

DilbertCartoonist@gmail.com

© 2014 Scott Adams, Inc./Dilbert by Universal Uclick

www.dilbert.com  
1-21-14

# 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
3. Hierarchical clustering

# Example Data Set:

## Cancer Research

[Home](#) [About](#) [Articles](#) [For Authors](#) [Alerts](#) [News](#)

---

Molecular and Cellular Pathobiology

### Activation of Wnt/ $\beta$ -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

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

1. `options(stringsAsFactors=F)`
2. Use the `read.delim()` function to read in “pedersenLog2RPKM20180817.txt” file into a data frame called “data1”
3. Use the `head()` and `dim()` function to find out about the structure of this data file

How many rows? What are the columns?

# 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

- 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
  
```

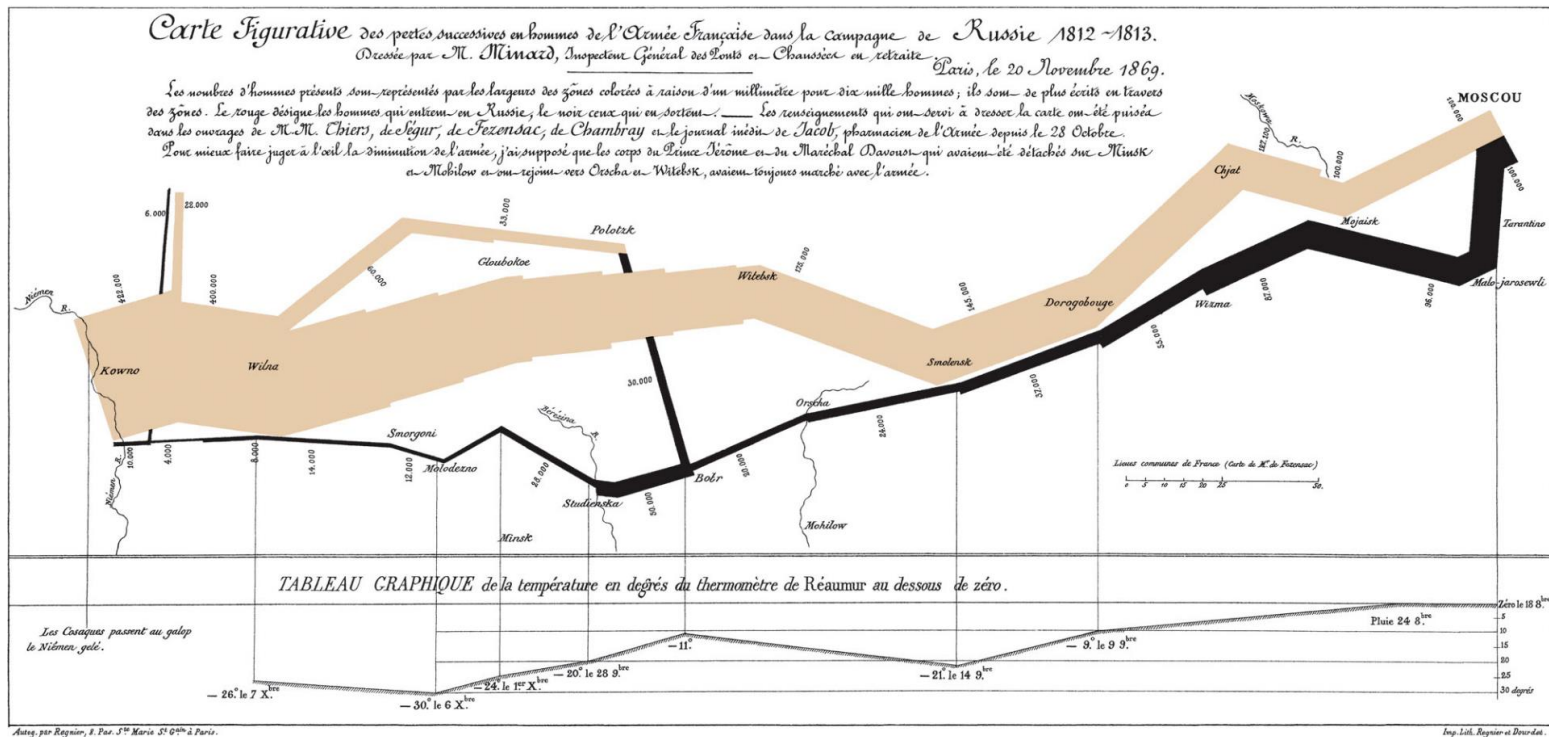
summary statistic

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

formula format:  $y \sim x1$

- $y$  is a numeric value
- $x1$  is a grouping variable
- possible to specify multiple groups as  $x1 + x2 + \dots$

# 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

# 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

# ggplot2

- ggplot2 is an R data visualization package created by Hadley Wickham
  - One of the most popular R packages
  - an alternative to R base graphics which are built into R
  - Breaks up graphs construction into additive functions
- 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()` 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()
```

