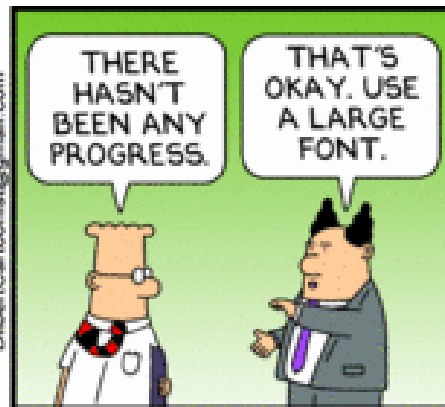
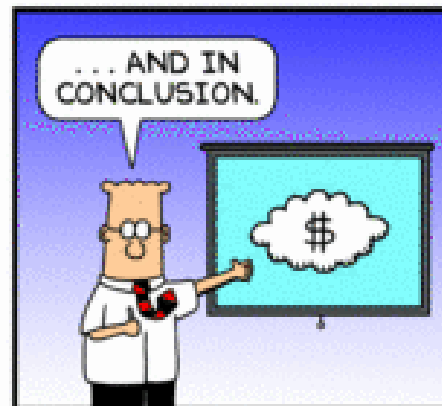
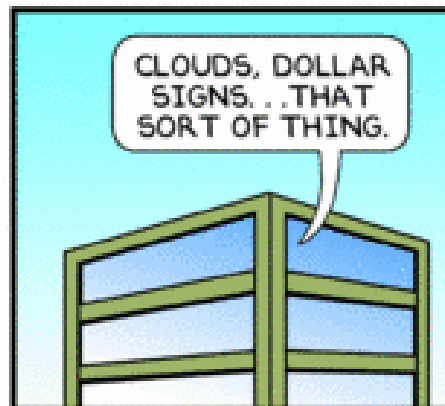
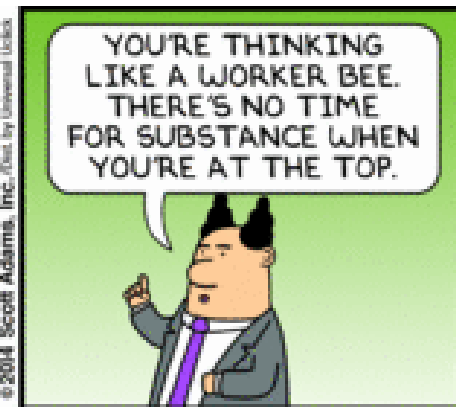




DilbertCartoonist@gmail.com



© 2014 Scott Adams, Inc. / Used 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 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

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?

