

## Data visualization with R

**Bioinformatics Training and Education Program** 

https://bioinformatics.ccr.cancer.gov/btep/

Instructors:

Alex Emmons, PhD

Joe Wu, PhD

Amy Stonelake, PhD

#### Lesson 1 disclaimer

- In this we will not be scripting or plotting, so just follow the PowerPoint presentation
- Hands-on portion will start with lesson 2 where we will learn to import data and subsequently generate common plots with the data

### Lesson 1 objectives

After this lesson, you should be able to

- Provide the rationale for using R
- Understand the benefits of plotting
- Be able to choose the appropriate visualization that will help communicate your data most effectively
- Understand the basic syntax for constructing data visualization with ggplot2

#### Course structure

- 6 lessons
  - meets on Tuesdays / Thursdays, 1 2:15 pm.
  - Class meeting link: <u>https://cbiit.webex.com/cbiit/j.php?MTID=m3a51f03a94c8074118076b578cc1bca6</u>
- First class on April 11th and last class on April 27th
- 1 1.25 hour of class followed by a 45 minute help session
- Each lesson will be recorded and made available on the BTEP Video Archive (<u>https://bioinformatics.ccr.cancer.gov/btep/btep-video-archive-of-past-</u> <u>classes/</u>)
- Course material can be found at <u>https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r</u>

#### Course goals

- We want to show you how to explore data with graphing using the programming language R, with emphasis on the ggplot2 package although there are other options such as base R plotting and <u>Lattice</u>
- You will learn how to create basic plots that form the basis of more complex analyses
- You won't leave the class an R or ggplot2 expert, but you will have the basic graphing skills to start exploring your own data

### Class working environment

- **DNAnexus** is a cloud platform for bioinformatics analysis
- We have installed R and R Studio (the Integrated Development Environment [IDE] for R)
  - $\odot \mbox{IDEs}$  are software that allow us to interface with a programming language
  - $\odot \text{IDEs}$  make scripting easier
- Because we will be using R Studio via DNAnexus

Everyone is working on the same platform, using the same R version, and has access to the same packages and example data
 There will be no software to install prior to class

## Setup DNAnexus account **DNAnexus**



Log in



- Sign up for DNAnexus account at <u>https://www.dnanexus.com</u> upon registering for course
- Send us your DNAnexus username by completing the survey posted at <u>https://www.surveymonkey.com/r/WZ52TSG</u>
- Stay after class if you have not done this or you are having trouble

#### R Studio on DNAnexus

$\leftarrow$ $\rightarrow$ C in job-g93b3x00k4ykgq22p2g61qfx.dnanexus.cloud	🖈 🕈 🗖 😩 🗄
R File Edit Code View Plots Session Build Debug Profile Tools Help	rstudio 🕞 🛛 🕲
Coto file/function Coto file/function Coto file/function	麘 Project: (None) 👻
Invironment History Connections Tutorial	_
🗘 🖒 🚛 🕞 Source on Save 🔍 🧨 📲 🔚 🕞 Source 🗸 🚖 🚔 🔚 Import Dataset 🗸 🎸	≣ List •   © •
1 Global Environment -	Q,
Environment is empty	
1:1 (Top Level)  R Script	
Console Terminal × Jobs ×	
~/ ∞ Solution with the second	
R version 3.6.3 (2020-02-29) "Holding the Windsock" Copyright (C) 2020 The R Foundation for Statistical Computing Platform: x86_64-conda-linux-gnu (64-bit)	Modified
R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type 'license()' or 'licence()' for distribution details.	
R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.	
Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.	

### Why plot using R?

- Scripting for reusability and reproducibility
- Avoid copy-paste errors
- Publication quality plots
- Hard to plot large dataset using Excel
- Excel autocorrect messes up gene names (<u>https://pubmed.ncbi.nlm.nih.gov/34389840/</u>)
- Strong community support and there are lots of packages that were made specifically for life sciences (<u>https://bioconductor.org</u>)



#### About *Bioconductor*

The mission of the *Bioconductor* project is to develop, support, and disseminate free open source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assays. We are dedicated to building a diverse, collaborative, and welcoming community of developers and data scientists.

*Bioconductor* uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. *Bioconductor* is also available as <u>Docker</u> images.

## Helpful resources for our plotting endeavors

For ggplot2 documentation see

<u>https://ggplot2.tidyverse.org</u>

To learn more about plots and see example code

- data-to-viz (<u>https://www.data-to-viz.com/</u>)
- R Graph Gallery (<u>https://www.r-graph-gallery.com</u>)

Tidy Tuesday (<u>https://github.com/rfordatascience/tidytuesday</u>) BTEP has licenses for the following:

- Coursera
- Dataquest
- See <a href="https://bioinformatics.ccr.cancer.gov/btep/self-learning/">https://bioinformatics.ccr.cancer.gov/btep/self-learning/</a> to get access to Coursera or Dataquest

# Recommended courses from Coursera and Dataquest

Coursera

- Data Visualization with R by IBM
- Getting Started with Data Visualization in R by Johns Hopkins University (instructor: Collin Paschall)
- Data Visualization in R with ggplot2 by Johns Hopkins University (instructor: Collin Paschall)

Dataquest

• Data Visualization with R

https://www.dataquest.io/path/data-visualization-with-r/

#### Overview

ggplot2 is a system for declaratively creating graphics, based on <u>The Grammar of Graphics</u>. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

