numeric, character, logical)
- `mylist <- c("3", "TP53", FALSE)`

Data Type: Factors

- Makes a vector nominal (able to be ordered by integers)
- Create a variable “gender” with 2 "male" entries and 4 "female" entries

```
> gender <- c(rep("male", 2), rep("female", 4))
> gender <- factor(gender) # stores gender as 2 2's and 4 1's
and associates
> gender
[1] male  male  female female female female
Levels: female male
```
- Now 1=female, 2=male internally (alphabetically)
- R now treats gender as a nominal variable

Data Type: Matrices

- Data must be all the same type (numeric, character, logical)
- Columns must have the same length
- Indexed by [row,column]
 - > `mymatrix[1,1]` #returns row 1, col 1
 - > `mymatrix[1,]` #returns all of row 1
 - > `mymatrix[,1]` #returns all of column 1

Data Type : Data Frames

- A subset of matrices
- Data can be mixed types (numeric, character, logical)
- You can give columns names so you can index by them!

```
> names(mydataframe) <- c("column1name",  
"column2name")
```

```
> mydataframe$column1name #works on column1
```

```
> mydataframe[,1] #works on column1
```

- To make a data frame into a matrix for certain operations:

```
> mymatrix <- as.matrix(mydataframe)
```

Useful functions:

- > `class(object)` #gives object class
- > `length(vector)` #gives length
- > `dim(object)` #gives matrix/dataframe dimensions
- > `nrow(object)` #gives number of rows
- > `ncol(object)` #gives number of columns
- > `str(object)` #gives object structure
- > `head(object)` #gives first 6 rows
- > `tail(object)` #gives last 6 rows
- > `summary()` #quick statistics

Adding and joining rows/columns

- “rbind” to add a row or another df/matrix to a pre-existing dataframe/matrix

```
> mymatrix <- rbind(mymatrix, newrow)
> mymatrix <- rbind(mymatrix, matrixtwo)
```
- “cbind” to add a column or another df/matrix to a pre-existing dataframe/matrix

```
> mymatrix <- cbind(mymatrix, newcol)
> mymatrix <- cbind(mymatrix, matrixtwo)
```

“for” loops

- Way to iterate over data
- R sometimes breaks with complicated “for” “if” “else” loops
- `> myvector <- c(1,3,5)`
- `> j <- NULL` #initialize it, good practice
- `> newvector <- NULL` #initialize it, good practice
- `> for (i in myvector) {
 j <- i + 20
 newvector <- c(newvector, j)
}`
`> newvector`

Functions

- Way to pack up commands into a repeatable format

```
> five = 5
```

```
> three = 3
```

```
> myfunction <- function(x,y) {
```

```
  z = x + y
```

```
  return(z) #what output is seen outside of function
```

```
}
```

```
> myfunction(five,three)
```

```
[1] 8
```

Apply

- Returns an object based on applying a function to a dataframe or matrix or list
- Format: `apply (to_what, how, function)`
- “1” is to apply over rows, “2” is to apply over columns
 - > `mymatrix <- apply(mymatrix, 1, function)`
 - > `mymatrix <- apply(mymatrix, 2, function)`
- Variants: `lapply`, `sapply`, `vapply`, `by`, `replicate`, `mapply`, `rapply`, `tapply`

Importing Data

- You can specify how your data is separated (comma separated: “,” tab: “\t” space: “ ”), and if the first row is a “header” row containing the column names)
 - `mydata <- read.table(file="PathToFile/filename.csv", header=TRUE, sep=",")`
 - `#add "row.names=1" to make column 1 the rownames`

Importing Data from MS Excel

- Read in the first worksheet from the workbook myexcel.xlsx
- First row contains variable (column) names
 - > library(xlsx) #install the first time from CRAN
 - > mydata <- read.xlsx("c:/myexcel.xlsx", 1)
- Read in the worksheet named mysheet
 - > mydata <- read.xlsx("c:/myexcel.xlsx", sheetName = "mysheet")

Importing Data from SPSS

- In SPSS: save SPSS dataset in transport format
get file='c:\mydata.sav'.
export outfile='c:\mydata.por'.
- in R

```
> library(Hmisc) #install the first time from CRAN  
> mydata <- spss.get("c:/mydata.por",  
use.value.labels=TRUE)  
  
# last option converts value labels to R factors
```