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
  
```

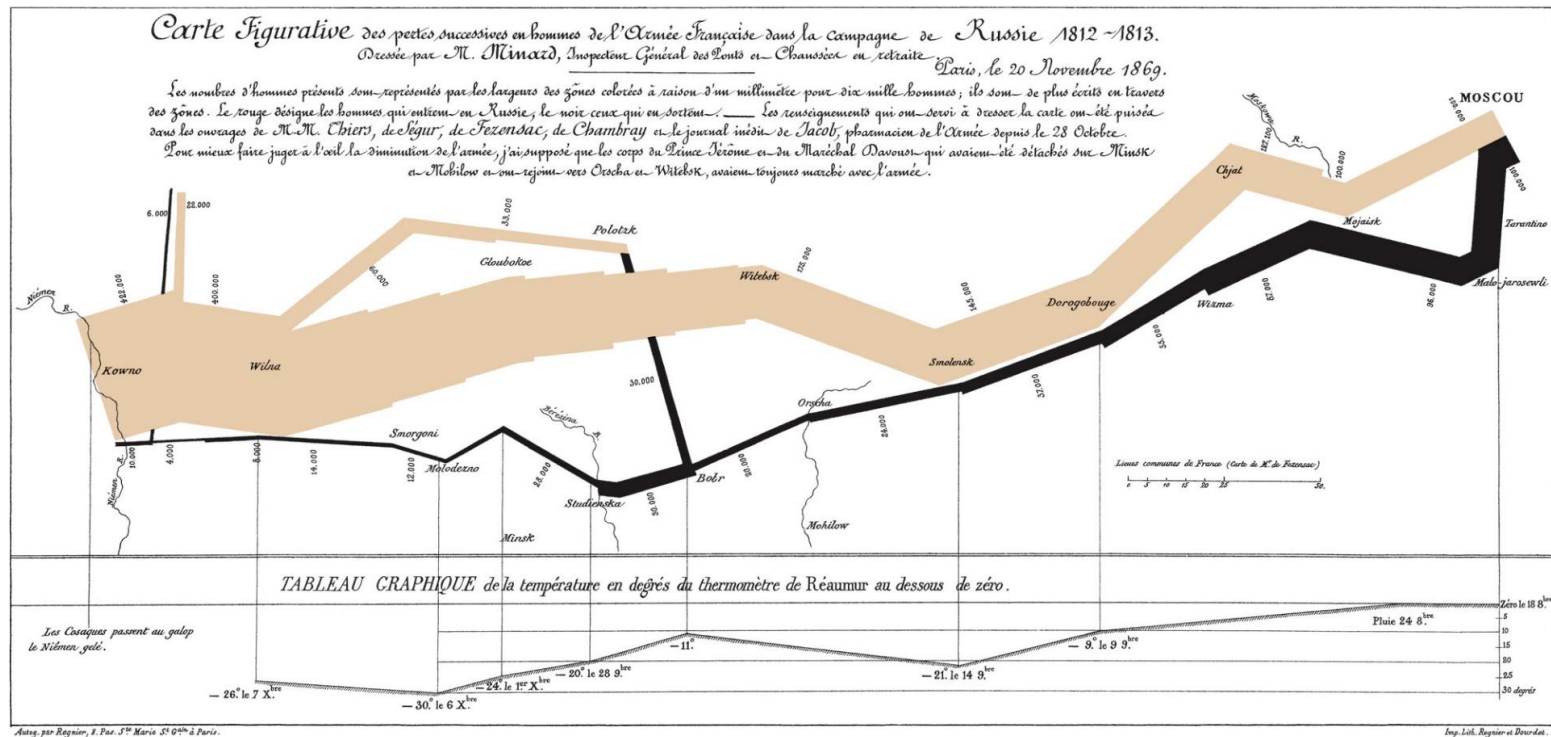
summary statistic

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
 - 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()` 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()
```