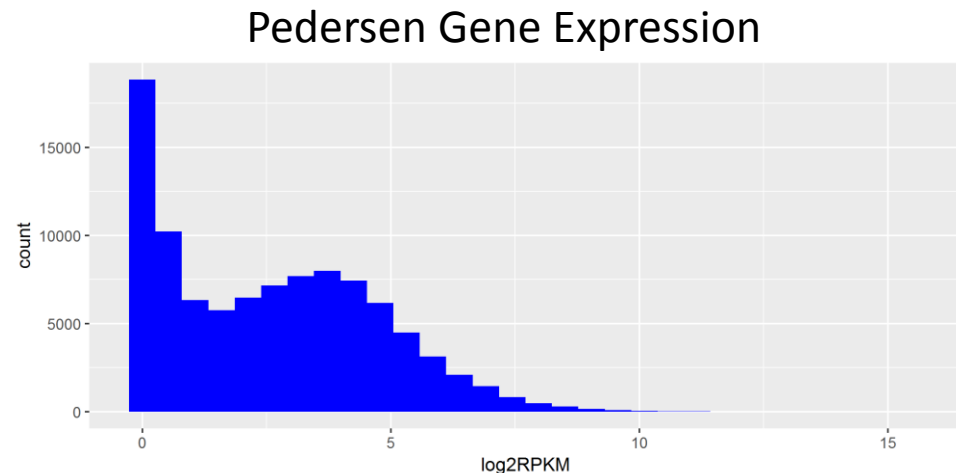
```
ggsave('histogram.png')
```

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



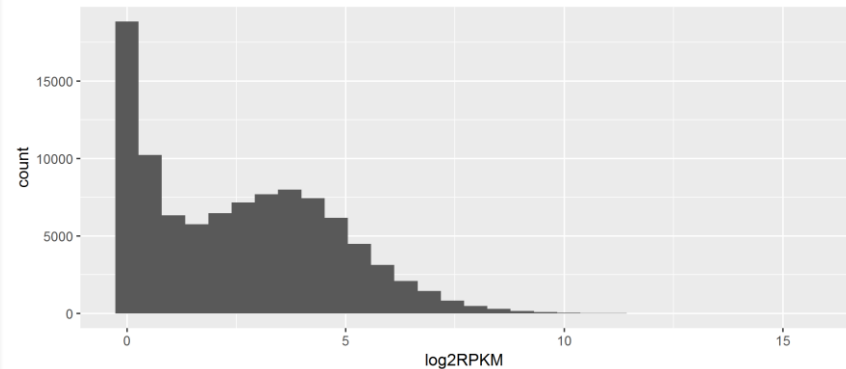
# How to Create a Basic Histogram Using ggplot2

```
options(stringsAsFactors=F)

library(ggplot2)

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

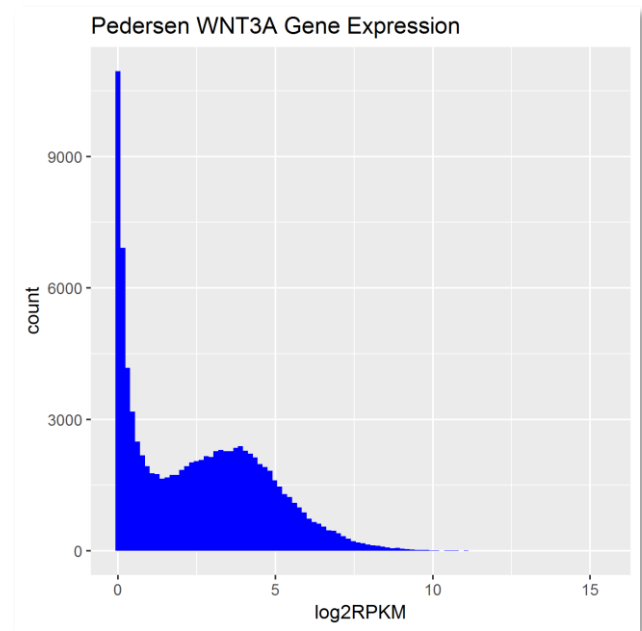
ggplot(data1,aes(x = log2RPKM)) +
  geom_histogram()
ggsave('histogram1.png')
```