Importing Data from SAS

- In SAS: save SAS dataset in transport format
libname out xport 'c:/mydata.xpt';
data out.mydata;
set sasuser.mydata;
run;
- In R
> library(Hmisc) #install the first time from CRAN
> mydata <- sasxport.get("c:/mydata.xpt")
character variables are converted to R factors

Importing Data from STATA

- In R: input Systat file
 - > library(foreign) #install the first time from CRAN
 - > mydata <- read.systat("c:/mydata.dta")

Exporting Data

- Easy way to export a variable (vector, dataframe, matrix, etc):

```
> write.table(nameofvariable, file="path/nameoffile.csv", sep=",")  
#sep="\t" or " " etc
```

Saving your workspace

- Save and pick up where you leave off – saves variables
 - > save.image()
 - > save(object list, file="mysaves.RData")
- Loading workspaces
 - > load("mysaves.RData")

Class Sample

- Download class data and R script to a folder from <http://hmsrc.me/RBioconductorCourse>
- Set your working directory to the folder where your data is

```
> setwd("pathtofolder/note/forward/slashes")
```

Class Sample – Import Data

- Import Rcoursetestdata1.csv as data frame, with headers and row names

```
> mydf <- read.table("Rcoursetestdata1.csv", header=TRUE,  
row.names=1, sep=",")
```

```
> head(mydf)
```

```
TNBC1 TNBC2 TNBC3 Normal1 Normal2 Normal3
```

```
ENSG000000008988 15258 15077 144720 12095 43544 46883
```

```
ENSG000000009307 14660 20767 8678 13774 23030 18917
```

```
ENSG000000019582 50866 55775 15089 6696 13754 86319
```

```
ENSG000000026025 21174 47966 26682 6068 21126 12728
```

```
ENSG000000034510 25645 31574 56403 29590 25216 37199
```

```
ENSG000000044574 23910 27200 13757 13364 10852 12378
```


Class Sample - Continued

- Get basic statistics on mydf

```
> summary(mydf)
```

```
TNBC1      TNBC2      TNBC3      Normal1
Min.   :  0  Min.   : 65  Min.   : 31  Min.   : 22
1st Qu.: 7888 1st Qu.: 9538 1st Qu.: 9324 1st Qu.: 5074
Median : 13034 Median : 16568 Median : 19108 Median : 10869
Mean   : 18596 Mean   : 26036 Mean   : 25646 Mean   : 14746
3rd Qu.: 23850 3rd Qu.: 28194 3rd Qu.: 30389 3rd Qu.: 18866
Max.   :103007 Max.   :351603 Max.   :272582 Max.   :89837

Normal2      Normal3
Min.   : 208  Min.   : 15
1st Qu.: 7124 1st Qu.: 8944
Median : 14005 Median : 17710
Mean   : 19425 Mean   : 25481
3rd Qu.: 21576 3rd Qu.: 32191
Max.   :212582 Max.   :244692
```



Class Sample – Transposing Data

- Need your data to read the other way? Turn it into a matrix, and transpose!
> `mymatrix <- as.matrix(mydf)`
> `myTmatrix<- t(mymatrix) #t = transpose`
> `myTdf <- as.data.frame(myTmatrix) #makes a data frame again`
> `myTdf`

ENSG00000008988 ENSG00000009307 ENSG00000019582 ENSG00000026025

TNBC1	15258	14660	50866	21174
TNBC2	15077	20767	55775	47966
TNBC3	144720	8678	15089	26682
Normal1	12095	13774	6696	6068
Normal2	43544	23030	13754	21126
Normal3	46883	18917	86319	12728