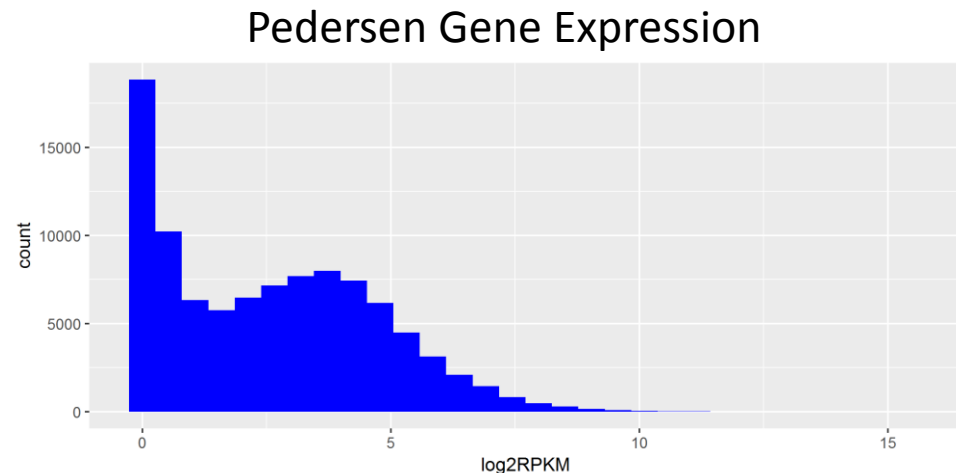
```
ggsave('histogram.png')
```

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



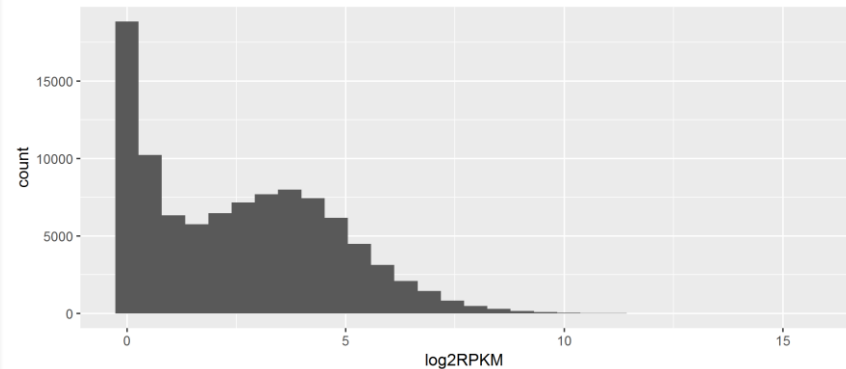
How to Create a Basic Histogram Using ggplot2

```
options(stringsAsFactors=F)

library(ggplot2)

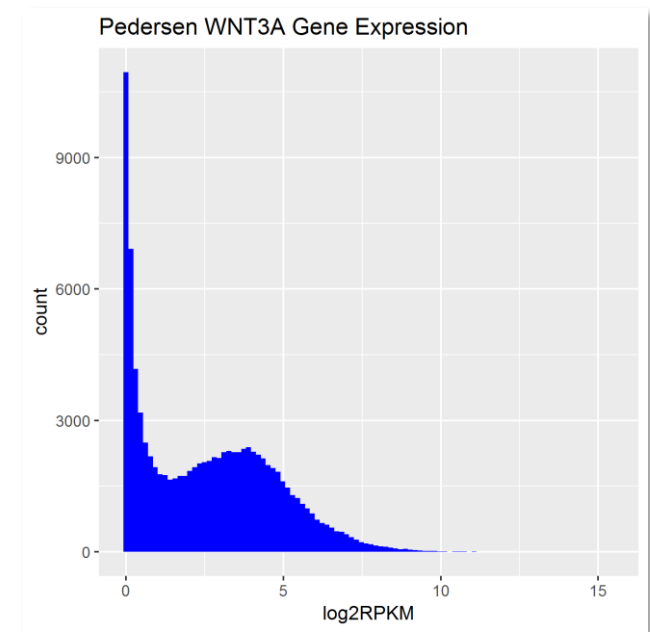
inFile1 = 'pedersenLog2RPKM20180817.txt'
data1 = read.delim(inFile1)

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:

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`

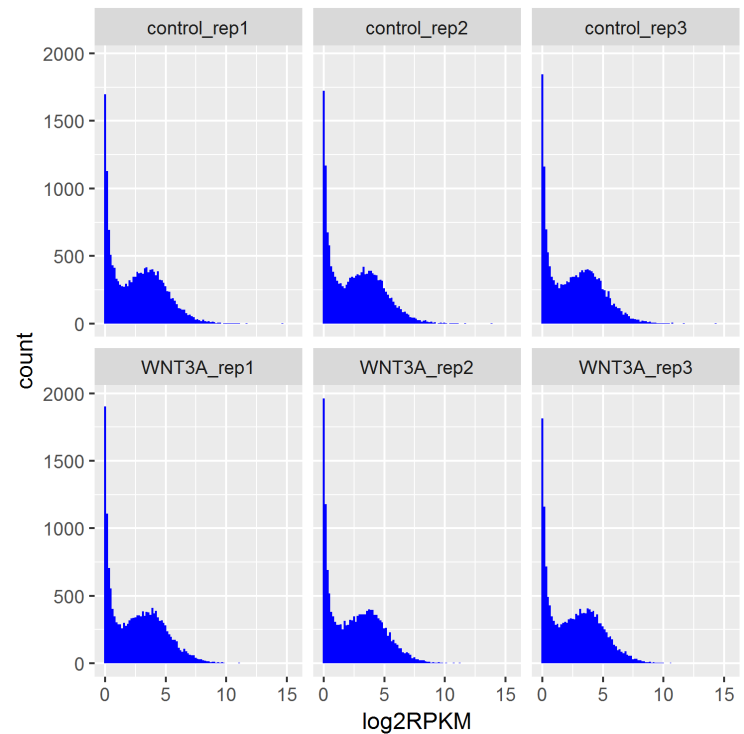


Create Sub-plots Using Facets

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

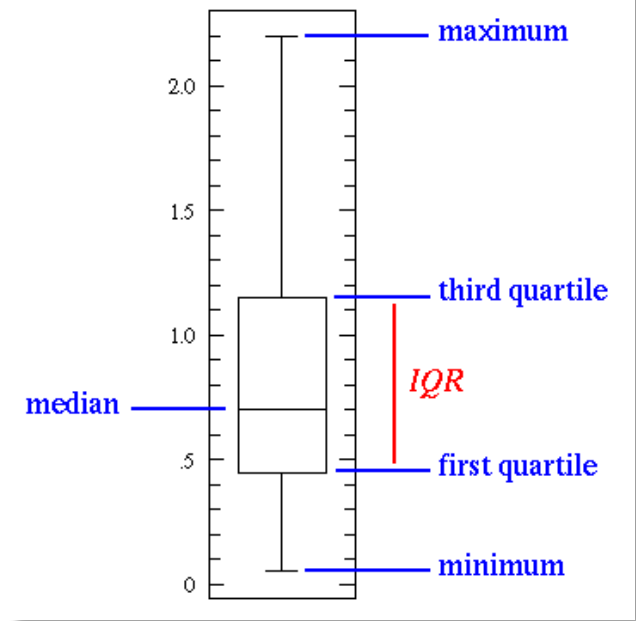
Exercise: Update your histogram visualization to facet on sample

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



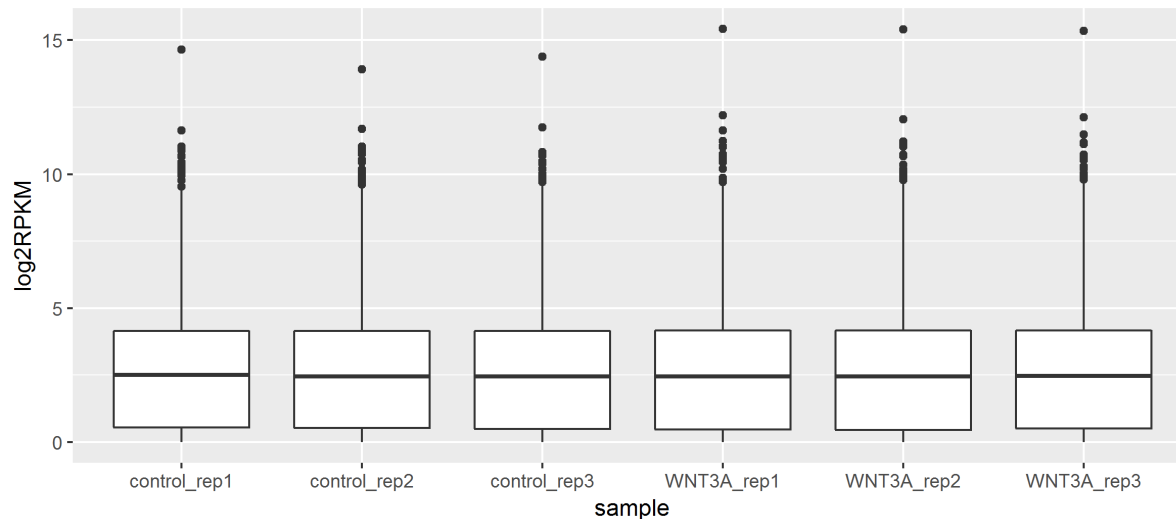
How to Create a Boxplot Using ggplot2