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

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`

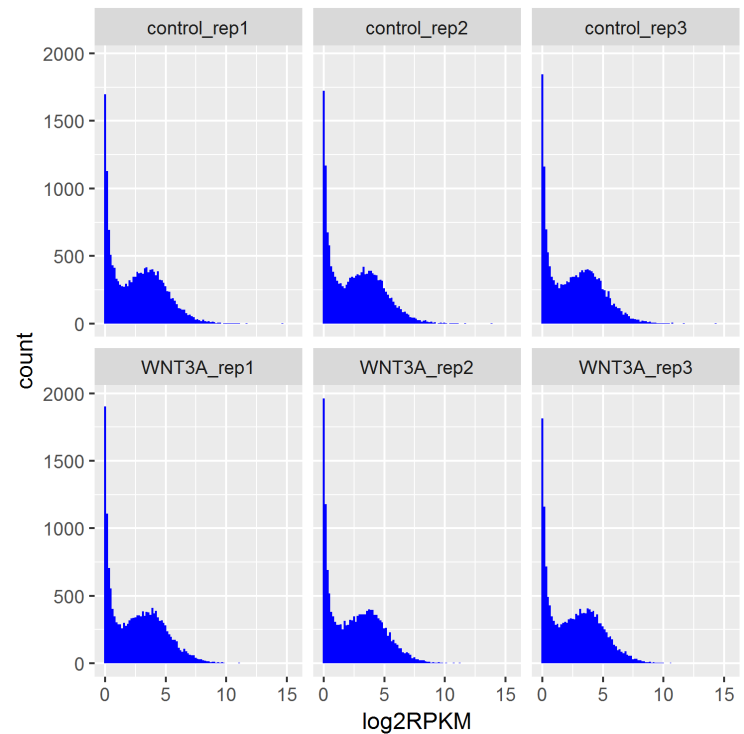


# Create Sub-plots Using Facets

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

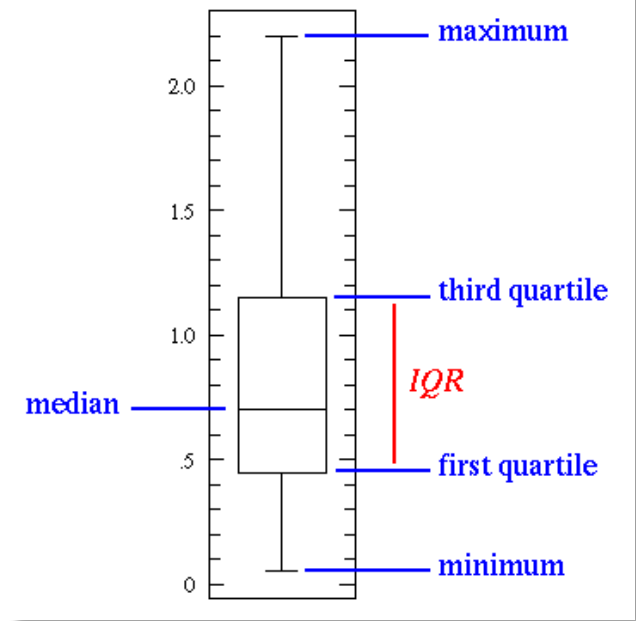
**Exercise: Update your histogram visualization to facet on sample**

