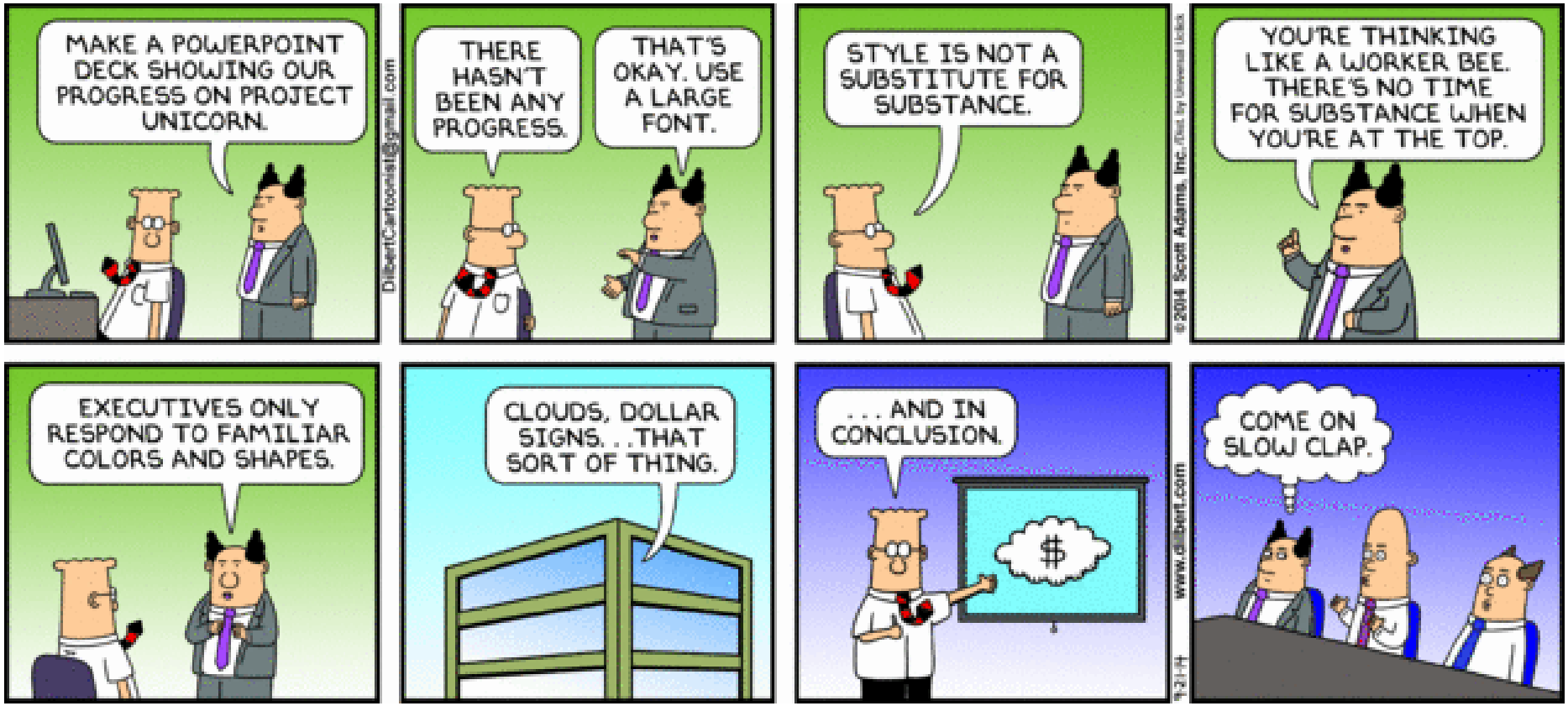

MAKE A POWERPOINT DECK SHOWING OUR PROGRESS ON PROJECT UNICORN.
DilbertCartoonist@gmail.com
THERE HASN'T BEEN ANY PROGRESS.
THAT'S OKAY. USE A LARGE FONT.
STYLE IS NOT A SUBSTITUTE FOR SUBSTANCE.
© 2014 Scott Adams, Inc. /Dist. by Universal Uclick
YOU'RE THINKING LIKE A WORKER BEE. THERE'S NO TIME FOR SUBSTANCE WHEN YOU'RE AT THE TOP.
EXECUTIVES ONLY RESPOND TO FAMILIAR COLORS AND SHAPES.
CLOUDS, DOLLAR SIGNS. . .THAT SORT OF THING.
. . . AND IN CONCLUSION.
$
www.dilbert.com
1-31-14
COME ON SLOW CLAP.

DCMB BioComputing BootCamp
Day 3, Lecture 3:
Data Exploration and Visualization in R

Armand Bankhead

bankhead@umich.edu

8/22/2018

Now What?

- Scenario:
  - You understand the fundamentals of R
  - You've read your data into an R data frame
- During this session we will talk about

1. Basic data summarization
2. Visualize data with plots

Overview

1. Summarizing Data in R
2. Creating plots in R Using ggplot2

Example Data Set:

Cancer Research

Home About Articles For Authors Alerts News

Molecular and Cellular Pathobiology

Activation of Wnt/β-Catenin in Ewing Sarcoma Cells Antagonizes EWS/ETS Function and Promotes Phenotypic Transition to More Metastatic Cell States

Elisabeth A. Pedersen, Rajasree Menon, Kelly M. Bailey, Dafydd G. Thomas, Raelene A. Van Noord, Jenny Tran, Hongwei Wang, Ping Ping Qu, Antje Hoering, Eric R. Fearon, Rashmi Chugh, and Elizabeth R. Lawlor

DOI: 10.1158/0008-5472.CAN-15-3422 Published September 2016

- Ewing's sarcoma: rare bone and soft tissue cancer occurring in children and teenagers
  - 70-80% survival
- *In vitro* CHLA25-7TGP ES cells stimulated to over-express WNT3A
- RNA-Seq profiling used to quantify gene expression

Download pedersenLog2RPKM20180817.txt and pedersenLog2_matrixRPKM20180817.txt from the Day3 course website.