Class Example – Data Ops Cont'd

- Get basic statistics on transposed data frame

> `summary(myTdf)`

ENSG00000008988 ENSG00000009307 ENSG00000019582 ENSG00000026025

Min. : 12095 Min. : 8678 Min. : 6696 Min. : 6068

1st Qu.: 15122 1st Qu.:13996 1st Qu.:14088 1st Qu.:14828

Median : 29401 Median :16789 Median :32978 Median :21150

Mean : 46263 Mean :16638 Mean :38083 Mean :22624

3rd Qu.: 46048 3rd Qu.:20305 3rd Qu.:54548 3rd Qu.:25305

Max. :144720 Max. :23030 Max. :86319 Max. :47966

Class Example: Simple t-test

- Do a t-test on TNBC1 and Normal1 from mydf

```
> t.test(mydf$TNBC1, mydf$Normal1)
```

```
tdata: mydf$TNBC1 and mydf$Normal1
```

```
t = 2.3766, df = 391.514, p-value = 0.01796
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
664.969 7034.341
```

```
sample estimates:
```

```
mean of x mean of y
```

```
18596.13 14746.48 .test(mydf$TNBC1, mydf$Normal1)
```

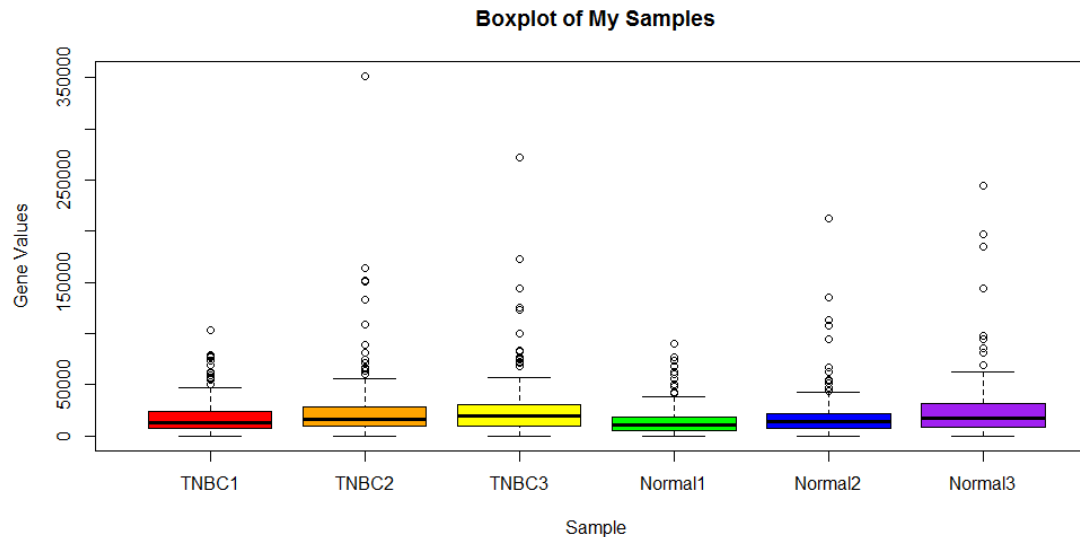
Saving plots

- Use “File->”save as” or,
- Open up a plot file (png, pdf, jpeg, tiff, bmp)
> png(file=“nameyourfile.png”)
- Make your plot
> plot()
- Turn plot off
> dev.off()

Class Example – Boxplot of Samples

- Boxplot of Samples

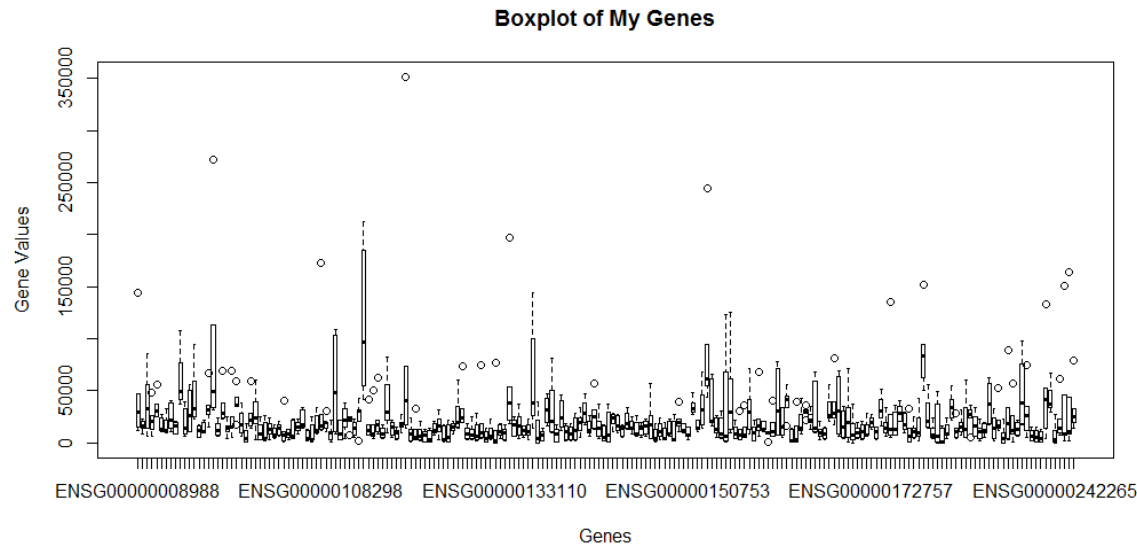
```
> boxplot(mydf, main="Boxplot of My Samples",  
xlab="Sample", ylab="Gene Values")
```