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



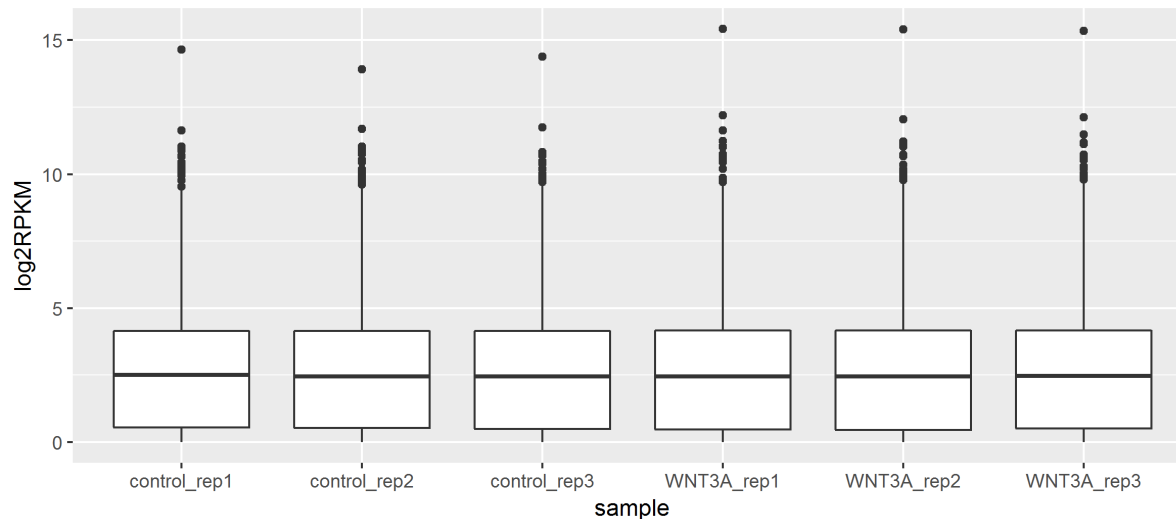
# How to Create a Boxplot Using ggplot2

```
options(stringsAsFactors=F)

library(ggplot2)

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

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



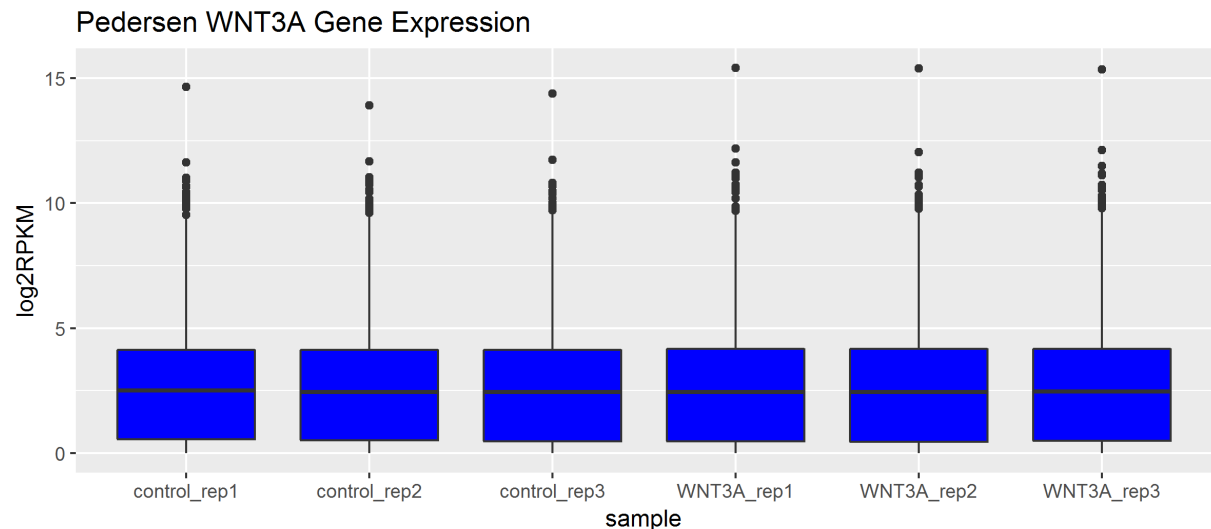