Exercise: Write a Script to Read Pedersen Gene Expression Data into a Data Frame

Do it Yourself!

1. Download both Pedersen data files
2. Use setwd() to move to the data file folder
3. options(stringsAsFactors=F)
4. Use the read.delim() function to read in "pedersenLog2RPKM20180817.txt" file into a data frame called "data1"
5. Use the head() and dim() function to find out about the structure of this data file

How many rows? What are the columns?

Do it Yourself!

Quickly Calculate Simple Statistics

- R has many built in statistical functions that use fast vector and matrix operations
- No need to write a for loop, sum, and then divide by n
- Just provide a vector of data to the mean() function:

```
> mean(data1$log2RPKM)
```

- With one line of code you have take the mean of 97,000 values!

Exercise: Use mean(), median(), max(), min(), summary() functions on the Pedersen data

Using the table() Function

- When exploring new datasets it is often useful to count the number of values
- table() can be used to build a contingency table of the counts of each value
  - For one column:

```
> table(data1$tx)

control   WNT3A
  48585   48585
```

  - For multiple columns:

```
> table(data1$tx,data1$rep)

             1     2     3
control 16195 16195 16195
WNT3A   16195 16195 16195
```

Using the aggregate() Function

Do it Yourself!

- Often times we want to perform a function on subsets of our data
  - example question: What is the mean expression for each sample?
- aggregate() splits data into subsets, computes summary statistics for each and returns the result
- aggregate() takes several arguments:

formula input data

```
> aggregate(log2RPKM ~ sample, data1, FUN='mean')
        sample log2RPKM
1 control_rep1 2.598991
2 control_rep2 2.583161
3 control_rep3 2.579987
4   WNT3A_rep1 2.593578
5   WNT3A_rep2 2.583571
6   WNT3A_rep3 2.598349
```

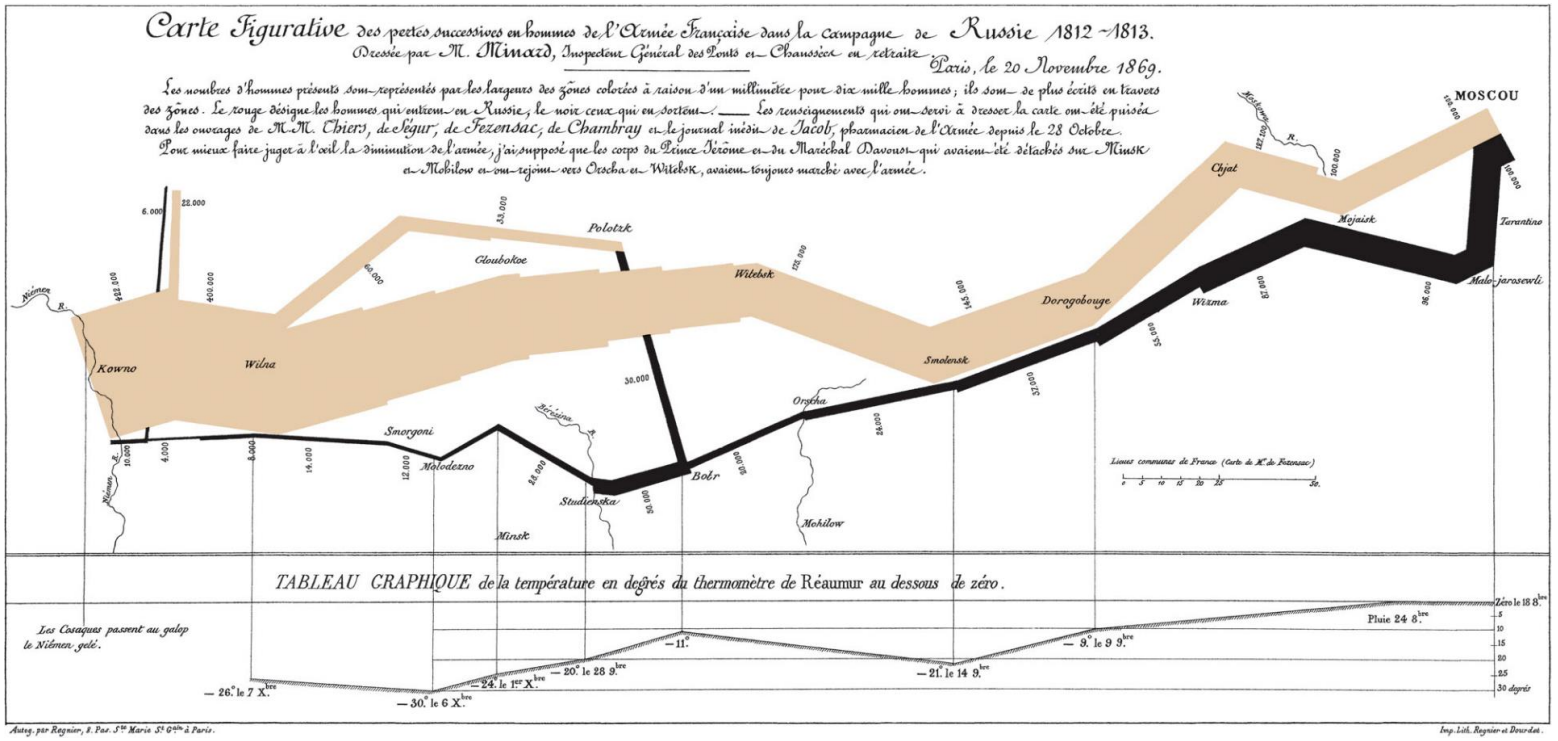
summary statistic

Exercise: Use aggregate() to calculate the maximum log2RPKM value per sample.