Class Example – Boxplot of Genes

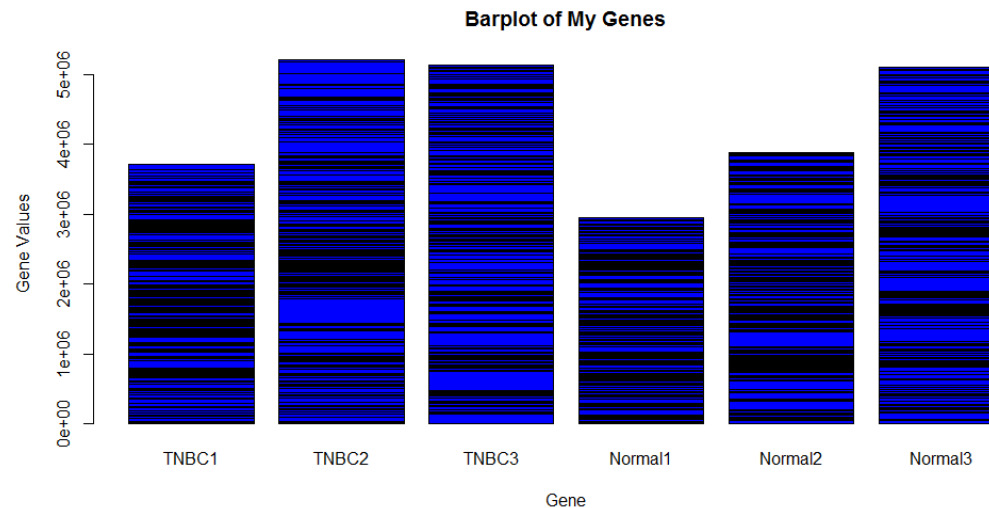
- Boxplot of Genes

```
> boxplot(myTdf, main="Boxplot of My Genes",  
xlab="Gene", ylab="Gene Values")
```



Class Example – Barplot of Genes

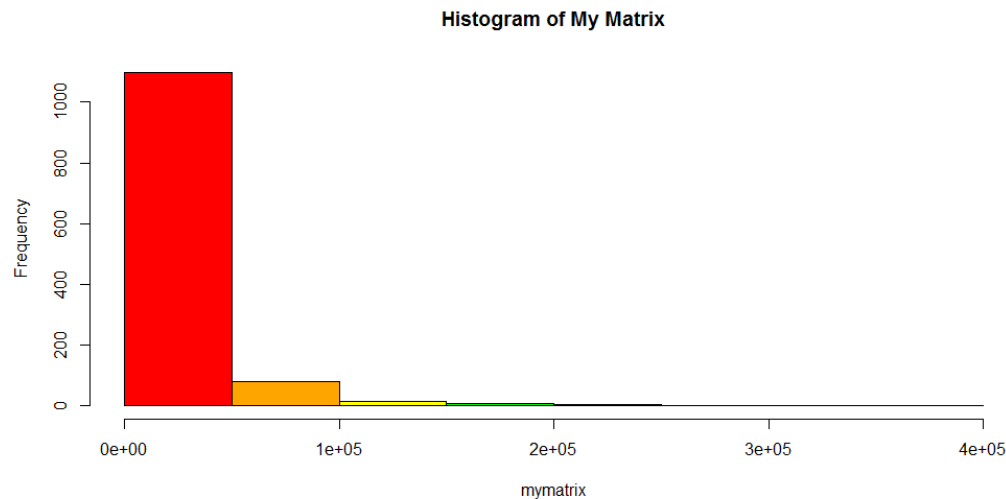
- Barplot of Genes #barplot needs a matrix
- ```
> barplot(myTmatrix, main="Barplot of My Genes",
xlab="Gene", ylab="Sample Values")
```