Exercise: Create a boxplot using the code from the previous page and update your visualization to have the following features:

1. Change the fill color

- HINT: `fill = 'blue'`

2. Add a title

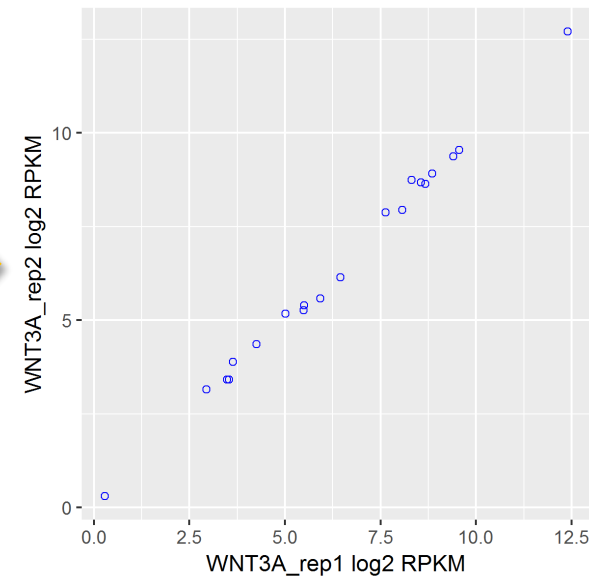
- HINT: `?labs`



# 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



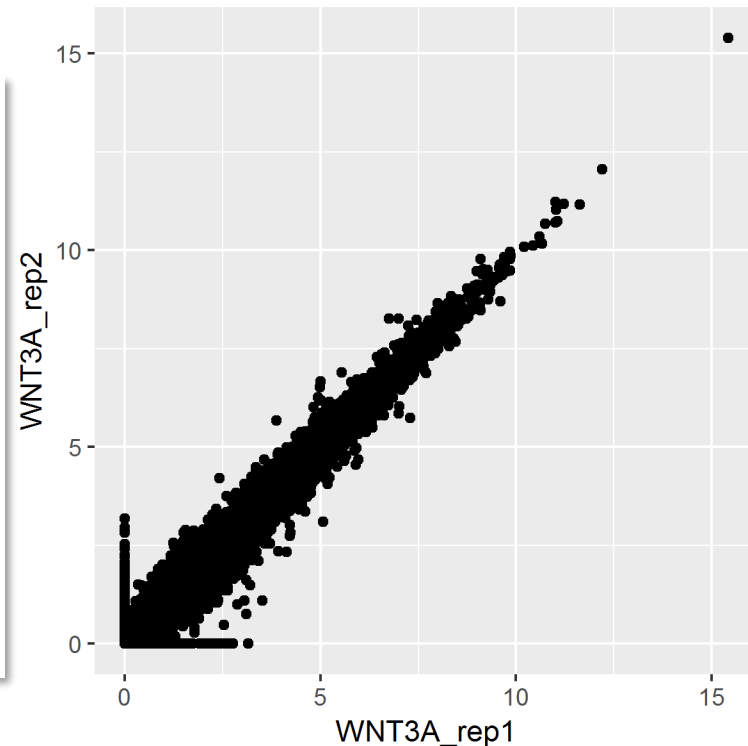
# How to Create a Scatter Plot Using ggplot2