formula format: y ~ x1

- **y is a numeric value**
- **x1 is a grouping variable**
- **possible to specify multiple groups as x1 + x2 + …**

Data Visualization Allows Researchers to Visually Present Data



- Minard's 1869 diagram of Napoleonic France's invasion of Russia
  - Line width indicates size of army
  - Color indicates army's course to and from Russia

Data Visualization Allows Researchers to Visually Present Data

- Data visualizations should:
  - Show the data
  - Avoid distorting the data
  - Present many numbers in a small space
  - Make large data sets coherent
  - Serve a reasonably clear purpose
  - Be closely integrated with the statistical and verbal descriptions of a data set

https://en.wikipedia.org/wiki/Data_visualization

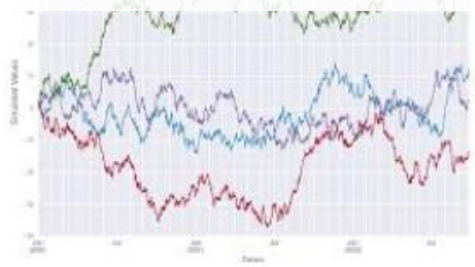
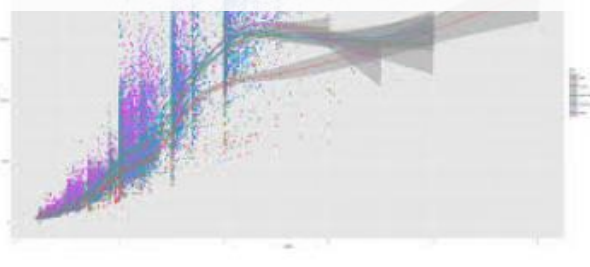
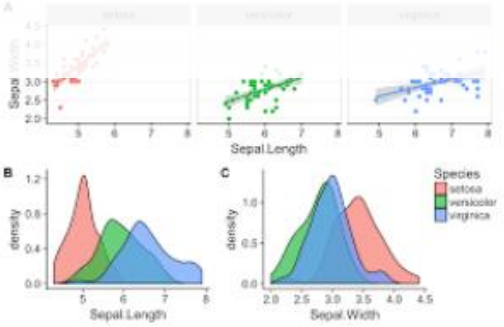
R Base Graphics Versus ggplot2

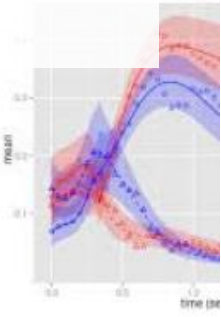
- R comes with “base graphics” built in to support commonly used data visualizations
- Today we will focus on using an alternative data visualization framework called ggplot2
- ggplot2 is an external package that must be downloaded, installed, and loaded with the library command
- A common practice is to use ggplot2 to construct publication quality graphs but still use base graphics to quickly visualize data

https://flowingdata.com/2016/03/22/comparing-ggplot2-and-r-base-graphics/

ggplot2

- ggplot2 is an R data visualization package created by Hadley Wickham
  - One of the most popular R packages
  - Breaks up graphs construction into additive functions called layers
- ggplot2 documentation and cheat sheet: https://www.rdocumentation.org/packages/ggplot2/versions/3.0.0



Creating a Visualization with ggplot2

- ggplot2 visualization function calls consist of several basic components:

1. ggplot()
2. geom_XXX()
3. optional layers
4. ggsave()

- Multiple function calls are combined together using "layers"
- aes(thetic) functions are used to map input data to plot features (e.g. x axis, y axis, colors)

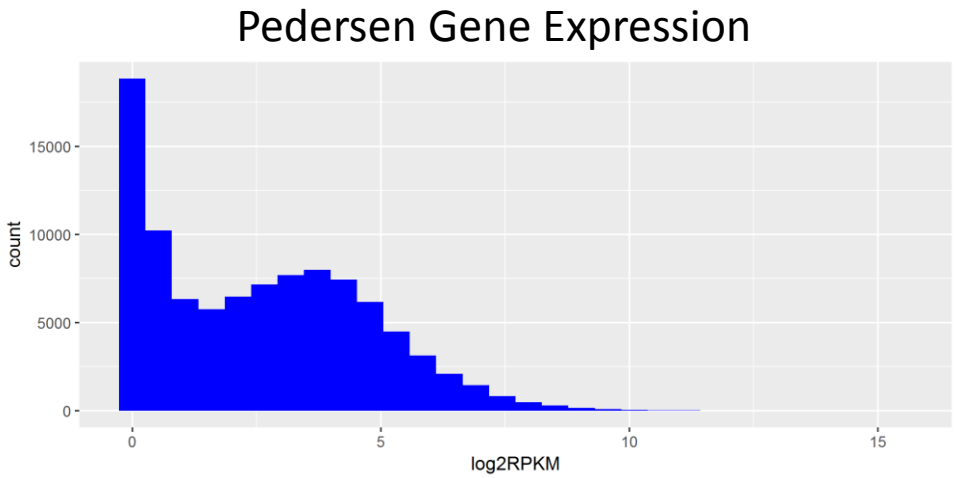
```
options(stringsAsFactors=F)

library(ggplot2)

inFile = 'data.txt'
data1 = read.delim(inFile)

ggplot(data1,aes(x = log2RPKM)) +
 geom_histogram()
ggsave('histogram.png')
```