```
options(stringsAsFactors=F)

library(ggplot2)

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

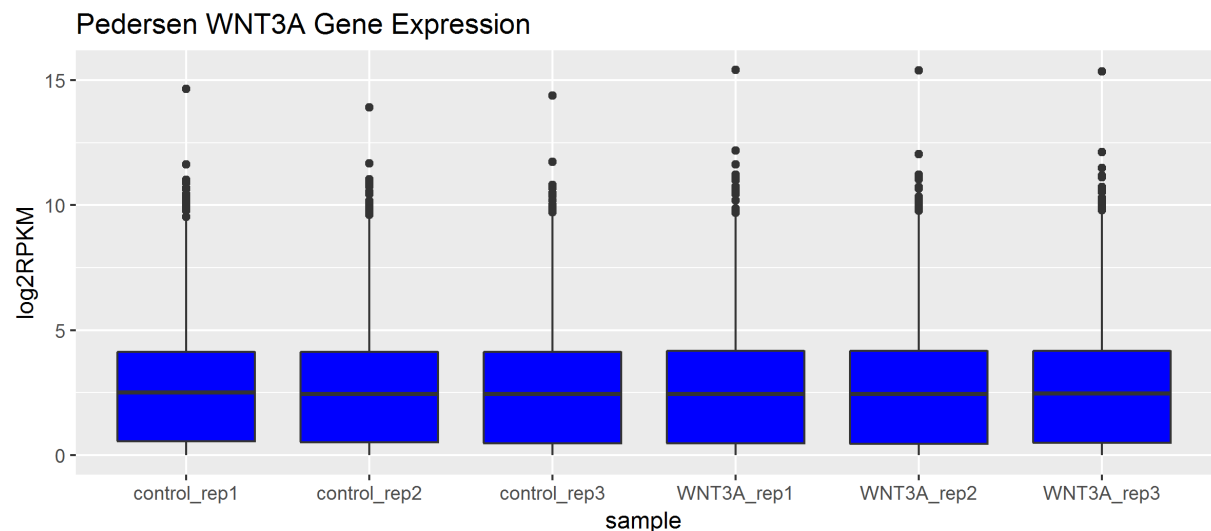
ggplot(data1,aes(x = sample, y = log2RPKM)) +
  geom_boxplot()
ggsave('boxplot1.png')
```



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`

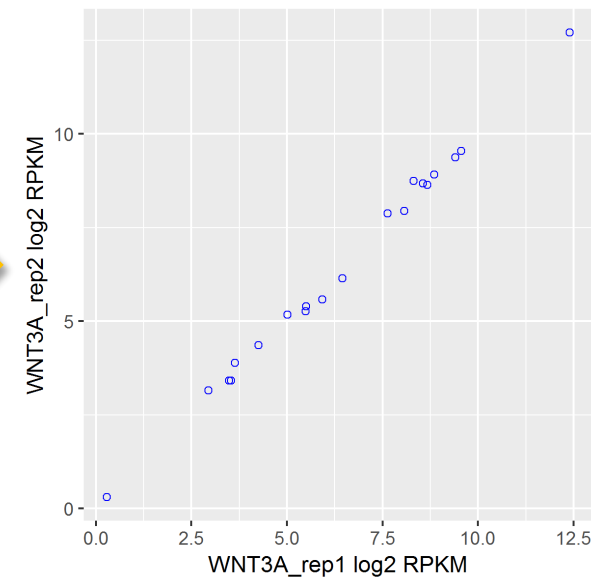
If time: Try creating a violin plot using `geom_violin()`



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



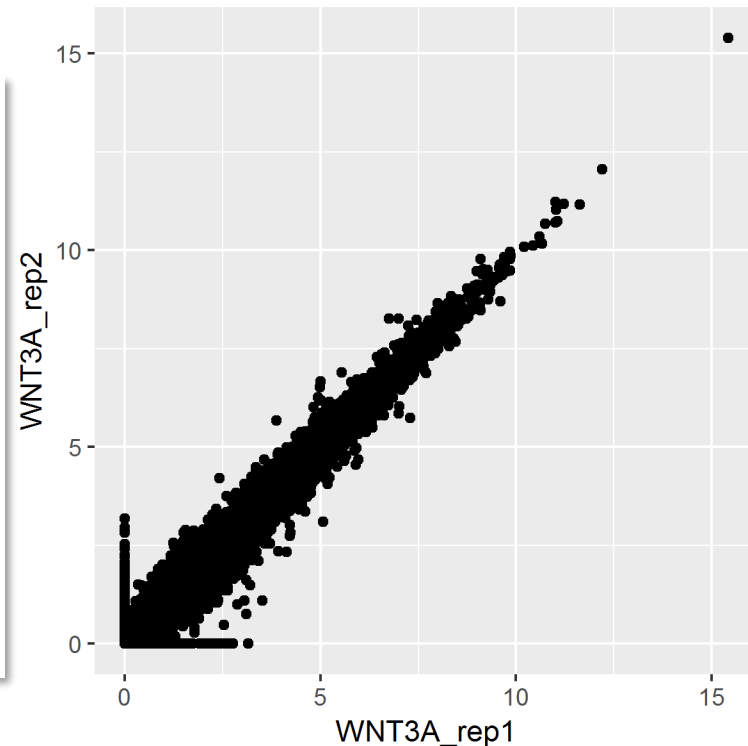
How to Create a Scatter Plot Using ggplot2

