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: https://cbiit.webex.com/cbiit/j.php?MTID=m3a51f03a94c8074118076b578cc1bca6
- 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 (https://bioinformatics.ccr.cancer.gov/btep/btep-video-archive-of-past-classes/)
- Course material can be found at https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r

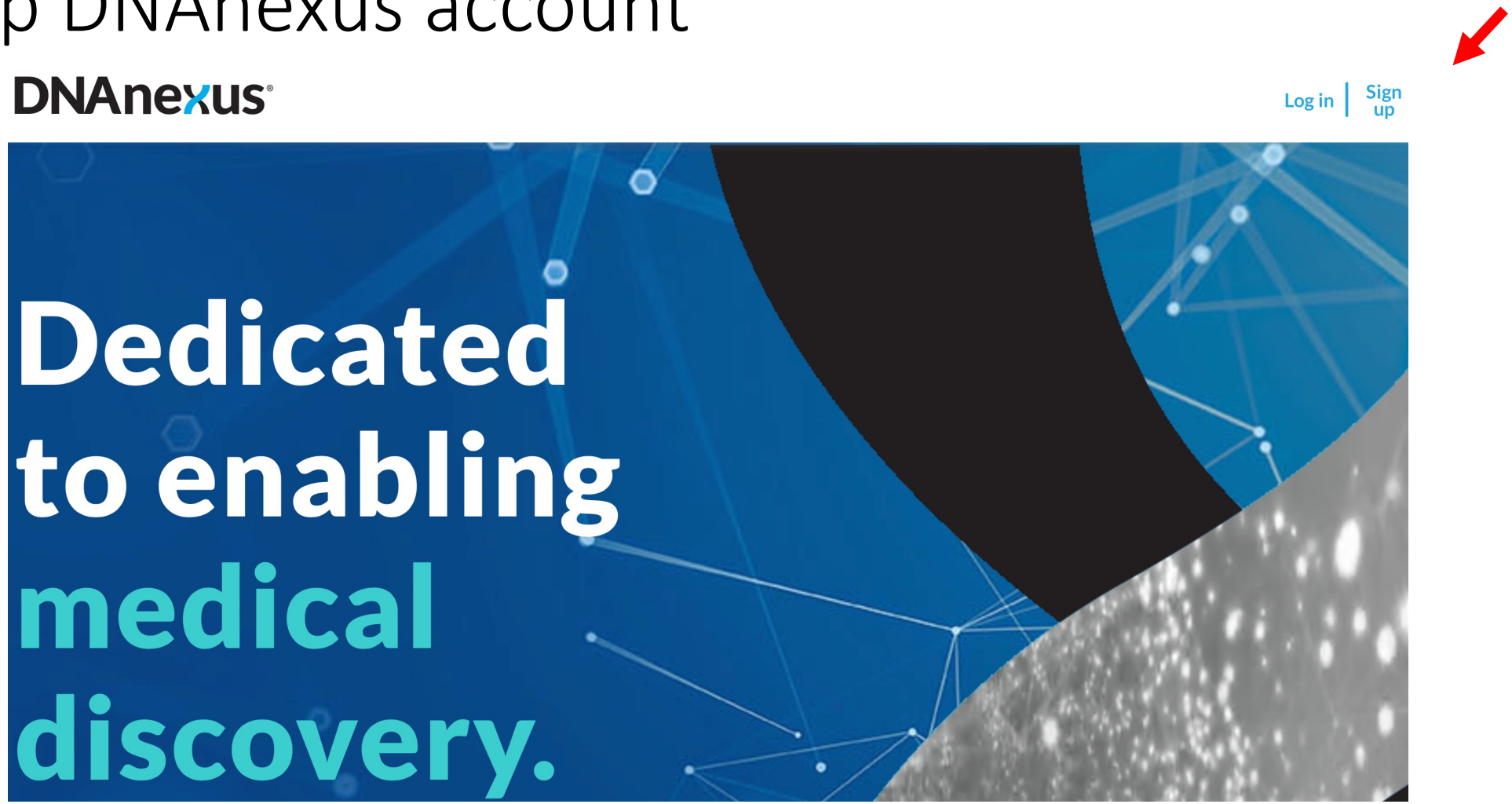
Course goals

- 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 Lattice
- 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)
  - IDEs are software that allow us to interface with a programming language
  - 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



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

R Studio on DNAnexus

job-g93b3x00k4ykgq22p2g61qfx.dnanexus.cloud

File Edit Code View Plots Session Build Debug Profile Tools Help

rstudio

Go to file/function | Addins | Project: (None)

Untitled1

Source on Save | Run | Source

1

1:1 (Top Level) R Script

Environment History Connections Tutorial

Import Dataset | List

Global Environment

Environment is empty

Console Terminal Jobs

~/

```
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)

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.

>
```

Files Plots Packages Help Viewer

New Folder | Upload | Delete | Rename | More

Home

Name	Size	Modified

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 (https://pubmed.ncbi.nlm.nih.gov/34389840/)
- Strong community support and there are lots of packages that were made specifically for life sciences (https://bioconductor.org)

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

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 Docker images.