input data frame

aesthetic

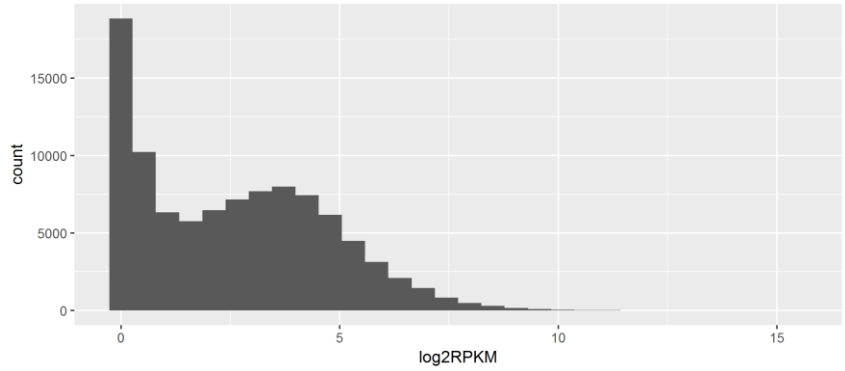
Visualizing Data Using Histograms

- histogram: a type of bar graph visualization in which data measurements are counted based on value
  - For **discrete** measures it shows the frequency of values in each category
  - For **continuous** measure it shows the frequency of values occurring in small intervals covering the whole range



How to Create a Basic Histogram Using ggplot2

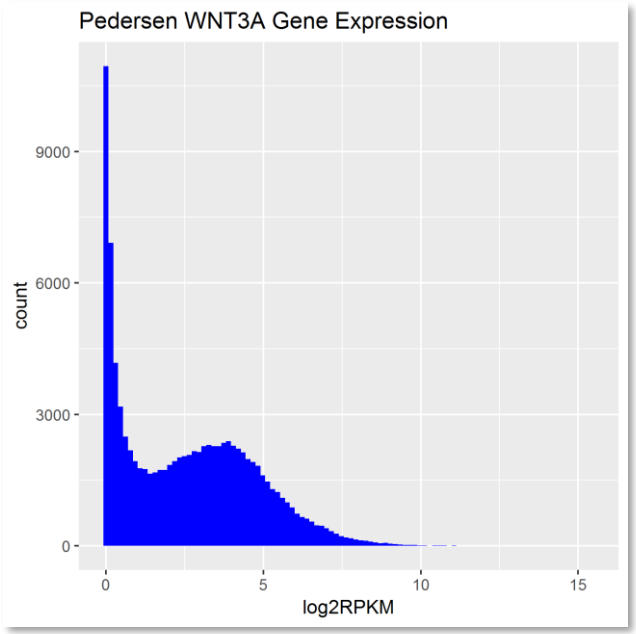
```
options(stringsAsFactors=F)

library(ggplot2)

inFile1 = 'pedersenLog2RPKM20180817.txt'
data1 = read.delim(inFile1)

ggplot(data1,aes(x = log2RPKM)) +
  geom_histogram()
ggsave('histogram1.png')
```



Do it Yourself!

Exercise: Create a histogram using the code from the previous page and update your visualization to:

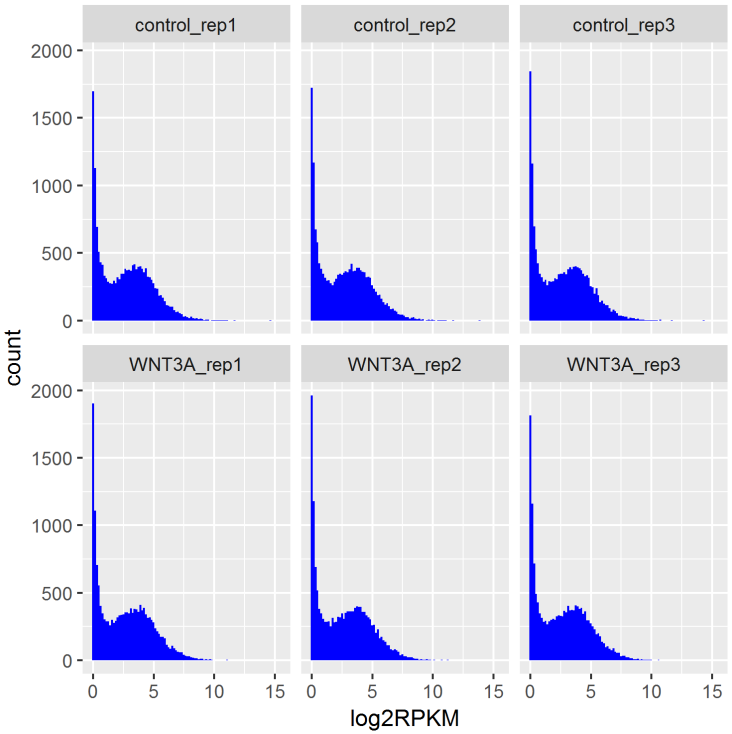
1. Change the color
   - HINT: ?geom_histogram
   - HINT: fill = “blue”
2. Set the image width and height to be 5 inches
   - HINT: ?ggsave
   - HINT: height = 5
3. Adjust the number of bins to 100
4. Add a title
   - HINT: ?labs



Do it Yourself!