# Class Example – Histogram of Values

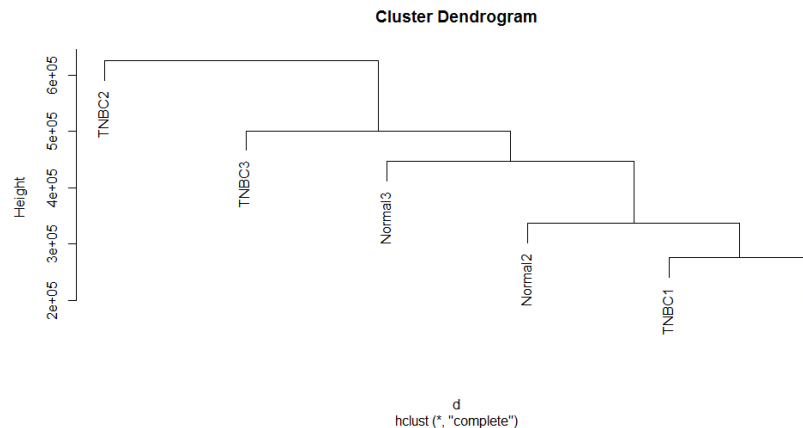
- Plot a histogram of the frequency of values in mymatrix  
`>hist(mymatrix)`



# Class Example - HCL

- Do hierarchical clustering of samples

```
> d <- dist(myTmatrix) #takes matrix as input
> hc <- hclust(d)
> plot(hc)
```



# Class Sample – Line Graphs

- #start a line plot with ENSG00000008988 "b" means both points and lines  

```
> plot(myTdf$ENSG00000008988, type="b", col="green",
ylim=c(10000,150000), main="Gene Values Over Samples",
xlab="Sample", ylab="Gene Values")
```
- #add a new line for ENSG00000009307 "lines" adds a line to the current plot  