#### Installation

# The easiest way to get ggplot2 is to install the whole tidyverse: install.packages("tidyverse")

# Alternatively, install just ggplot2: install.packages("ggplot2")

# Or the development version from GitHub: # install.packages("devtools") devtools::install\_github("tidyverse/ggplot2")

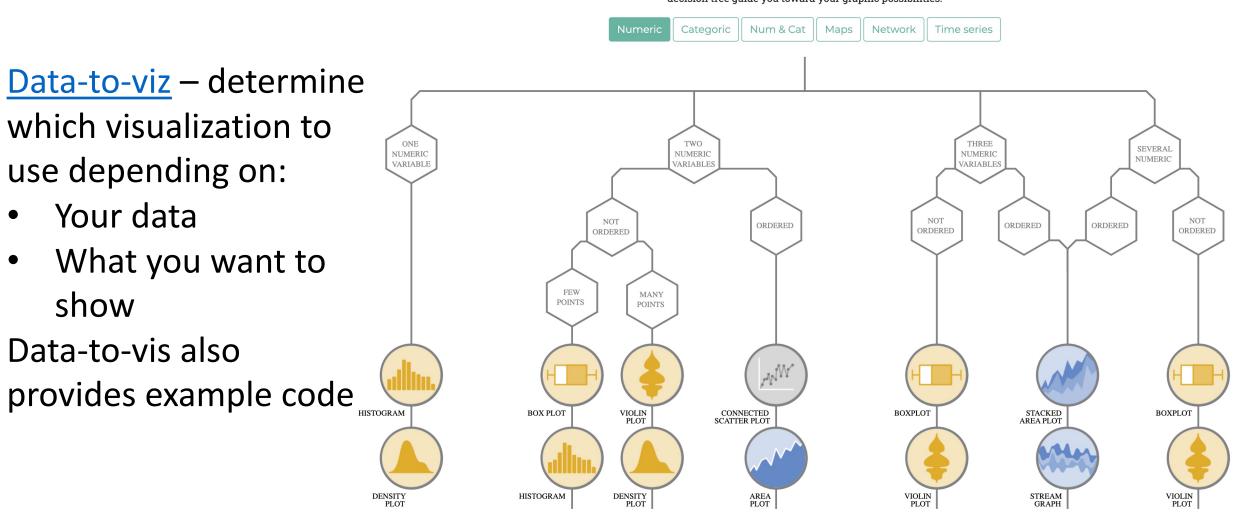
#### Cheatsheet

aca visaatiza	cion witch 66pt	ot2 : : снеат she		Stats An alternative way to build a layer.	Scales override defaults with scales package.	Coordinate Systems	Faceting
Sics		da points, use the geory's aesthetic properties to represent a	ggplot2	A stat builds new variables to plat (e.g., court, prop).	Scales map data values to the visual values of an senthetic. To change a mapping, add a new scale. In <= d + geom_bar(sen(50 + fc))	r + cart, cartadas)din + c(1, 5) - din, ylin The default cartesian coordinate system.	Fourth chicke a plat with subplicits based on the values of one of more discovering these
co can build every graph from the same onents: a data set, a coordinate system, anna - visual marks that represent data points.	GRAPHICAL PRIMITIVES a <- ggplotisconomics, analdute, unemploy() b <- ggplotiseals, aesis = long, y = lot()	TWO VARIABLES both continuous e (- ggiot(mpg, amin(y, hwy))	continuous bivariate distribution h ggiot(durondi, am)oral, prioc)	data stat gann coordinate plot		ratio, stim, yon - Cartesian coordinates with feed aspect ratio between x and y units.	t <- ggpict(npg, ani(ity, hey) + geon, point)
+ = ×	**prom, Mariki) and **expand, Smiths) Ensure limits include values across all pilots. **prom, narred(antiymed + list + 1, alpha, angin, constant + 10 - x, and, 5, pend, alpha, angin, constant, insteger, score	<ul> <li>** prom. Job el(seclubel = (t), nudge, s+1, nudge, s+2) = n, tabec, apa, angle, case, family, forthor, host, branc, bett, open, family, forthor, host, brane, size, open</li> <li>** prom. point() s, stable, color, 61, shape, size, strate</li> </ul>	h • pass, biologi (h • (825, 500) • , y, sifes, cales, 50, bergge, cas, weight 3 • pass, density, 240 • , y, sifes, cales, prop, lengue, size • , soon back	Valuations a real type throughing the default state of a general function, paper, and interfere "searing" on type using a date. function, and a search general "family, which can's a default growth translate a larger improvingent to a general function. Usestatesyntain to map state variables to perthetics.		gplot(reg.sec)+f()+peen(3ec)     Fip-cartenia-coordinates by reh/hing     said y and/tele mapping.     r assol, palar() via + 'Y', direction-1)     thete, start, direction-Polar coordinates.	1 - facet, pellowie + werkfill     Facet into solutions based on fi     - facet, pellowie + vanigenel()     Facet into more based on pare     1 - facet, pellowie + vanigenel()     1 - facet, pellowie + vanigenel()     1 - facet, pellowie + vanigenel()
lay values, map variables in the data to visual less of the groom (averbatical like size, color, and a values,	<ul> <li>* prom, path(lineed + "but", Ineptite + "round", linemitte + () s, y, alpha, color, group, faultype, size</li> </ul>	a • geom, quantila) x, y, alpha, (olor, group, livelype, size, weight	x, y, sight, color, 81, size	() () () () () () () () () () () () () (	Use with most aeathetics scale, ", continueus(). Wap cont' values to visual ones, scale ", discrete). Was discrete values to visual ones.	Transformed cartesian coordinates, Set atrans and yours to the name of a window function.	t + facet, wrap/wars(%) thrap facets into a rectangular layout.
+ = pist	<ul> <li>** pass, ph/specifics/b/pi * 552 - 5, s/ ph/s.</li> <li>** pass, ph/specifics/b/pi * 552 - 5, s/ ph/s.</li> <li>** pass, performance interport, size inte</li></ul>	**peak_regides **** ** peak_rest. Introduces the ** peak_rest. Introduces the ** peak_rest. Introduces the ** peak_rest. Introduces the ************************************	continuous function i cogginiticanowski, aeolytek, unerging() i cogen, state i cogen, face) i cogen, face) i cogen, face) i cogen, face) i cogen, face)	<pre>e=stat_bio[cond0++], box-day=s(0 e=stat_bio[cond0++], box-day=s(0 e=stat_cond0++), area(-, area(-</pre>	Matar, "	Position Adjustments	Set eachs to be and limits very access fronts. 1 · fact, grid (house · accellence, ach · access(b), access · first of a set to individual fronts · first, act - access fronts algust · first, act - access fronts algust · first, act - access fronts algust · first, access fronts algust · first, access fronts algust · for fact access fronts algust · for fact access fronts access fronts algust · first access fronts reacted. Societation : tables for the indi-