Helpful resources for our plotting endeavors

For ggplot2 documentation see

- https://ggplot2.tidyverse.org

To learn more about plots and see example code

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

Tidy Tuesday (https://github.com/rfordatascience/tidytuesday)

BTEP has licenses for the following:

- Coursera
- Dataquest
- See https://bioinformatics.ccr.cancer.gov/btep/self-learning/ 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/

[ggplot2 website](#) :

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

Create Elegant Data Visualisations Using the Grammar of Graphics • ggplot2

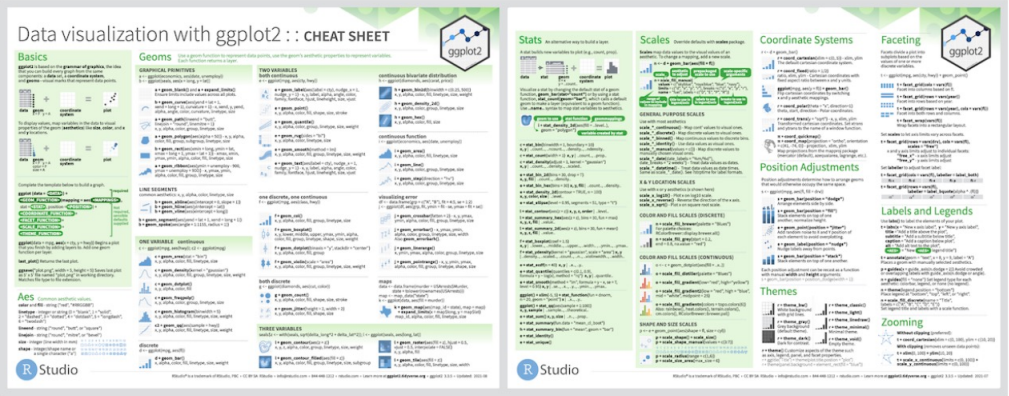
Overview

ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. 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")
```

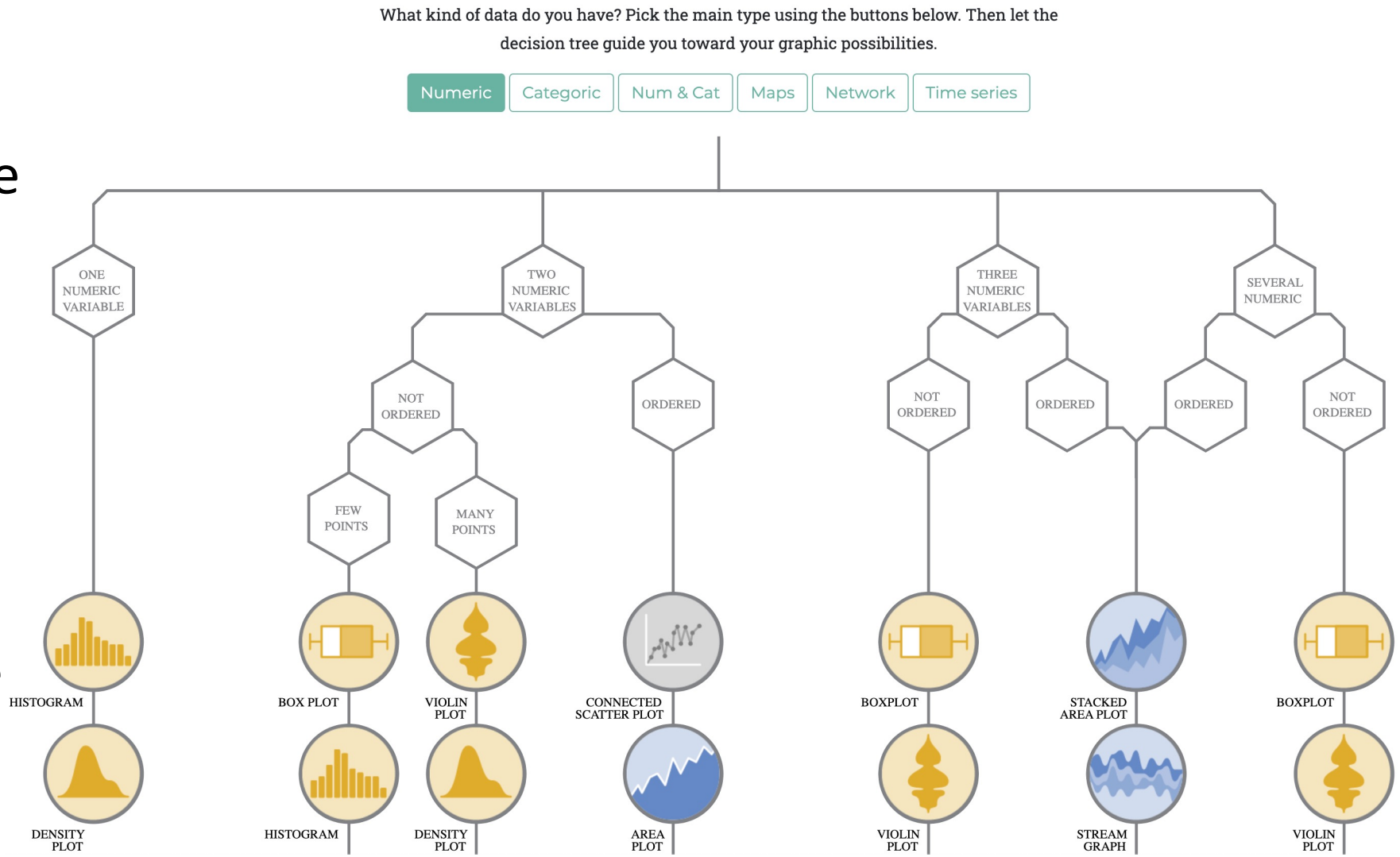
Cheatsheet



Data-to-viz – determine which visualization to use depending on:

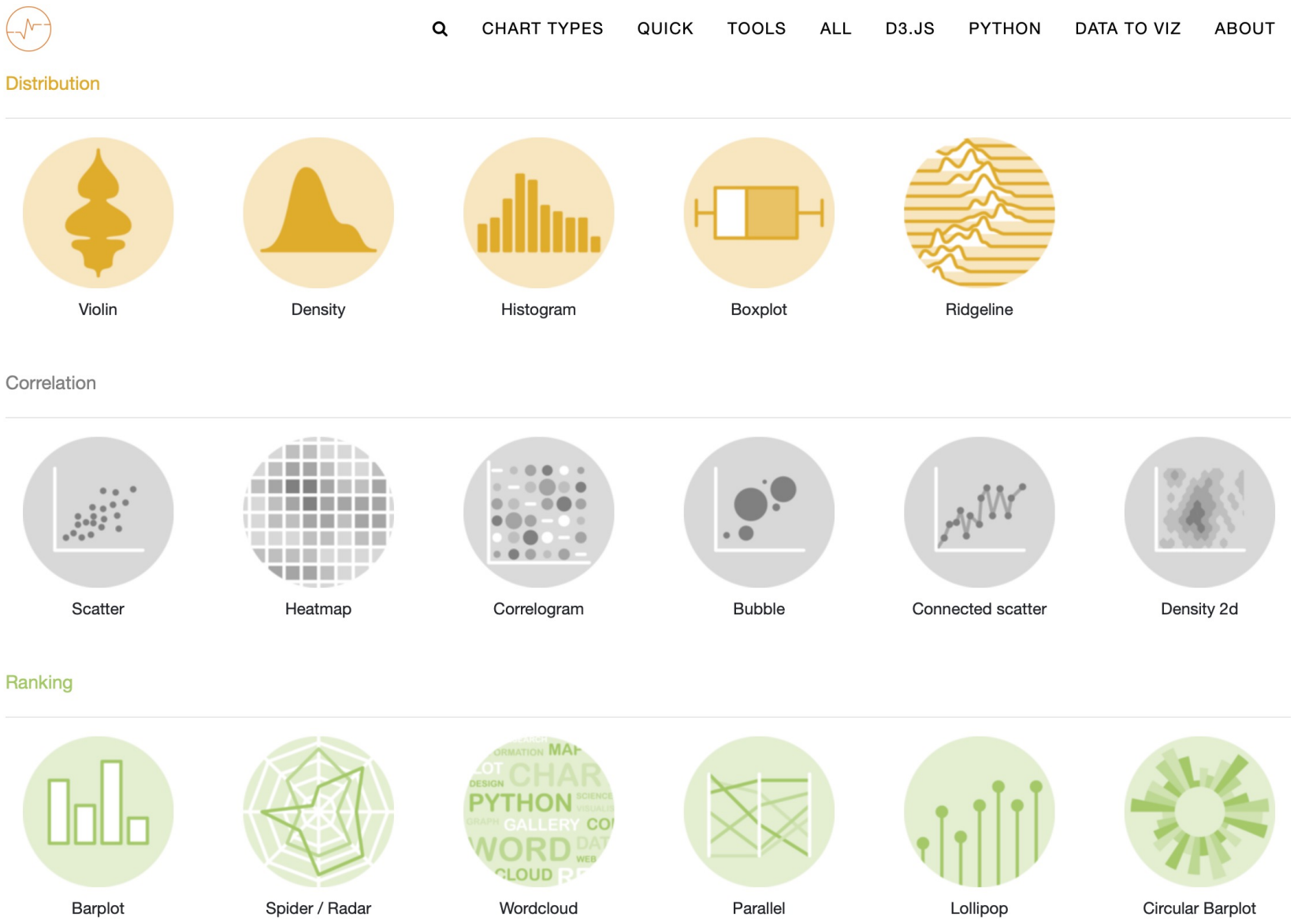
- Your data
- What you want to show

Data-to-vis also provides example code



[R graph gallery]():

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

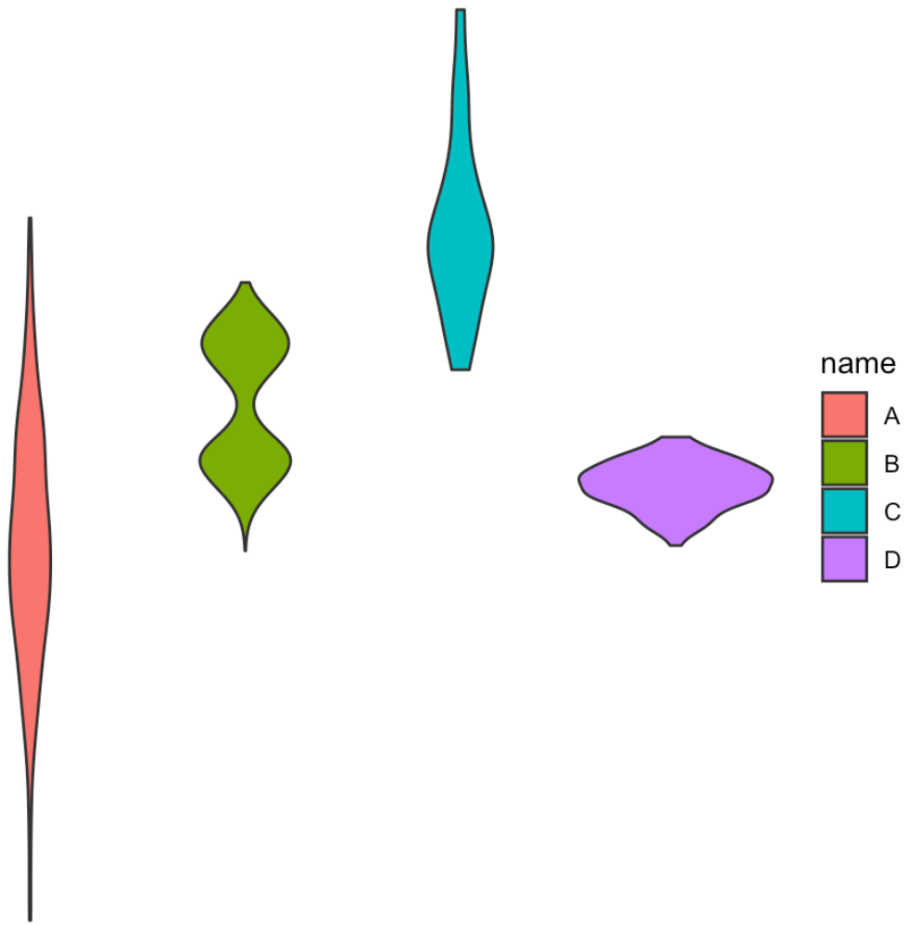


R graph gallery – example code

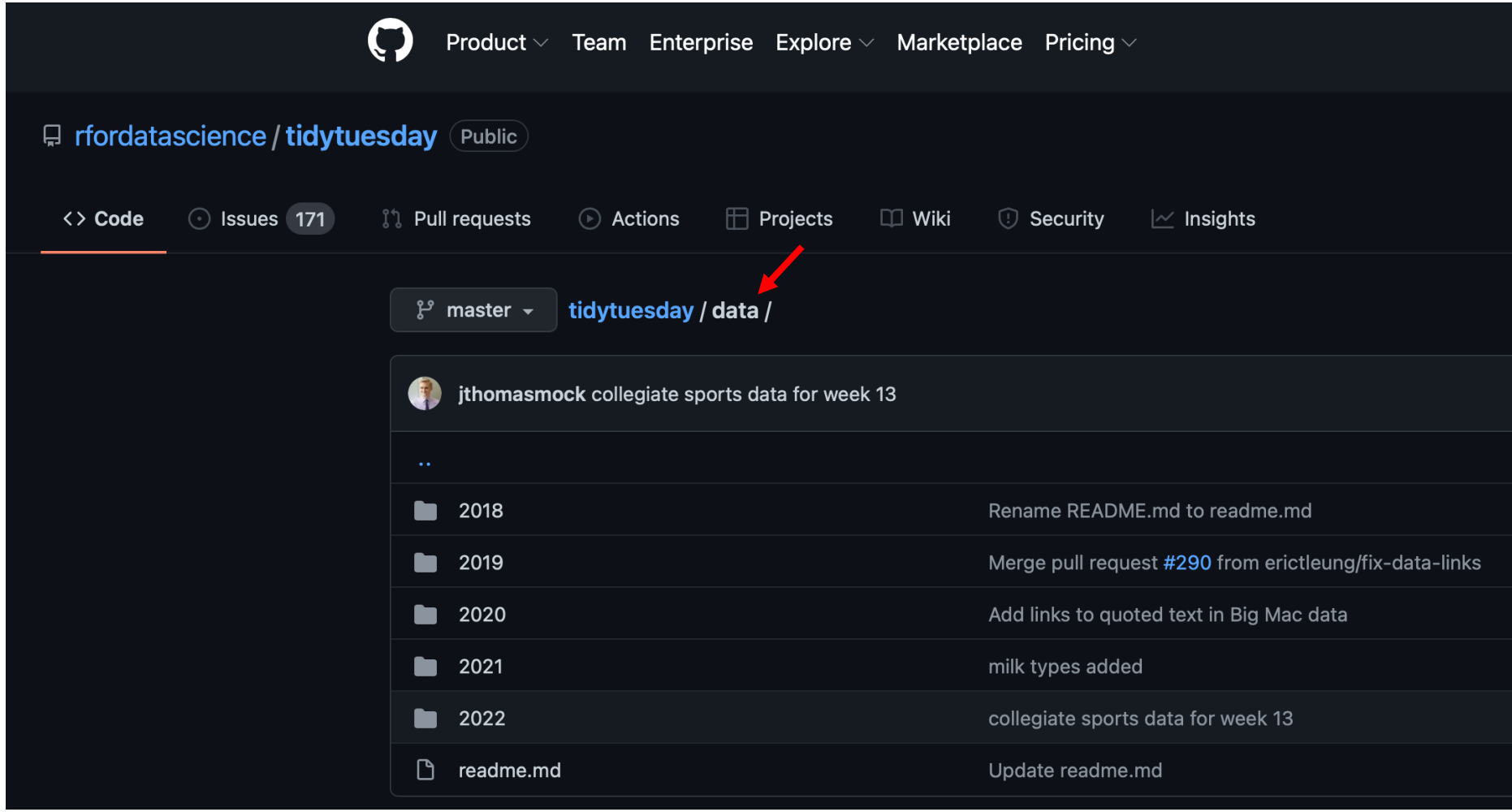
Basic violin plot

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
  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()

#p
```