Create Sub-plots Using Facets

- Sub-plots can easily be created using facet layers:
  - facet_wrap()
  - facet_grid()



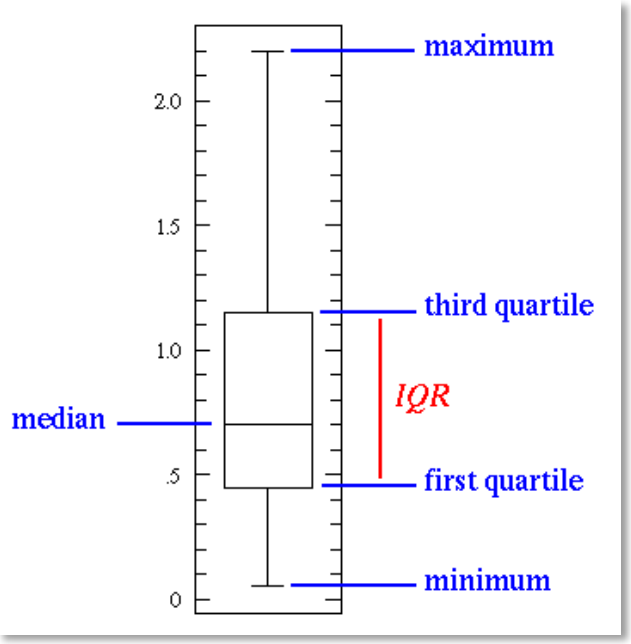
Exercise: Update your histogram visualization to facet on sample

1. Add a facet_wrap() layer
   - HINT: + facet_wrap(~sample)

Visualizing Data Using Boxplots

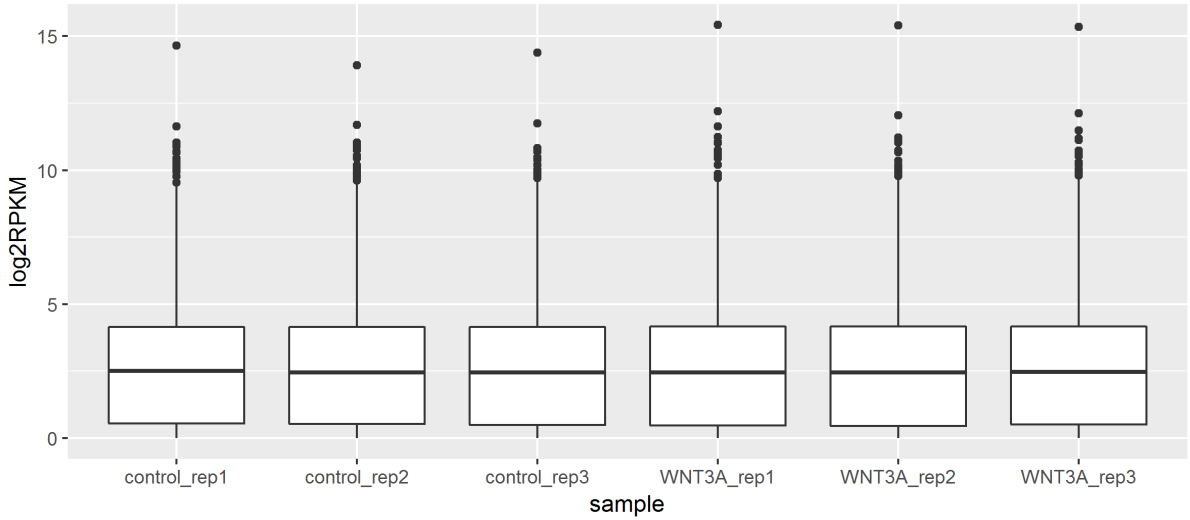
- boxplot: graphically represents data distributions using quartiles
- box-and-whisker plot: includes boxplots with lines extending from boxes to indicate variability outside the upper and lower quartiles

- Why it is useful?
  - Summarize the main characteristics of the data: Mean/median, quartile, spread, symmetry and outliers.
  - Efficient – less complicated than histogram
  - Allows us to represent multiple data distributions in the same graph



How to Create a Boxplot Using ggplot2

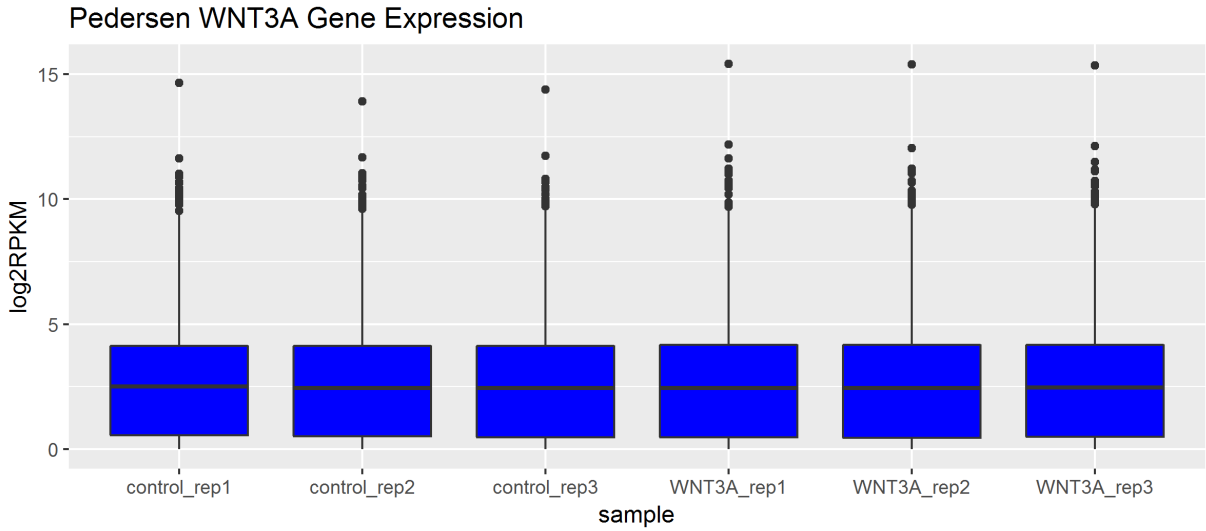
```
options(stringsAsFactors=F)

library(ggplot2)

inFile1 = 'pedersenLog2RPKM20180817.txt'
data1 = read.delim(inFile1)

ggplot(data1,aes(x = log2RPKM)) +
  geom_boxplot()
ggsave('boxplot1.png')
```



Do it Yourself!