```
options(stringsAsFactors=F)

library(ggplot2)

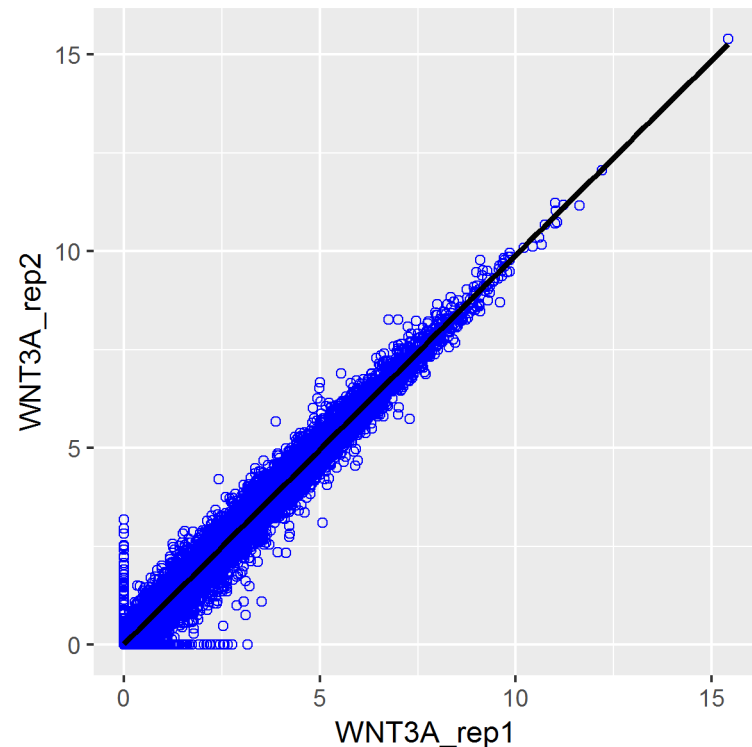
inFile2 = 'pedersenLog2RPKM_matrix20180817.txt'
data2 = read.delim(inFile2)