to the securities before to build a graph. Inter a security of the security o	LINE SEGMENTS unners and/below to a single state and the second	ene discrete, one continueus tre gamerage ansites, tangi	<ul> <li>(*) Sum years offic state in the first particular state of the state o</li></ul>	$\label{eq:constraints} \begin{split} & \mathbf{v}_{1}(\mathbf{x}_{1}^{(1)}) = (\min_{i=1}^{n}, \min_{i=1}^{n-1}, \min_{i=1}^{n-1}, \min_{i=1}^{n-1}, \min_{i=1}^{n-1}, \min_{i=1}^{n-1}, \min_{i=1}^{n-1}, \max_{i=1}^{n-1}, \max_{i=1}^{n-$	LE STUCKTION RCALE The unit and unit a	Public delatamente delamante dela monte para la consegna guerra tar escal al forma dela consegna fue asar escal a escal a consegna a la consegna del asar escal a escal a consegna della consegna d	the set of the se
et() Returns the last plot. ("plot ang," width + 5, height + 5) Seves last plot. I file samed "plot ang," in working directory. I file type to file edension.	<ul> <li>c., upta.coint.BL.tempje.com</li> <li>c., post.coint.BL.tempje.com</li> <li>c., post.coint.BL.tempje.com</li> <li>c., post.coint.BL.tempje.com</li> <li>c., post.coint.BL.tempje.com</li> <li>c., post.coint.BL.tempje.com</li> </ul>	for general, sinilarly color "server")     s_g, alpha, color, thit, group, treetype, clas, weight     both discrete     cr applications of a sector, color)	1 2 per per anticipante per constructiones de la construcción de la co	$\label{eq:statestatest} \begin{split} & \mathbf{x} + statestatestatestatestatestatestatestat$	a+scale, RL, det(int) arc(int - s) a+scale, RL, det(int) arc(int - Ther?) a+scale, RL, gealer() arc^1of?, high="yellow")	Each settiments an type of one another. Each position adjustment can be result as a function with makes width and height apprents; 1 - prom, targetester - position, disspecietter - 13	Places a prom with manually selected arethetics. <b>a</b> • <b>printing</b> : • <b>print</b> , unlish dodge • (2) Antic constant or semisping labels with golds, solits dodge or angle). <b>a</b> • <b>printing</b> (1) • "soury" [select signed free the each <b>a</b> • <b>printing</b> (1) • "soury" [select signed free each <b>a</b> • <b>printing</b> (1) • "soury" [select signed free each <b>a</b> • <b>printing</b> (1) • "soury" [select signed free each <b>a</b> • <b>printing</b> (1) • "soury" [select signed free each <b>a</b> • <b>printing</b> (1) • "soury" [select signed free each
Common sestivatic values.	<ul> <li>x, y, stpha, colin, fill</li> <li>c + genera, frequely()</li> <li>x, y, stpha, colin, group, freques, size</li> </ul>	• • • • • • • • • • • • • • • • • • •	state + tolower(rownames(/SAmesti()) mag <- map, data("state") k <- ggalat/data, aes(fil + murder()	ggpiet) + sliet(-5, 1) + stat, Auscien()(-n + drorn, n+20, grown * [rem?] #   -k., -y., generated + stat, and enclosed + 1, 1000	a + scale, fill, gradiant@low + "red", high + "blue", mid + "white", melpoint + 20 a + acale, fill, gradiantelision + tops colora(5)	Themes	a - theme:legand.acuition - "hotoer"] Place legand at "bottom", "top", "with, or "right", a - acuit, 50, discutsion or "Title", bottom://www.com/acuition.or - "Title",
<ul> <li>Integrate and the global state of the state</li></ul>	C+genn, Nidegram(hiraddh+1) 1. p. sipha, color, NJ, Dretype, clas, weight	a + geom, jitte-O-sight + 3, width + 2] s, y, sighta, color, 63, shape, size	k * peom, map(sec(nap, id + state), map + map) * expand, Benite(x + mapString, y + mapStri) map, id, sight, color, 65, Energies, size	x, y, xample   .sample., .theoretical. s + stat_num() x, y, star) .A., .prop.	Alto: tsinbew(, heat.oslovs), tertain.oslovs), on.oslovs(, #Calor@rewer.brewer.pat)	and white background and rethere, light) (** theme, light) (** theme, light) (** theme, light) (** theme, light)	Set agent the and labels with a scale function.
<ul> <li>string ("round", "butt", or "square")</li> <li>string ("round", "mitre", or "bevel")</li> </ul>	Cl + prom_apgionitample + hegil x_y_alpha_color_NL livetype_size_weight	THREE VARIABLES seab(c <= with)main, sprijdella_iorg*2 + della_iof*2);( <= g		<pre>+ stat, summary(fun.dots + "mean.d.boot") b+ stat, summary_bin(fun + "mean", gourn + "bar") e+ stat, identify()</pre>	SHAPE AND SIZE SCALES p+-e+grom_point(acc)dage+f, size+(yf) p+rate_shape()+rate_size()	Grant Statignord	Wheel cligging (ordered)
Anger (ine widt) in ment	discrete d <- gg/dat/mpg_aes(ft)	() 1 • pase, carbox(lard) • () 4, y. s. alpha, calor, group, lowtype, size, weight	I + geom, restar(sec)(II + d, h)art + 0.5, spat + 0.5, interpolate + FA(50) h, y, alpha, 50	erstal,unique)	p+ sole shape manual (state + c)(7) p+ sole reflecting = c(1,6)	Dark for contrast	1 + coord_cartinizacjotm + c(0, 100, you + c(10, With clipping transvers unseen data points) 1 + stimult, 1001 + pland/d, 201
	d+geom,bar() x, apha.color,fill, linetype, size, weight	1+ green, contour, filled(arc)(1+z) x, y, alpha, color, fill, group, Trenype, size, subgroup	1+ geom, tile(sec(51+g))	R Studio	p+state_state_state_state=0	r + golde (104) + thereplotific postion + plot() + Renetional Society of + stemant, world + "Silar"	t + main x continuend into + cli, 1003 +