Exercise: Create a boxplot using the code from the previous page and update your visualization to:

1. Change the fill color
   - HINT: fill = 'blue'
2. Add a title
   - HINT: ?labs

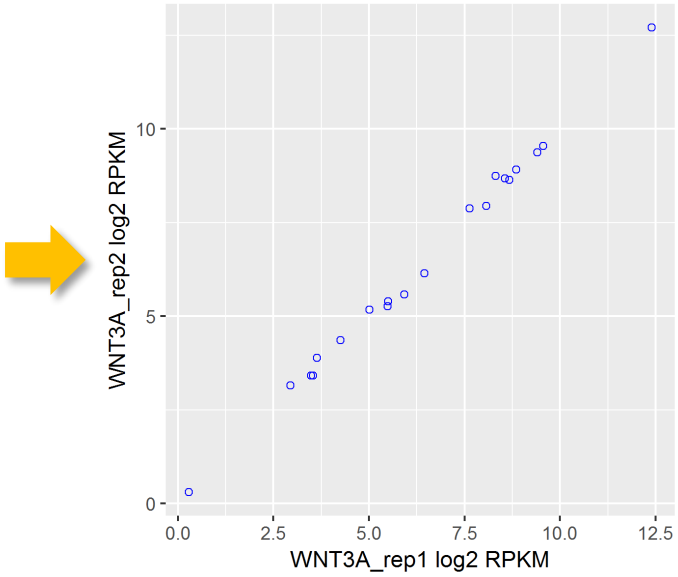


If time: Try creating a violin plot using geom_violin()

Visualizing Data Using Scatter Plots

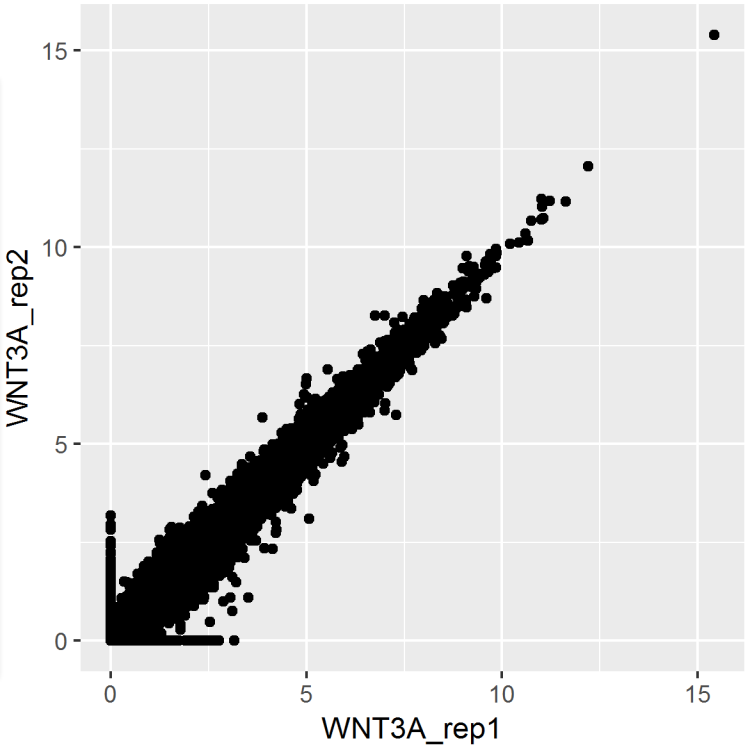
- Scatter Plots are visualizations that display two data values for the same measurement
  - example: two sample replicates expression values for each gene
- Data points that are not on the diagonal indicate disagreement
- We expect strong agreement between sample replicates

gene	WNT3A_rep1	WNT3A_rep2
A1BG	2.38	1.64
A1BG-AS1	0.83	0.58
A1CF	0.02	0.00
A2M	0.30	0.67
A2M-AS1	0.09	0.10
A2ML1	0.33	0.73
A2MP1	0.00	0.00
A4GALT	4.20	4.82
A4GNT	0.00	0.00
AAAS	5.57	5.53
AACS	2.70	2.52
AACSP1	0.13	0.15
AADAC	0.12	0.00
AADACL2	0.00	0.00
AADACL4	2.30	2.03
AADAT	0.94	1.19
AAED1	1.20	1.51
AAGAB	4.36	4.29
AAK1	1.68	1.97
AAMDC	4.41	3.78