Tidy Tuesday releases small dataset every Tuesday that could be used to practice ggplot2



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 https://www.data-to-viz.com
- 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
  - Missing data points
  - Size of the data
  - 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

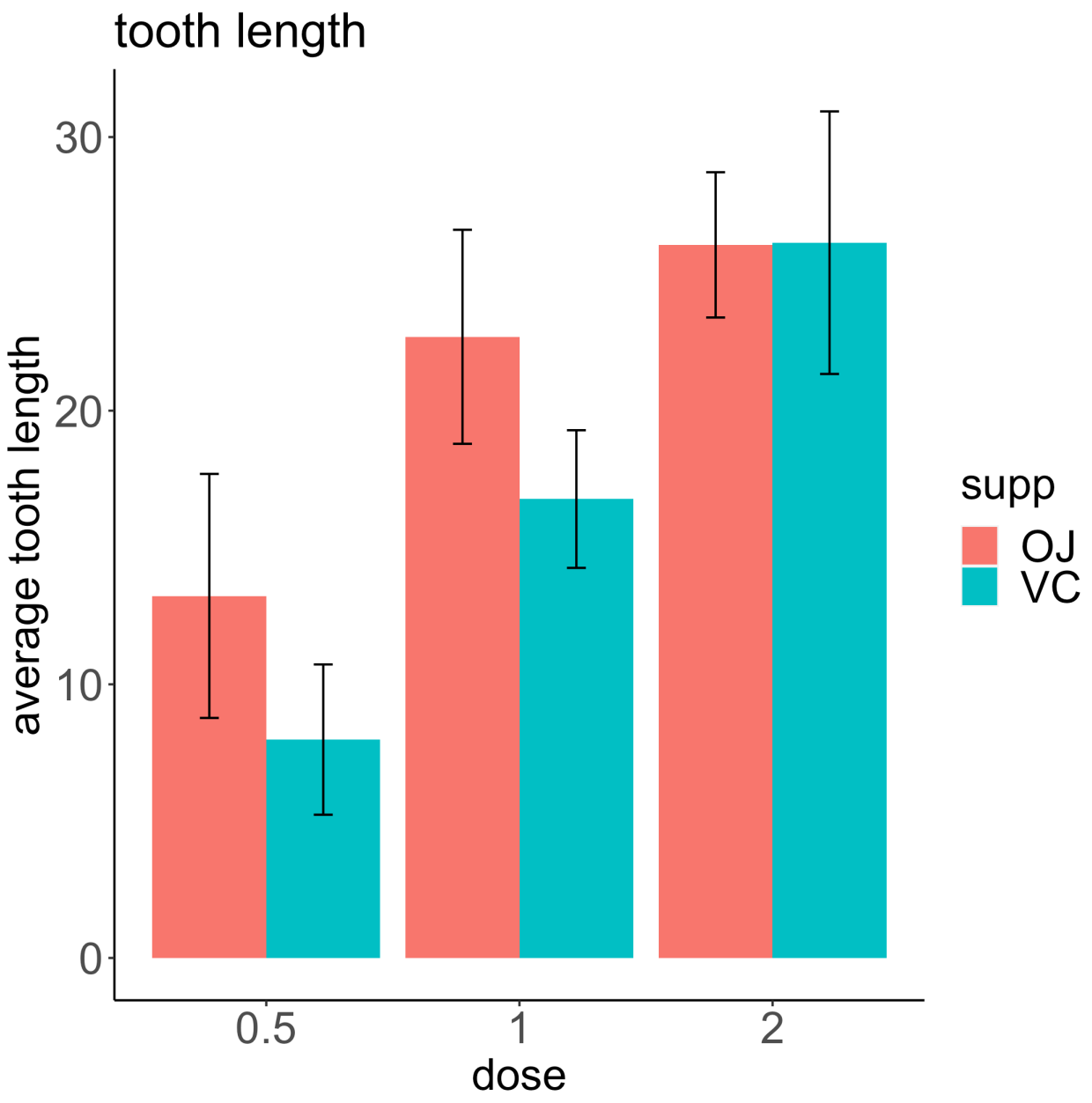
Plotting condenses large datasets

- 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

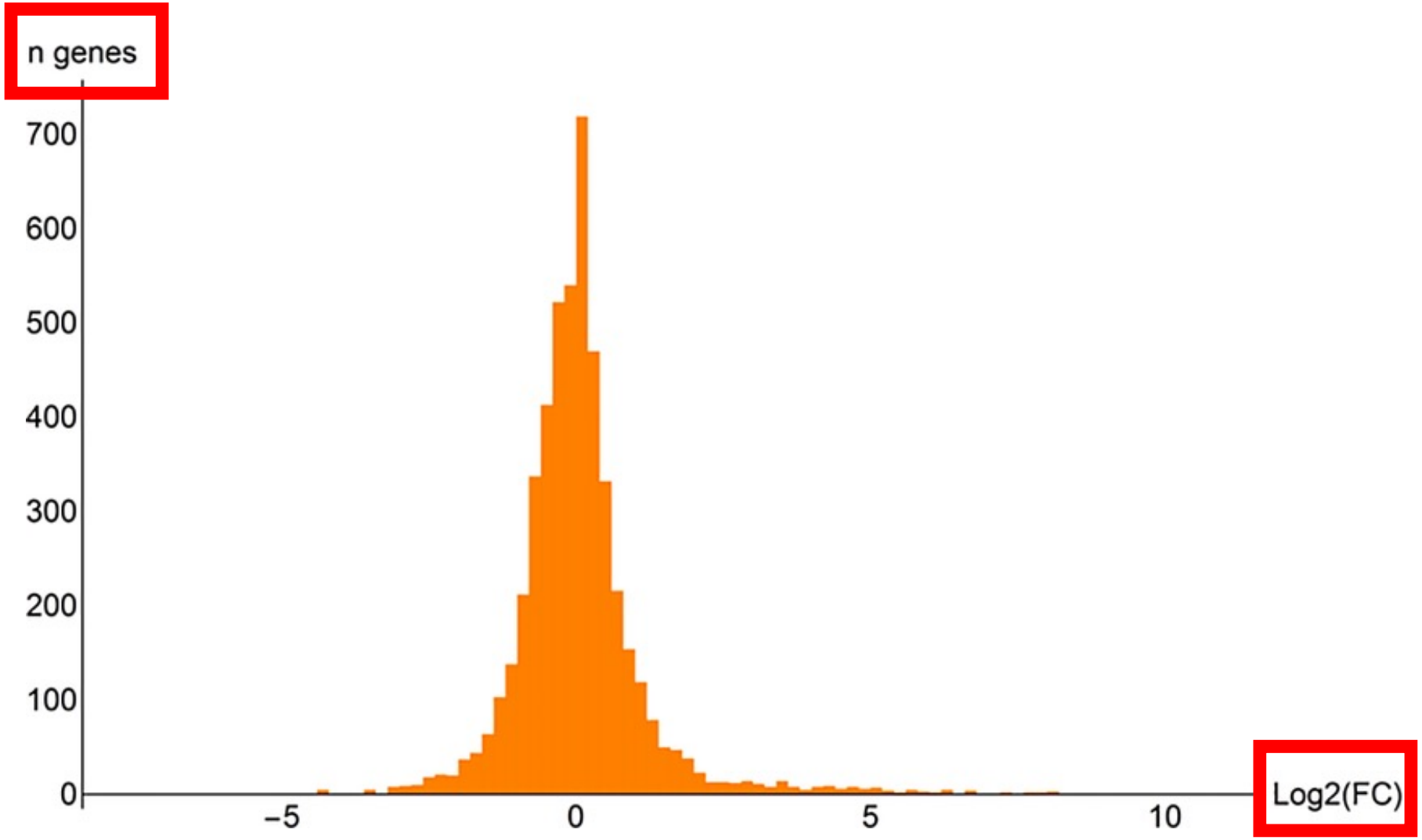
Shows magnitude or mean/variation across categorical or discrete variables