#### ggplot2 website :

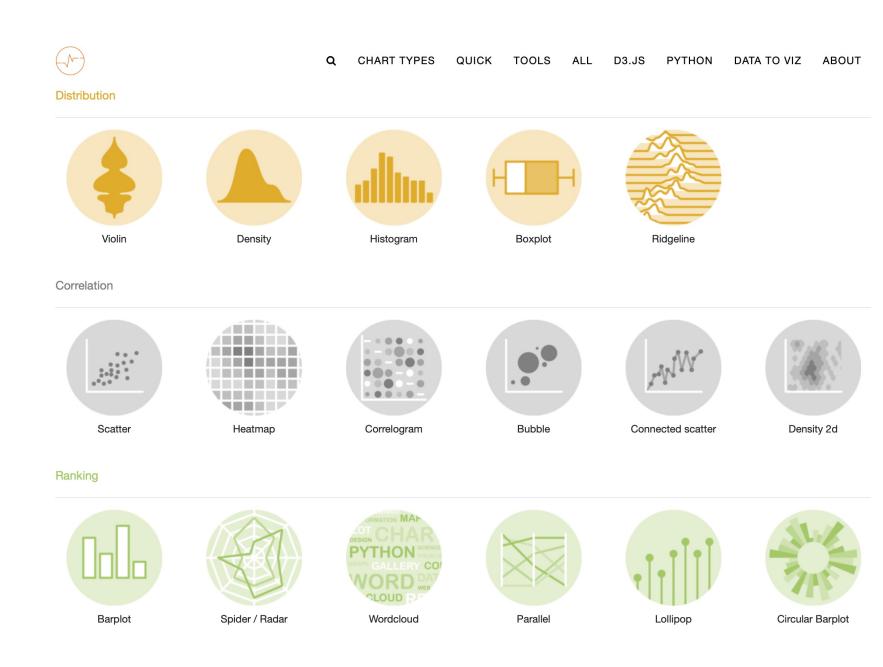
- Instructions for installation
- Cheat sheet for quick reference
- Help documents

What kind of data do you have? Pick the main type using the buttons below. Then let the decision tree guide you toward your graphic possibilities.



#### <u>R graph gallery</u>:

- suggests plot type to use depending on what you want to show
- provides example code



← the R Graph Gallery

#### Q CHART TYPES QUICK TOOLS ALL D3.JS PYTHON DATA TO VIZ ABOUT

#### Basic violin plot

R graph gallery – example code Building a violin plot with ggplot2 is pretty straightforward thanks to the dedicated geom\_violin() function.

# Library
library(ggplot2)

#### # create a dataset

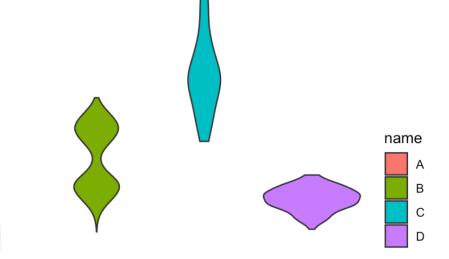
```
data <- data.frame(
    name=c( rep("A",500), rep("B",500), rep("B",500), rep("C",20</pre>
```

value=c( rnorm(500, 10, 5), rnorm(500, 13, 1), rnorm(500, 18)

#### # Most basic violin chart

```
p <- ggplot(data, aes(x=name, y=value, fill=name)) + # fill=na
geom_violin()
```

#р



<u>Tidy Tuesday</u> releases small dataset every Tuesday that could be used to practice ggplot2

$\mathbf{O}$	Product $arphi$ Team Enterprise Explore $arphi$ Marketp	blace Pricing $\vee$
🛱 rfordatascience / tidytuesday	y Public	
<> Code 🕢 Issues 171 🕴 Pu	ull requests 🕟 Actions 🗄 Projects 🕮 Wiki	🕐 Security 🛛 🗠 Insights
ۍ ۲	ి master - tidytuesday / data /	
	<b>jthomasmock</b> collegiate sports data for week 13	
	2018	Rename README.md to readme.md
	2019	Merge pull request #290 from erictleung/fix-data-links
	2020	Add links to quoted text in Big Mac data
	2021	milk types added
	2022	collegiate sports data for week 13
	] readme.md	Update readme.md

### Plots that will be covered in this course

- Here, we will present some common plot types and where we might see these
- The list presented here is not exhaustive and a good resource is <u>https://www.data-to-viz.com</u>
- The goal is to get everyone familiar enough with ggplot2 so that you feel comfortable and are encouraged to continue learning and/or apply ggplot2 to make your own plots
- Remember practice and repetition is the key to mastering a new skill

## Purpose of plotting data

Plots serve as diagnostic tools before downstream analysis

- Quality (example: quality of sequencing data)
- Structure
  - $\odot$  Missing data points
  - $\odot$  Size of the data
  - $\odot$  Number and type of variables
- Distribution
  - Determines appropriate statistical approach
  - How to model the data (RNA seq differential gene expression packages model data as a negative binomial distribution)
- The above can be classified as exploratory analysis

### Plots are used to convey research findings

- Whether it is a poster presentation or publication, we use plots to convey our research findings
- This step is known as explanatory analysis

#### Plotting condenses large datasets

- To facilitate interpretation
- To the right we have the first 13 observations (out of 60) in a study that examined dietary supplements (supp) at various doses (dose) on tooth length (len)