```
options(stringsAsFactors=F)

library(ggplot2)

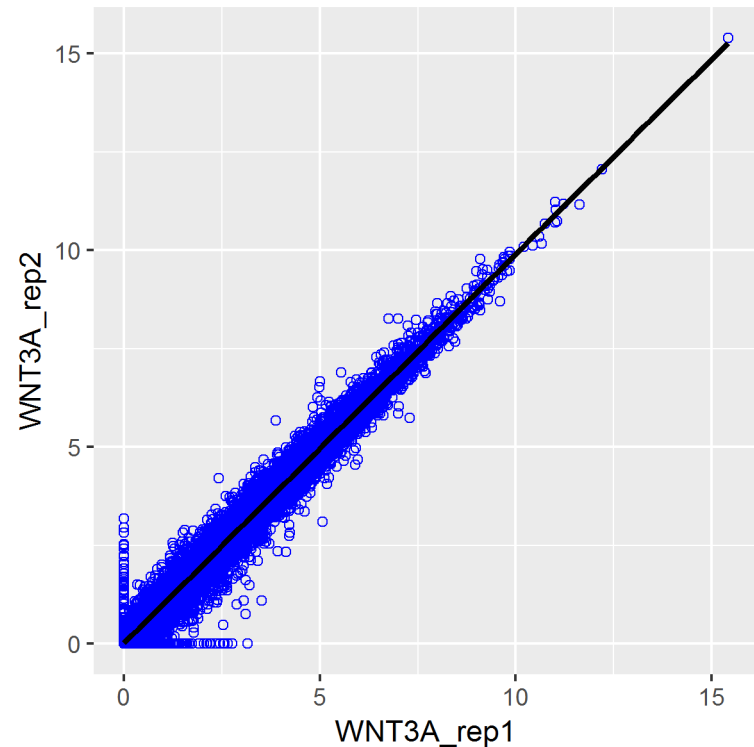
inFile2 = 'pedersenLog2RPKM_matrix20180817.txt'
data2 = read.delim(inFile2)