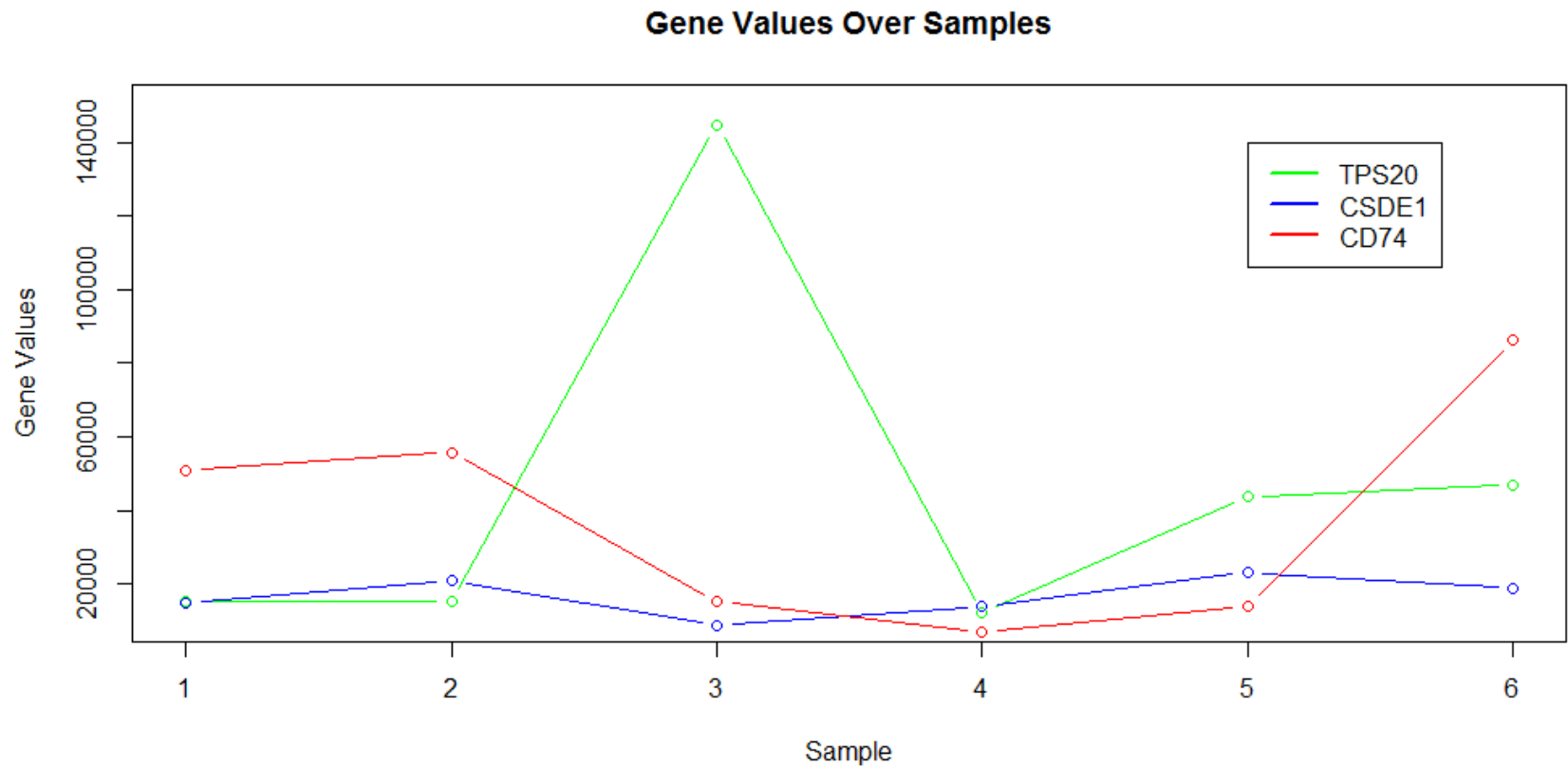
```
> lines(myTdf$ENSG00000009307, type="b", col = "blue")
```
- #add a new line for ENSG00000019582  

```
> lines(myTdf$ENSG00000019582, type="b", col="red")
```

# Class Sample – Line Graphs Legend

```
> legend(5, 140000, #positions x, y
c("TPS20", "CSDE1", "CD74"), #line names
lty=c(1,1), #specifies lines
lwd=c(2.5,2.5), #specifies line width
col=c("green", "blue", "red") #add colors
) #ends legend
```

# Class Sample – Line Graphs



# Class Example – Count Data

- Import count data

```
> mytable<-read.table("Rcoursetestdata2.csv", header=T, row.names=1, sep=",")
```

```
> mytable
```

|    | neversmoke | smoke | pastsmoke | gender |
|----|------------|-------|-----------|--------|
| 1  | 0          | 1     | 0         | male   |
| 2  | 0          | 1     | 0         | female |
| 3  | 0          | 0     | 0         | male   |
| 4  | 1          | 0     | 0         | female |
| 5  | 0          | 0     | 1         | male   |
| 6  | 0          | 1     | 0         | female |
| 7  | 0          | 0     | 1         | male   |
| 8  | 0          | 0     | 1         | female |
| 9  | 1          | 0     | 0         | male   |
| 10 | 0          | 0     | 1         | female |

# Class Example – Count Data Cont'd

- Get summary statistic on mytable

```
> summary(mytable)
```

```
> summary(mytable)
```

```
 neversmoke smoke pastsmoke gender
```

```
Min. :0.0 Min. :0.00 Min. :0.0 female:5
```

```
1st Qu.:0.0 1st Qu.:0.00 1st Qu.:0.0 male :5
```

```
Median :0.0 Median :0.00 Median :0.0
```

```
Mean :0.2 Mean :0.30 Mean :0.4
```

```
3rd Qu.:0.0 3rd Qu.:0.75 3rd Qu.:1.0
```

```
Max. :1.0 Max. :1.00 Max. :1.0
```

```
>
```