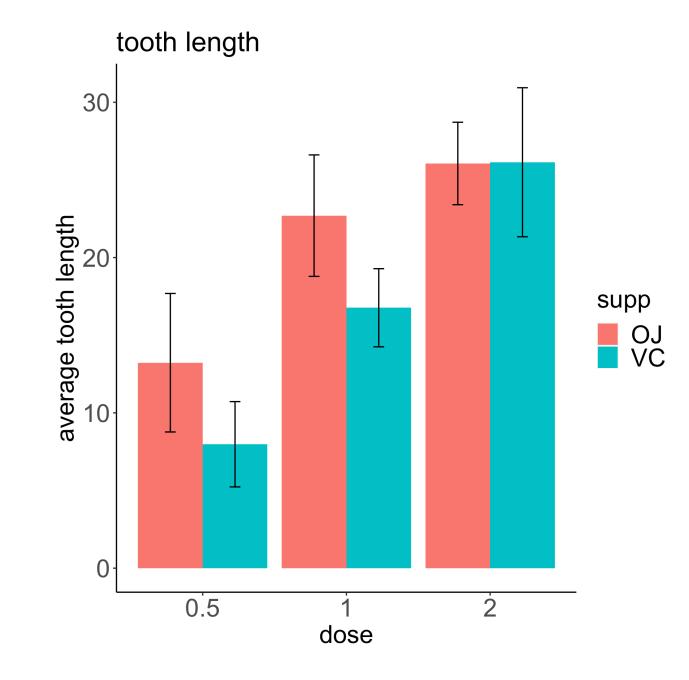
   it will be hard to make sense of the data just by looking at the table

len	supp	dose
4.2	VC	0.5
11.5	VC	0.5
7.3	VC	0.5
5.8	VC	0.5
6.4	VC	0.5
10	VC	0.5
11.2	VC	0.5
11.2	VC	0.5
5.2	VC	0.5
7	VC	0.5
16.5	VC	1
16.5	VC	1
15.2	VC	1

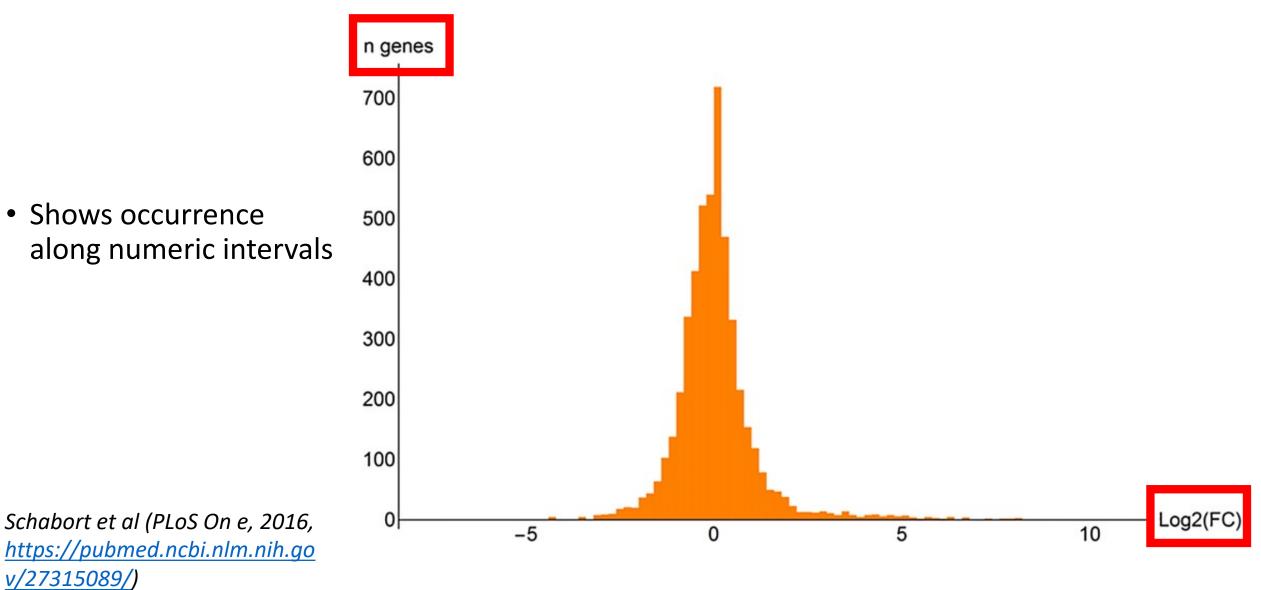
### Bar plot

Shows magnitude or mean/variation across categorical or discrete variables

The bar chart used data from the ToothGrowth dataset built into R. The study examined the effect that orange juice and vitamin c at various doses had on tooth growth.