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



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)`



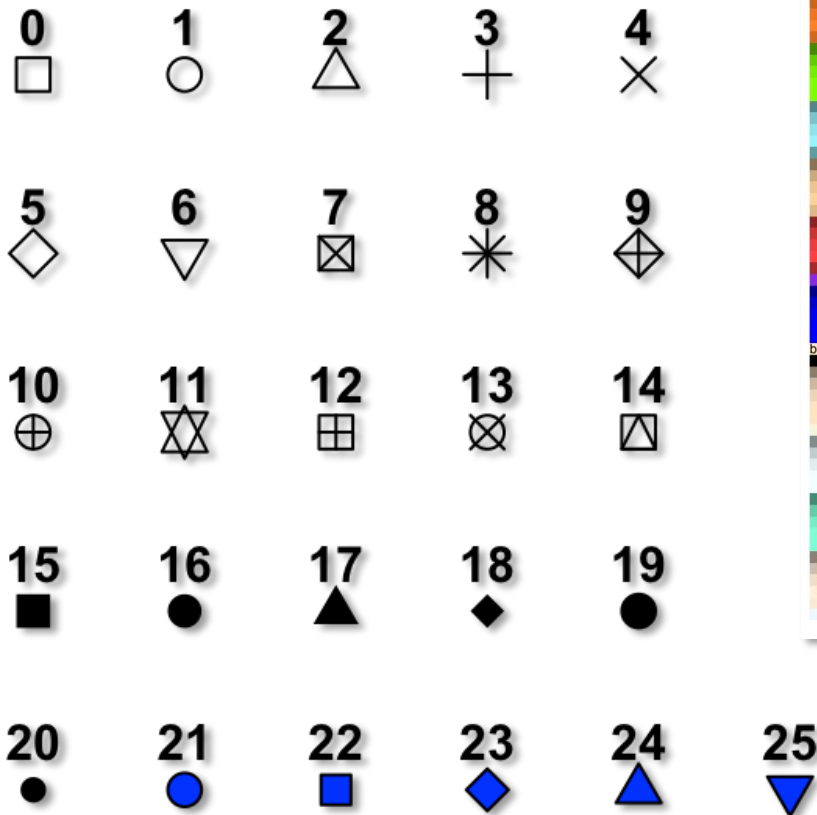
ggplot2

- 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



R Graphics Colors

grDevices::colors																			
coral3	deeppink4	gray27	gray87	gray99	lightpink1	mistyrose1	pink4	slategray1											
coral2	deeppink3	gray26	gray86	gray98	lightpink	mistyrose	pink3	slategray											
coral1	deeppink2	gray25	gray85	gray97	lightpink	mintcream	pink2	slategray4											
coral	deeppink	gray24	gray84	gray96	lightgreen	mediumslateblue	pink1	slateblue3											
chocolate4	darkviolet	gray23	gray83	gray95	lightgray	mediumslateblue	pink	slateblue2											
chocolate3	darkslategray4	gray22	gray82	gray94	lightgoldenrod1	mediumturquoise	peru	slateblue1											
chocolate2	darkslategray3	gray21	gray81	gray93	lightgoldenrod2	mediumspringgreen	peachpuff4	slateblue											
chocolate1	darkslategray2	gray20	gray80	gray92	lightgoldenrod3	mediumslateblue	peachpuff3	skyblue4											
chocolate	darkslategray1	gray19	gray79	gray91	lightgoldenrod4	mediumseagreen	peachpuff2	skyblue3											
chartreuse4	darkslategray4	gray18	gray78	gray90	lightgoldenrod5	mediumpurple4	peachpuff1	skyblue2											
chartreuse3	darkslategray3	gray17	gray77	gray89	lightcyan4	mediumpurple3	peachpuff	skyblue1											
chartreuse2	darkslategray2	gray16	gray76	gray88	lightcyan3	mediumpurple2	papayawhip	wheat4											
chartreuse1	darkslategray1	gray15	gray75	gray87	lightcyan2	mediumpurple1	palevioletred4	sienna4											
cadetblue4	darkseagreen4	gray14	gray74	gray86	lightcyan1	mediumpurple	palevioletred3	sienna3											
cadetblue3	darkseagreen3	gray13	gray73	gray85	lightcyan	mediumorchid4	palevioletred2	sienna2											
cadetblue2	darkseagreen2	gray12	gray72	gray84	lightcyan	mediumorchid3	palevioletred1	sienna1											
cadetblue1	darkseagreen1	gray11	gray71	gray83	lightcoral	mediumorchid2	palevioletred	sienna											
cadetblue	darkseagreen	gray10	gray70	gray82	lightblue4	mediumorchid1	paleturquoise4	seashell4											
burlywood4	darksalmon	gray9	gray69	gray81	lightblue3	mediumorchid	paleturquoise3	seashell3											
burlywood3	darkred	gray8	gray68	gray80	lightblue2	mediumblue	paleturquoise2	seashell2											
burlywood2	darkorchid4	gray7	gray67	gray79	lightblue1	mediumaquamarine	paleturquoise1	seashell1											
burlywood1	darkorchid3	gray6	gray66	gray78	lightblue	maroon4	paleturquoise	seashell											
burlywood	darkorchid2	gray5	gray65	gray77	lightblue	maroon3	palegreen4	seagreen4											
brown4	darkorchid1	gray4	gray64	gray76	lemonchiffon4	maroon2	palegreen3	seagreen3											
brown3	darkorchid	gray3	gray63	gray75	lemonchiffon3	maroon1	palegreen2	seagreen2											
brown2	darkorange4	gray2	gray62	gray74	lemonchiffon2	maroon	palegreen1	seagreen1											
brown1	darkorange3	gray1	gray61	gray73	lemonchiffon1	magenta4	palegreen	seagreen											
brown	darkorange2	gray0	gray60	gray72	lawngreen	magenta3	palegoldenrod	sandybrown											