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

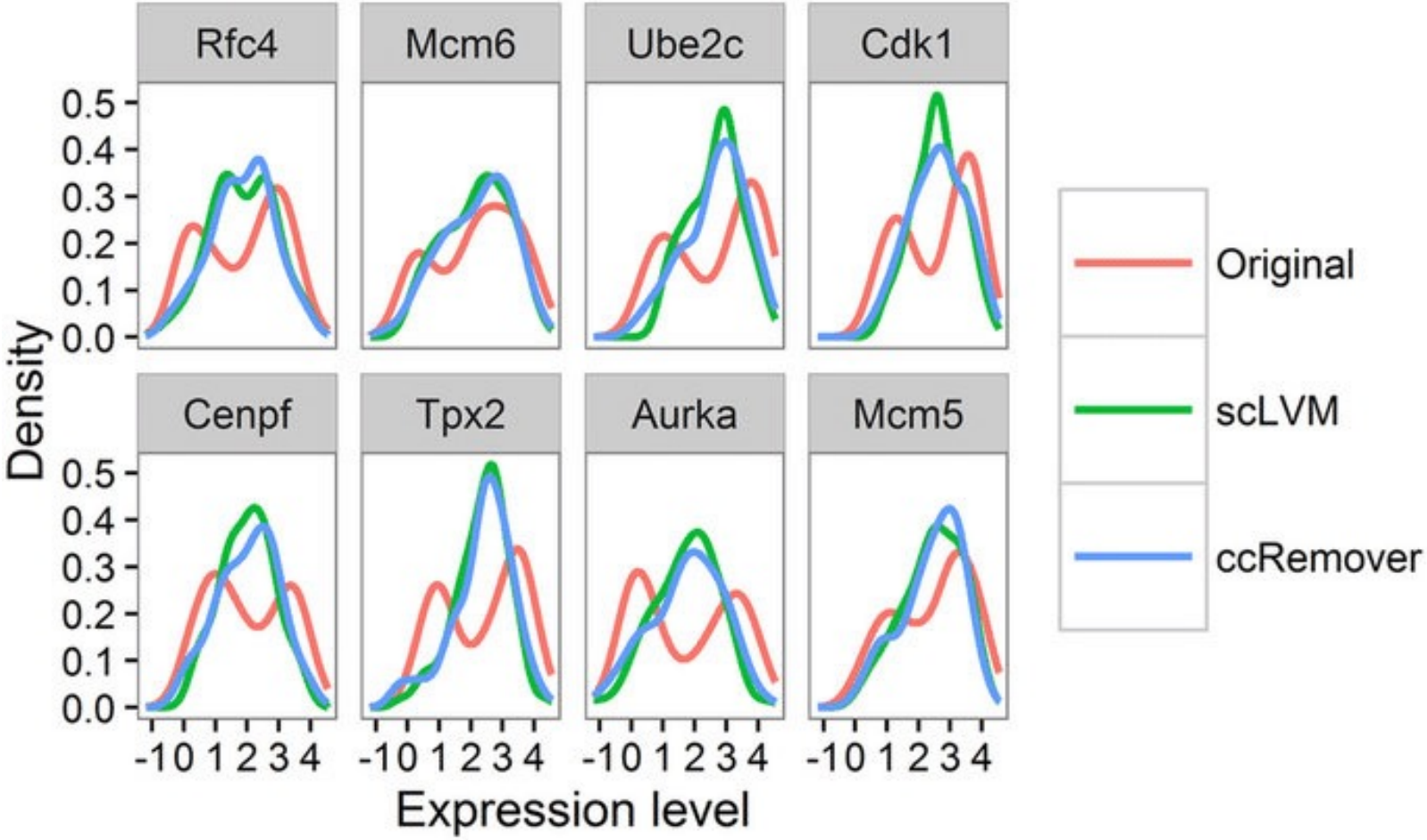
- Shows occurrence along numeric intervals



Schabort et al (PLoS On e, 2016, https://pubmed.ncbi.nlm.nih.gov/27315089/)

Density plot

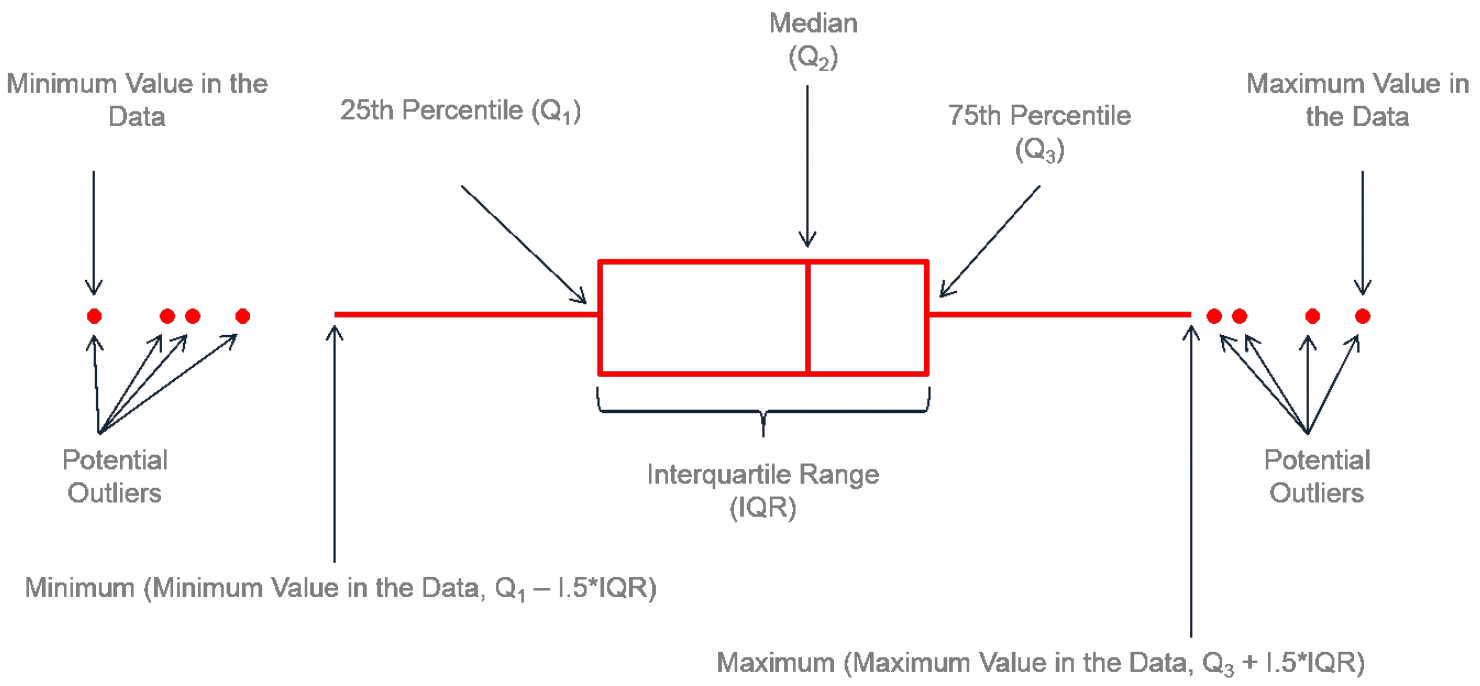
- Shows distribution distribution along numeric intervals (smoothed out version of a histogram)
- We will learn about faceting, which generates subplots in lesson 2