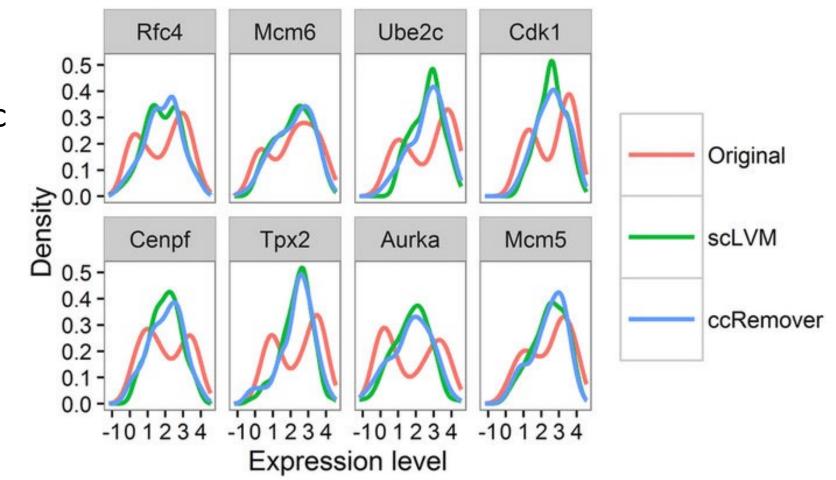


### Histogram



## Density plot

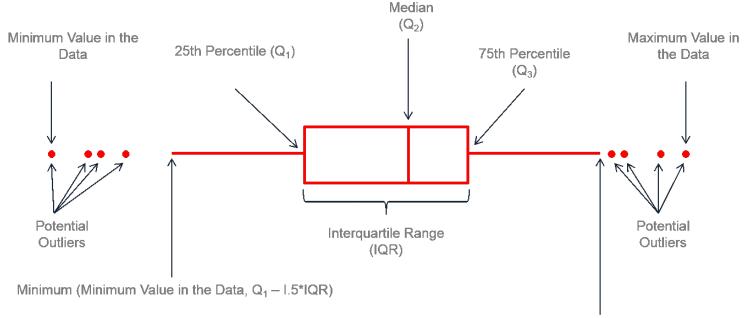
- Shows distribution distribution along numeric intervals (smoothed out version of a histogram)
- We will learn about faceting, which generates subplots in lesson 2



Barron and Li (2016, Nature Scientific Reports, <a href="https://pubmed.ncbi.nlm.nih.gov/27670849/">https://pubmed.ncbi.nlm.nih.gov/27670849/</a>)

Box and whisker plot

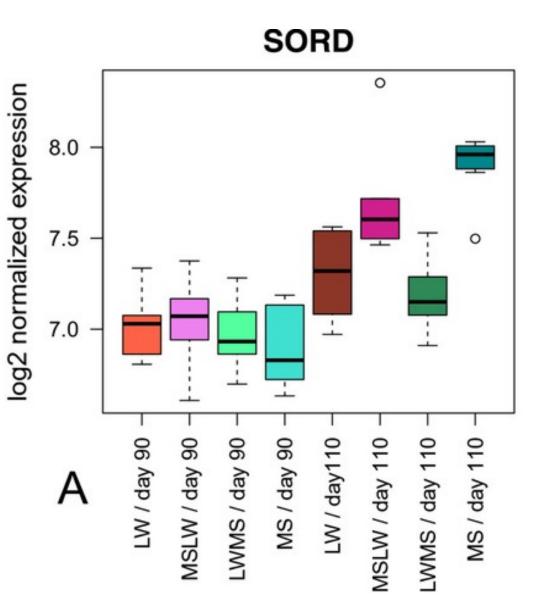
- Shows summary statistics
- Outliers
- Can incorporate many variables at the same time

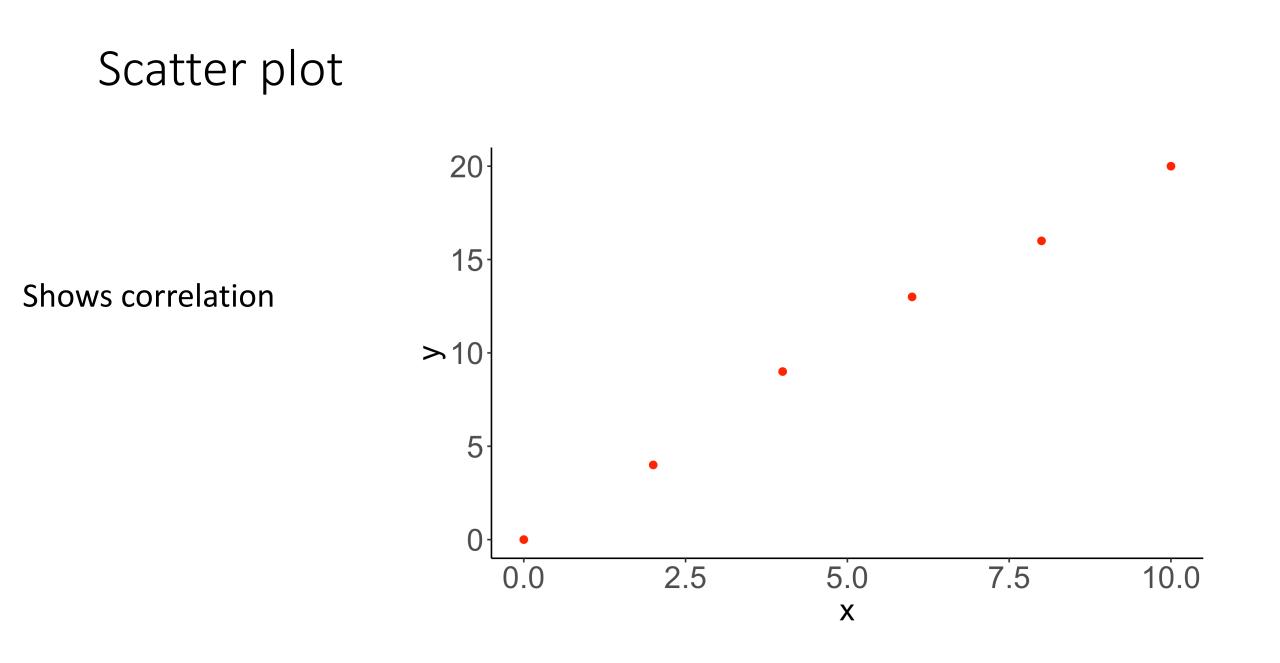


Maximum (Maximum Value in the Data, Q<sub>3</sub> + 1.5\*IQR)