blues4	darkorange1		gray59	gray71	lavenderblush4	magenta2	orchid4	salmon4											
blues3	darkolivegreen4	goldenrod4	gray58	gray70	lavenderblush3	magenta1	orchid3	salmon3											
blues2	darkolivegreen3	goldenrod3	gray57	gray69	lavenderblush2	magenta	orchid2	salmon2											
blues1	darkolivegreen2	goldenrod2	gray56	gray68	lavenderblush1	linen	orchid1	salmon1											
blues	darkolivegreen1	goldenrod1	gray55	gray67	lavender	lightgreen	orchid	salmon											
blanchedalmond	darkolivegreen	gold4	gray54	gray66	lavender	lightyellow4	orchid4	sandybrown											
black	darkviolet	gold3	gray53	gray65	khaki4	lightyellow3	orange4	royalblue4											
bisque4	darkkhaki	gold2	gray52	gray64	khaki3	lightyellow2	orange3	royalblue3											
bisque3	darkgray	gold1	gray51	gray63	khaki2	lightyellow1	orange2	royalblue2											
bisque2	darkgray	ghostwhite	gray50	gray62	khaki1	lightyellow	orange1	royalblue1											
bisque1	darkgray	gainsboro	gray49	gray61	khaki	lightsteelblue4	orange4	royalblue											
beige	darkgoldenrod4	forestgreen	gray48	gray60	ivory4	lightsteelblue3	orange3	rosybrown4											
azure4	darkgoldenrod3	floralwhite	gray47	gray59	ivory3	lightsteelblue2	orange2	rosybrown3											
azure3	darkgoldenrod2	firebrick4	gray46	gray58	ivory2	lightsteelblue1	orange1	rosybrown2											
azure2	darkgoldenrod1	firebrick3	gray45	gray57	ivory1	lightsteelblue	orange	rosybrown1											
azure1	darkcyan4	firebrick1	gray44	gray56	indianred4	lightslategray	olivedrab4	rosybrown											
aquamarine4	darkcyan3	firebrick2	gray43	gray55	indianred3	lightslategray	olivedrab3	red3											
aquamarine3	darkcyan2	firebrick2	gray42	gray54	indianred2	lightslategray	olivedrab2	red2											
aquamarine2	darkcyan1	firebrick1	gray41	gray53	indianred1	lightslategray	olivedrab1	red1											
aquamarine1	darkcyan	firebrick1	gray40	gray52	indianred	lightslategray	olivedrab	red											
aquamarine	darkcyan	dodgerblue4	gray39	gray51	indianred	lightskyblue2	olivedrab	red											
antiquewhite4	cornsilk4	dodgerblue3	gray38	gray50	hotpink4	lightskyblue1	navyblue	purple4											
antiquewhite3	cornsilk3	dodgerblue2	gray37	gray49	hotpink3	lightskyblue	navyblue	purple3											
antiquewhite2	cornsilk2	dodgerblue1	gray36	gray48	hotpink2	lightskyblue	navajowhite4	purple2											
antiquewhite1	cornsilk1	dodgerblue	gray35	gray47	hotpink1	lightskyblue	navajowhite3	purple1											
antiquewhite	cornsilk	dodgerblue	gray34	gray46	hotpink	lightsalmon4	navajowhite2	purple											
aliceblue	cornsilk4	deepskyblue4	gray33	gray45	honeydew4	lightsalmon3	navajowhite1	plum4											
white	cornsilk3	deepskyblue3	gray32	gray44	honeydew3	lightsalmon2	navajowhite	plum3											
	cornsilk2	deepskyblue2	gray31	gray43	honeydew2	lightsalmon1	navajowhite	plum2											
	cornsilk1	deepskyblue1	gray30	gray42	honeydew1	lightsalmon	moccasin	plum1											
	coral4	deepskyblue	gray29	gray41	honeydew	lightpink4	mistyrose4	plum											
	coral3	deepskyblue	gray28	gray40	lightpink2	mistyrose3	mistyrose2	plum											

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