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



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

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)`

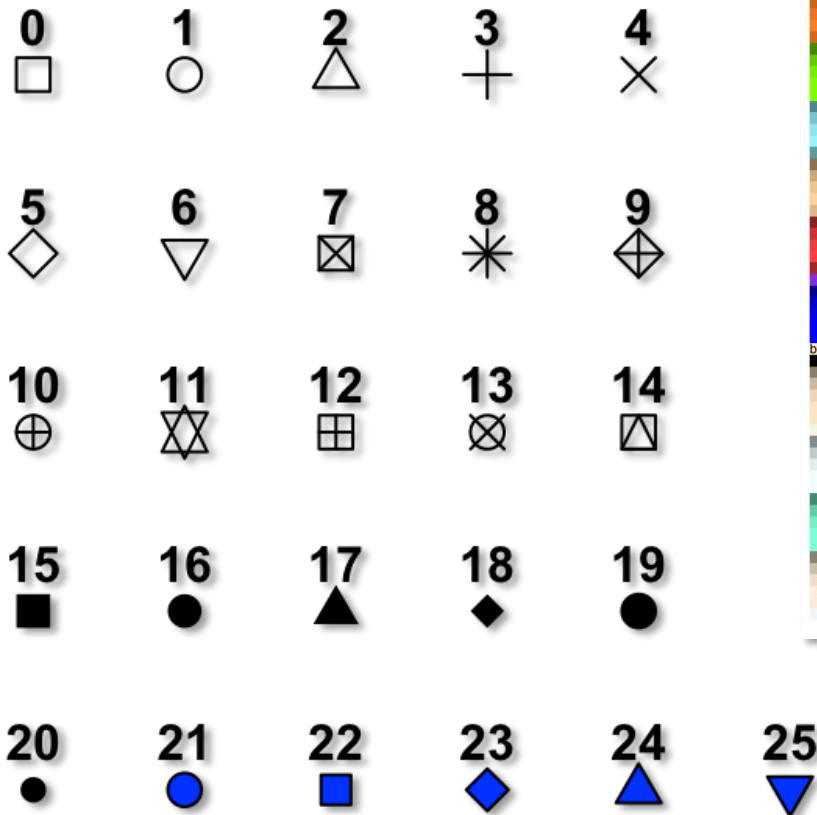




# References

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

## R Graphics Shapes



## R Graphics Colors

grDevices::colors																			
coral3	deeppink4	gray27	gray87	gray98	lightpink1	mistyrose1	pink4	slategray1											
coral2	deeppink3	gray26	gray86	gray97	lightpink2	mistyrose2	pink3	slategray2											
coral1	deeppink2	gray25	gray85	gray96	lightpink3	mistyrose3	pink2	slategray3											
coral	deeppink1	gray24	gray84	gray95	lightpink4	mistyrose4	pink1	slategray4											
chocolate4	darkviolet	gray23	gray83	gray94	lightgreen	mediumslateblue	peru	slateblue1	yellowgreen										
chocolate3	darkturquoise	gray22	gray82	gray93	lightgoldenrod1	mediumslateblue	peachpuff4	slateblue2	yellow										
chocolate2	darkslategray4	gray21	gray81	gray92	lightgoldenrod2	mediumslateblue	peachpuff3	slateblue3	yellow2										
chocolate1	darkslategray3	gray20	gray80	gray91	lightgoldenrod3	mediumslateblue	peachpuff2	slateblue4	yellow3										
chocolate	darkslategray2	gray19	gray79	gray90	lightgoldenrod4	mediumslateblue	peachpuff1	slateblue5	yellow4										
chartreuse4	darkslategray1	gray18	gray78	gray89	lightcyan4	mediumslateblue	peachpuff	skyblue1	whitesmoke										
chartreuse3	darkslategray2	gray17	gray77	gray88	lightcyan3	mediumslateblue	peachpuff	skyblue2	wheat4										
chartreuse2	darkslategray3	gray16	gray76	gray87	lightcyan2	mediumslateblue	peachpuff	skyblue3	wheat3										
chartreuse1	darkslategray4	gray15	gray75	gray86	lightcyan1	mediumslateblue	peachpuff	skyblue4	wheat2										
cadetblue4	darkseagreen4	gray14	gray74	gray85	lightcyan2	mediumslateblue	peachpuff	skyblue5	wheat1										
cadetblue3	darkseagreen3	gray13	gray73	gray84	lightcyan1	mediumslateblue	peachpuff	skyblue6	wheat										
cadetblue2	darkseagreen2	gray12	gray72	gray83	lightcyan2	mediumslateblue	peachpuff	skyblue7											
cadetblue1	darkseagreen1	gray11	gray71	gray82	lightcyan1	mediumslateblue	peachpuff	skyblue8											
cadetblue	darkseagreen	gray10	gray70	gray81	lightcyan2	mediumslateblue	peachpuff	skyblue9											
burlywood4	darksalmon	gray9	gray69	gray80	lightblue4	mediumslateblue	peachpuff	skyblue10											
burlywood3	darkred	gray8	gray68	gray79	lightblue3	mediumslateblue	peachpuff	skyblue11											
burlywood2	darkorchid4	gray7	gray67	gray78	lightblue2	mediumslateblue	peachpuff	skyblue12											
burlywood1	darkorchid3	gray6	gray66	gray77	lightblue1	mediumslateblue	peachpuff	skyblue13											
burlywood	darkorchid2	gray5	gray65	gray76	lightblue	mediumslateblue	peachpuff	skyblue14											
brown4	darkorchid1	gray4	gray64	gray75	lightblue	mediumslateblue	peachpuff	skyblue15											
brown3	darkorchid2	gray3	gray63	gray74	lightblue	mediumslateblue	peachpuff	skyblue16											
brown2	darkorchid3	gray2	gray62	gray73	lightblue	mediumslateblue	peachpuff	skyblue17											