# Box and whisker plot - example

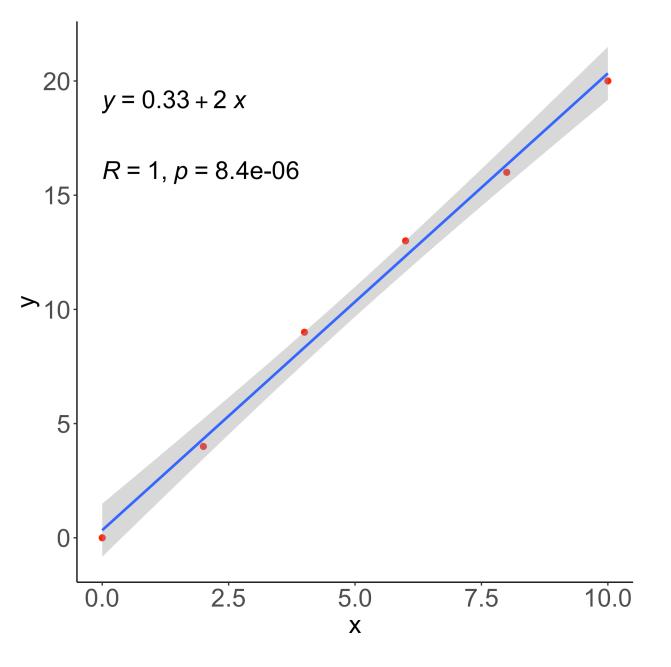
Voillet et al (2014, BMC Genomics, https://pubmed.ncbi.nlm.nih.gov/2 5226791/)





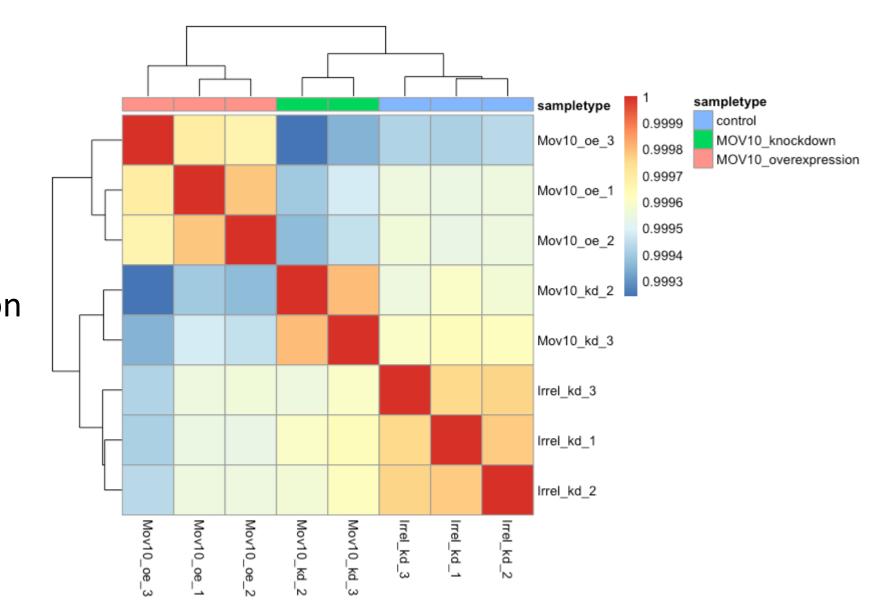
# Scatter plot with regression line

Same plot as previously shown, but ggplot2 makes it easy to add additional information to the plot such as regression line along with confidence interval (gray shading)



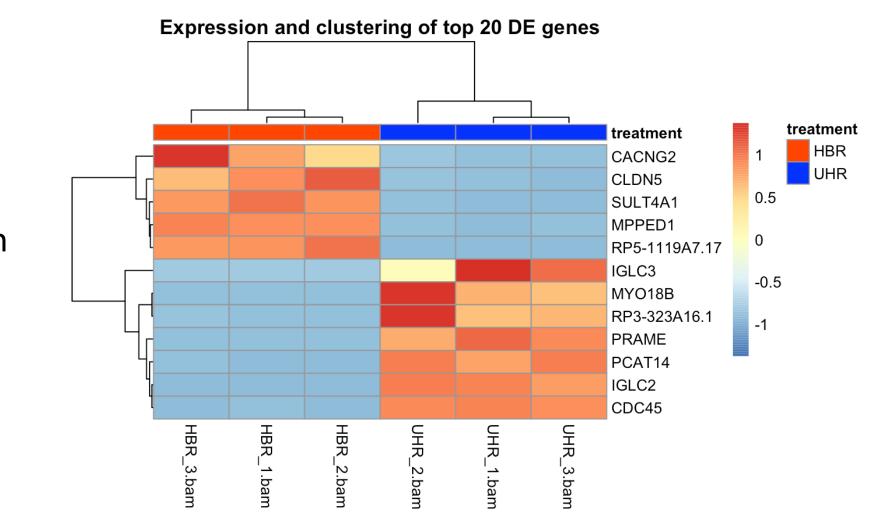
# Heatmap and dendrogram

 Allows for visualization of correlations and clustering



Source: <u>https://hbctraining.github.io/DGE\_workshop\_salmon/lessons/03\_DGE\_QC\_analysis.html</u>

# Heatmap and dendrogram

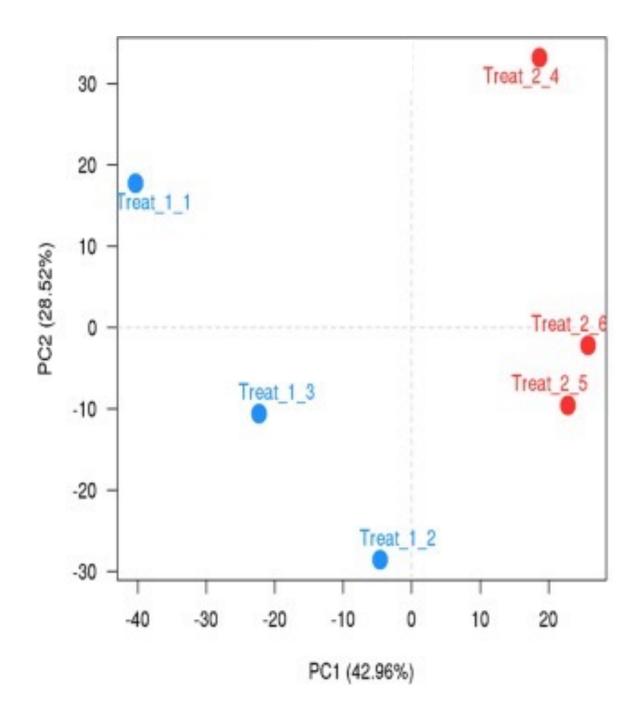


 Allows for visualization of correlations and clusters

Heatmap of gene expression from the Human Brain Reference and Universal Human Reference RNA sequencing data (<u>https://rnabio.org/module-01-inputs/0001/05/01/RNAseq\_Data/</u>)