# Class Example – Tables Cont'd

- Take a subset of a table to do statistics on

```
> genderVsmoke <- table(mytable$gender, mytable$smoke)
```

```
> genderVsmoke
```

```
 0 1
```

```
female 3 2
```

```
male 4 1
```



# Class Example – Tables Cont'd

- `summary(genderVsmoke)` does a Chi Square Test of Independence

> `summary(genderVsmoke)`

Number of cases in table: 10

Number of factors: 2

Test for independence of all factors:

Chisq = 0.4762, df = 1, p-value = 0.4902

Chi-squared approximation may be incorrect

# Class Example – Tables Cont'd

- Get the frequencies of a table via prop.table

```
> propgenderVsmoke <- prop.table(genderVsmoke)
> propgenderVsmoke
```

|        | 0   | 1   |
|--------|-----|-----|
| female | 0.3 | 0.2 |
| male   | 0.4 | 0.1 |

# Class Example – Plotting Tables

- Do a mosaicplot of table genderVsmoke  
> `mosaicplot(genderVsmoke)`



# Bonus Plot: Heatmap

- Try installing packages “gplots” and “RColorBrewer”  
> `install.packages(“gplots”, “RColorBrewer”)`
- Call the libraries  
> `library(“gplots”)`  
> `library(“RColorBrewer”)`
- Check out the code!

# R on Orchestra

- Open a high-memory R session – better than a desktop!
- Log in to Orchestra with X11 enabled (important for graphics)
- Mac: Xquartz installed, in console  
`ssh -X -l user123 orchestra.med.harvard.edu`
- Linux  
`ssh -X user123@orchestra.med.harvard.edu`
- Windows: Xming, Putty installed (check X11 forwarding)  
`user123@orchestra.med.harvard.edu`

# LSF and Orchestra

- LSF is how we interact with the cluster
- Simple interactive session:

```
mfk8@balcony:~$ bsub -ls -q interactive bash
```

- For more RAM:

```
mfk8@balcony:~$ bsub -ls -q interactive -R "rusage[mem=16000]"
bash
```

(where 16000 is memory requested in MB (16000 = 16GB))

- Rmpi: How to run R scripts over multiple cores

# R Versions

- `mfk8@balcony:~$ module avail stats/R`
- Why does it matter what version of R you run?

Downstream packages may only work with certain versions of R.

- How to load a version of R:

`mfk8@balcony:~$ module load stats/R/version`

- Unloading R

`module unload stats/R/version`

- Starting R:

`mfk8@clarinet001:~$ R`

# Managing your R packages on Orchestra

- It is best to manage your own R packages to work with the version of R you select. In doing so, there are no disruptions to your workflow.
- Setting up your Orchestra R library:

```
mfk8@balcony:~$ mkdir -p ~/R/library
```

```
mfk8@balcony:~$ echo
'R_LIBS_USER=~"/R/library"'> $HOME/.Renvron
```

```
mfk8@balcony:~$ export R_LIBS_USER="/home/user123/R/library"
```

- If you must manually download a package (not through Bioconductor/CRAN etc), put the package in the set up location (/home/user123/R/library)
- Accessing packages manually uploaded to your Orchestra R library (first time)  

```
> install.packages("name-of-your-package",lib=~"/R/library")
```



# Useful R packages

---

- “RColorBrewer” – define colors and pallets
- “gplots” and “ggplots” – great for plotting
- “genefilter” – useful to apply filters over matrices
- “edgeR” and “DESeq2” – RNAseq differential analysis alternatives using count data as input
- “cummeRbund” – front end to extract, manage, and plot cuffDiff
- “biomaRt” – cross-annotate samples

# Contact Information

Survey!

<http://hmsrc.me/RBioconductorSurvey3>

HMS Research Computing

[rchelp@hms.harvard.edu](mailto:rchelp@hms.harvard.edu)

Office Hours Wed 1-3p

500 Gordon Hall

<http://rc.hms.harvard.edu/>