brown1	darkorchid4	gray1	gray61	gray72	lightblue	mediumslateblue	peachpuff	skyblue18											
brown	darkorange4	gray0	gray60	gray71	lightblue	mediumslateblue	peachpuff	skyblue19											
blues4	darkorange3	gray	gray59	gray70	lightblue	mediumslateblue	peachpuff	skyblue20											
blues3	darkorange2	gray	gray58	gray69	lightblue	mediumslateblue	peachpuff	skyblue21											
blues2	darkorange1	gray	gray57	gray68	lightblue	mediumslateblue	peachpuff	skyblue22											
blues1	darkorange	gray	gray56	gray67	lightblue	mediumslateblue	peachpuff	skyblue23											
blues	darkolivegreen4	gray	gray55	gray66	lightblue	mediumslateblue	peachpuff	skyblue24											
blanchedalmond	darkolivegreen3	gray	gray54	gray65	lightblue	mediumslateblue	peachpuff	skyblue25											
black	darkolivegreen2	gray	gray53	gray64	lightblue	mediumslateblue	peachpuff	skyblue26											
bisque4	darkolivegreen1	gray	gray52	gray63	lightblue	mediumslateblue	peachpuff	skyblue27											
bisque3	darkolivegreen	gray	gray51	gray62	lightblue	mediumslateblue	peachpuff	skyblue28											
bisque2	darkolivegreen	gray	gray50	gray61	lightblue	mediumslateblue	peachpuff	skyblue29											
bisque1	darkolivegreen	gray	gray49	gray60	lightblue	mediumslateblue	peachpuff	skyblue30											
beige	darkolivegreen	gray	gray48	gray59	lightblue	mediumslateblue	peachpuff	skyblue31											
azure4	darkgoldenrod4	gray	gray47	gray58	lightblue	mediumslateblue	peachpuff	skyblue32											
azure3	darkgoldenrod3	gray	gray46	gray57	lightblue	mediumslateblue	peachpuff	skyblue33											
azure2	darkgoldenrod2	gray	gray45	gray56	lightblue	mediumslateblue	peachpuff	skyblue34											
azure1	darkgoldenrod1	gray	gray44	gray55	lightblue	mediumslateblue	peachpuff	skyblue35											
azure	darkgoldenrod	gray	gray43	gray54	lightblue	mediumslateblue	peachpuff	skyblue36											
aquamarine4	darkcyan	gray	gray42	gray53	lightblue	mediumslateblue	peachpuff	skyblue37											
aquamarine3	darkcyan	gray	gray41	gray52	lightblue	mediumslateblue	peachpuff	skyblue38											
aquamarine2	darkcyan	gray	gray40	gray51	lightblue	mediumslateblue	peachpuff	skyblue39											
aquamarine1	darkcyan	gray	gray39	gray50	lightblue	mediumslateblue	peachpuff	skyblue40											
aquamarine	darkcyan	gray	gray38	gray49	lightblue	mediumslateblue	peachpuff	skyblue41											
antiquewhite4	cornsilk4	gray	gray37	gray48	lightblue	mediumslateblue	peachpuff	skyblue42											
antiquewhite3	cornsilk3	gray	gray36	gray47	lightblue	mediumslateblue	peachpuff	skyblue43											
antiquewhite2	cornsilk2	gray	gray35	gray46	lightblue	mediumslateblue	peachpuff	skyblue44											
antiquewhite1	cornsilk1	gray	gray34	gray45	lightblue	mediumslateblue	peachpuff	skyblue45											
antiquewhite	cornsilk	gray	gray33	gray44	lightblue	mediumslateblue	peachpuff	skyblue46											
aliceblue	coral4	gray	gray32	gray43	lightblue	mediumslateblue	peachpuff	skyblue47											
white	coral3	gray	gray31	gray42	lightblue	mediumslateblue	peachpuff	skyblue48											

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