Principal component analysis (PCA)

- Allows for visualization of clusters
- If using a differential expression package, we can import PCA results into ggplot2 to make the plot



Bakhtiarizadeh et al (2019, Nature Scientific Reports, <u>https://pubmed.ncbi.nlm.nih.gov/31235755/</u>)

### Volcano plot

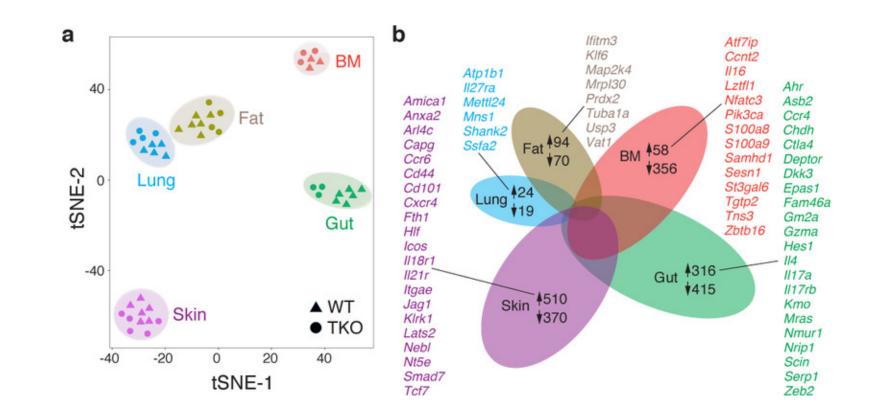
Differential gene expression results from the airway dataset (https://bioconductor.org/pack ages/release/data/experiment/ html/airway.html) were used to generate this volcano plot. The airway study looked at the transcriptomic profile of airway smooth muscle without or with dexamethasone treatment.

ZBTB16 6 ANGPTL7 -log10FDR ALOX15B • significant down no significance up Fold change=2 FDR=0.05 0 10 5 logFC Fold change=0.5

airway sm dexmethasone versus control

## tSNE and Venn diagram

- tSNE also shows clustering
- Venn diagram shows commonality
- We will learn about multi-panel plots in lesson 6



Ricardo-Gonzalez et al (2018, Nature Immunology, <u>https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6202223/</u>)

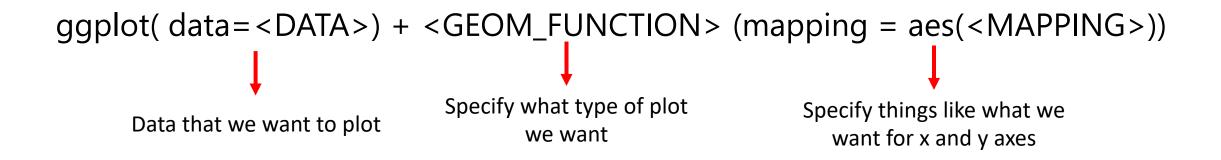
### Overview of ggplot2

- Popular tool used for plotting in R
- Generates publication quality plots

### Quick glance at usage of ggplot2

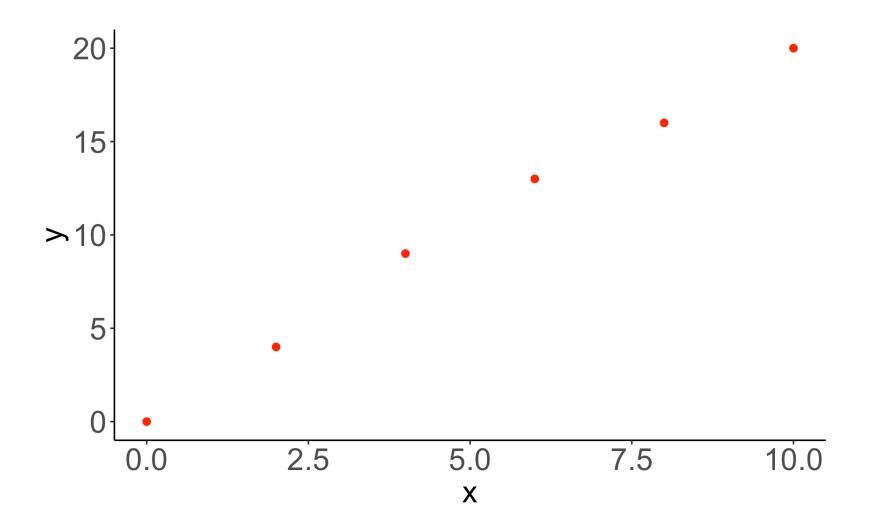
install.packages("ggplot2") # installs ggplot2

library(ggplot2) # loads ggplot2 in to our R work environment



ggplot(data=my\_data)+geom\_point (mapping=aes(x=x, y=y))

#### ggplot(data=my\_data)+geom\_point (mapping=aes(x=x, y=y))



#### Lesson 1 recap

- Went over course objective and hopeful outcome
- Went over plots that we will work with for this course series
- Saw an introduction to the structure/syntax for ggplot2, which will be reiterated through this course series
- For detailed lesson plans see the course material at <u>https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r</u>
- Please stay after for help with obtaining and/or setting DNA Nexus account

### Sneak preview

- From lesson 2 onward, the class will be hands-on so we will be coding OImporting dataset
  - Generating plots from the dataset so we will have plenty of opportunity to become familiar with ggplot2
- Lesson 2 will highlight
  - $\circ$  Basic ggplot2 syntax
  - $\circ \text{Geoms}$
  - $\circ$  Faceting
- For detail course outline, visit the course page at <a href="https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r">https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r</a>