How to Create a Scatter Plot Using ggplot2

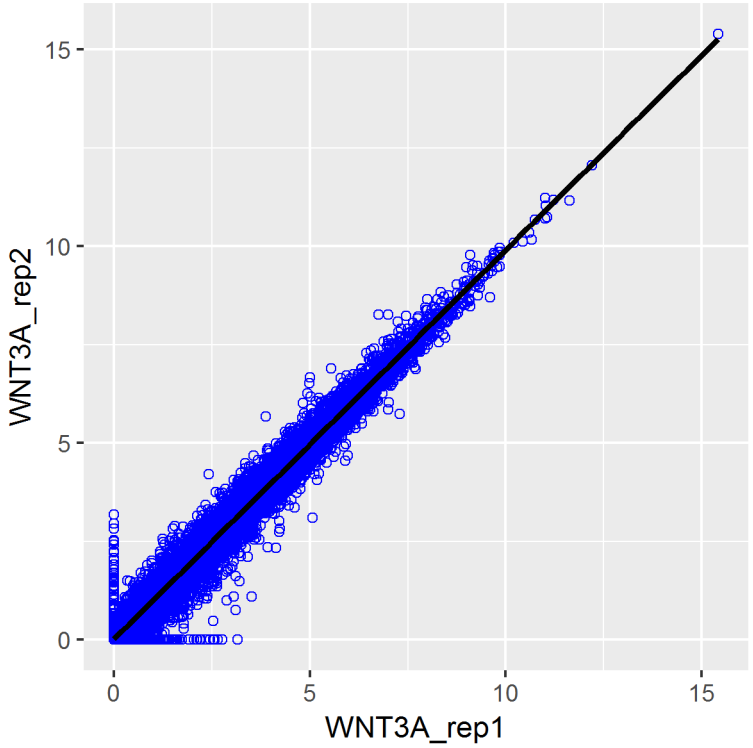
```
options(stringsAsFactors=F)

library(ggplot2)

inFile2 = 'pedersenLog2RPKM_matrix20180817.txt'
data2 = read.delim(inFile2)

ggplot(data2,aes(x = WNT3A_rep1, y = WNT3A_rep2)) +
  geom_point()
ggsave('scatter2.png')
```



Do it Yourself!

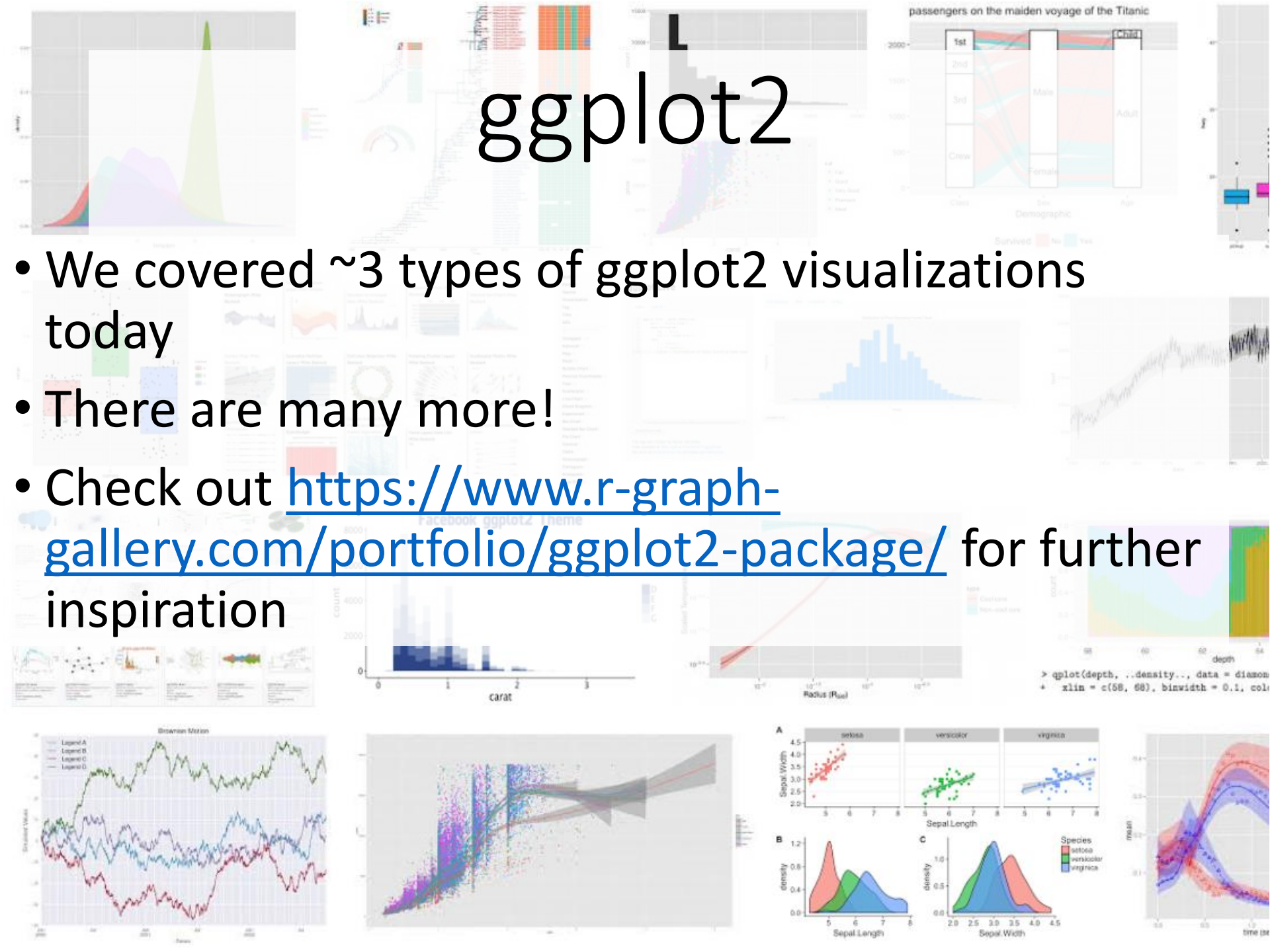
Exercise: Create a boxplot using the code from the previous page and update your visualization to:

1. Change the color and shape of scatter plot points
   - HINT: color = 'blue'
   - HINT: shape = 1
2. Add a black linear regression line using a geom_smooth layer
   - HINT: geom_smooth(method = lm)



ggplot2

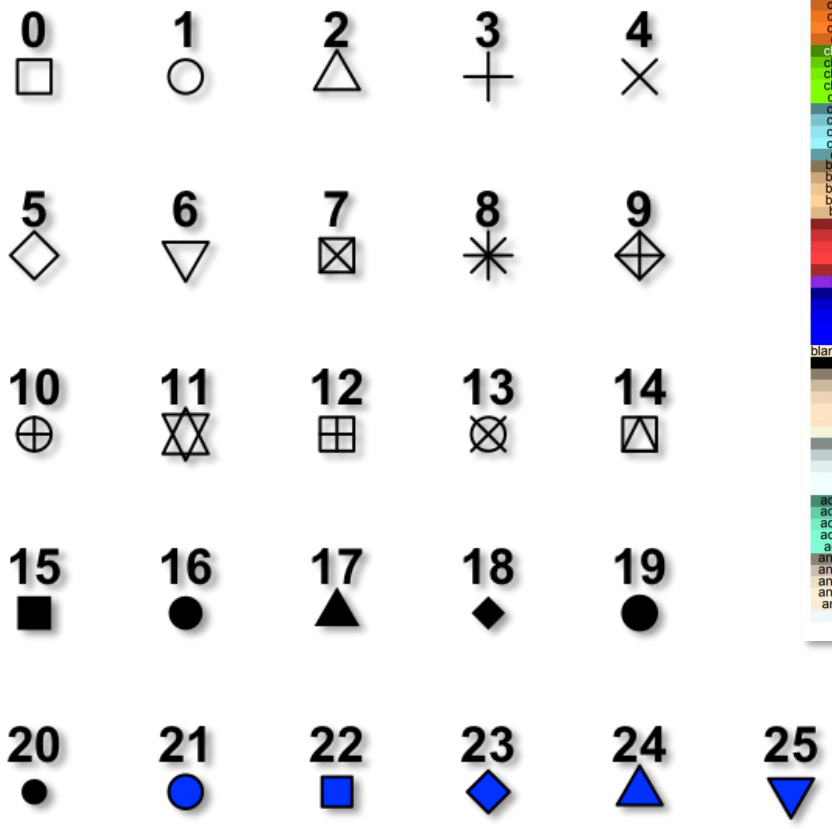
- We covered ~3 types of ggplot2 visualizations today
- There are many more!
- Check out https://www.r-graph-gallery.com/portfolio/ggplot2-package/ for further inspiration



References

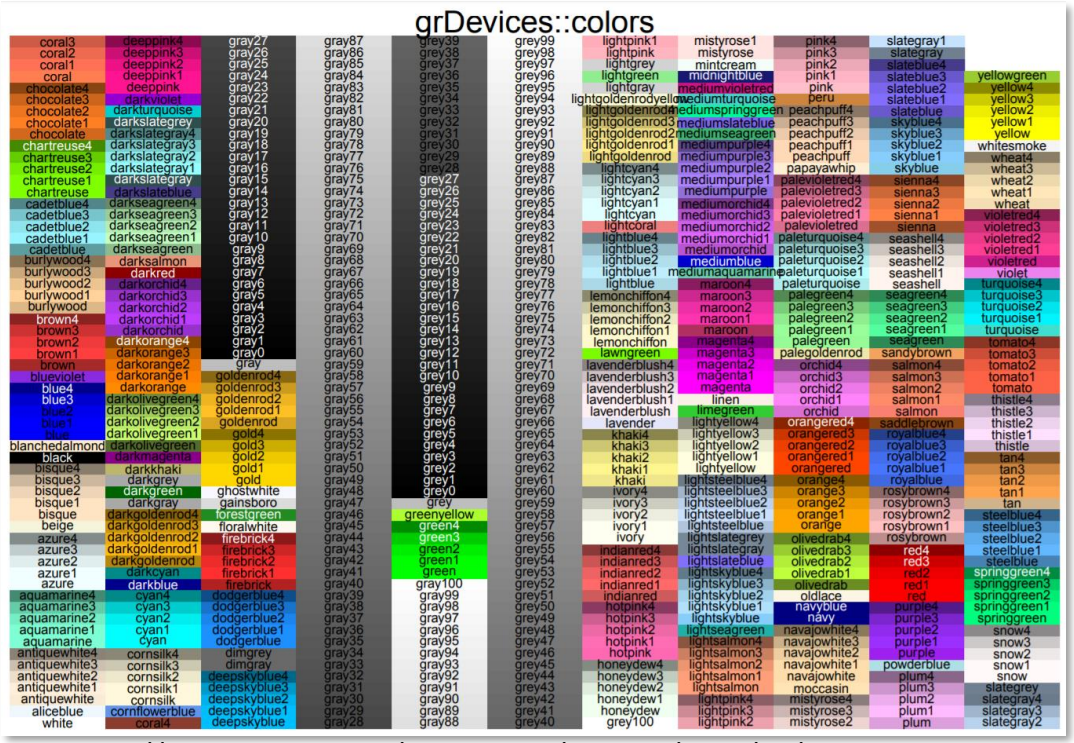
- Gentleman, Robert. R Programming for Bioinformatics. CRC Press, 2009.
- Slides Partially Sourced from Jacob Kitzman and Barry Grant

R Graphics Shapes



http://www.sthda.com/english/wiki/r-plot-pch-symbols-the-different-point-shapes-available-in-r

R Graphics Colors



http://bc.bojanorama.pl/wp-content/uploads/2013/04/rcolorsheet.pdf