Barron and Li (2016, Nature Scientific Reports, https://pubmed.ncbi.nlm.nih.gov/27670849/)

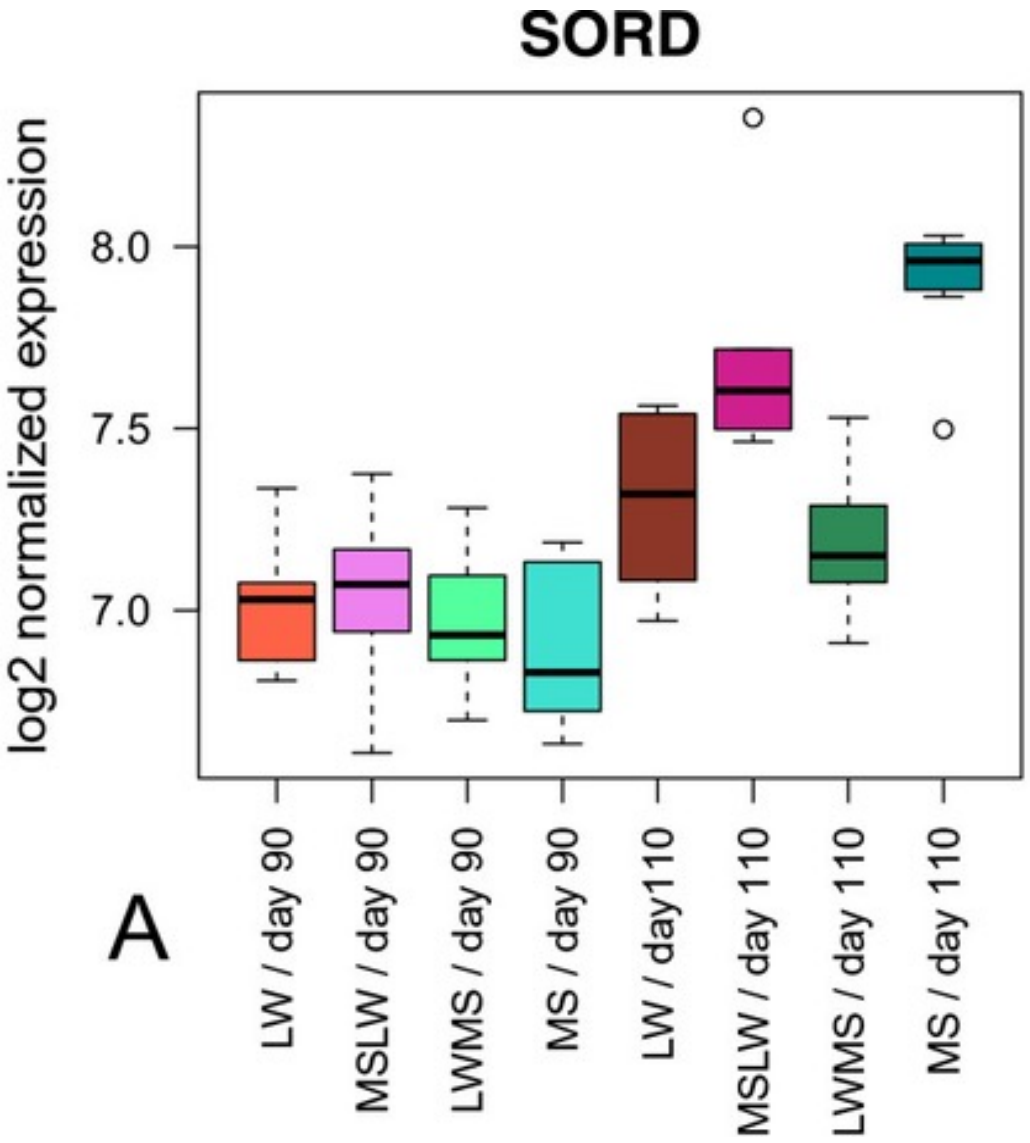
Box and whisker plot

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



https://r-graph-gallery.com/boxplot

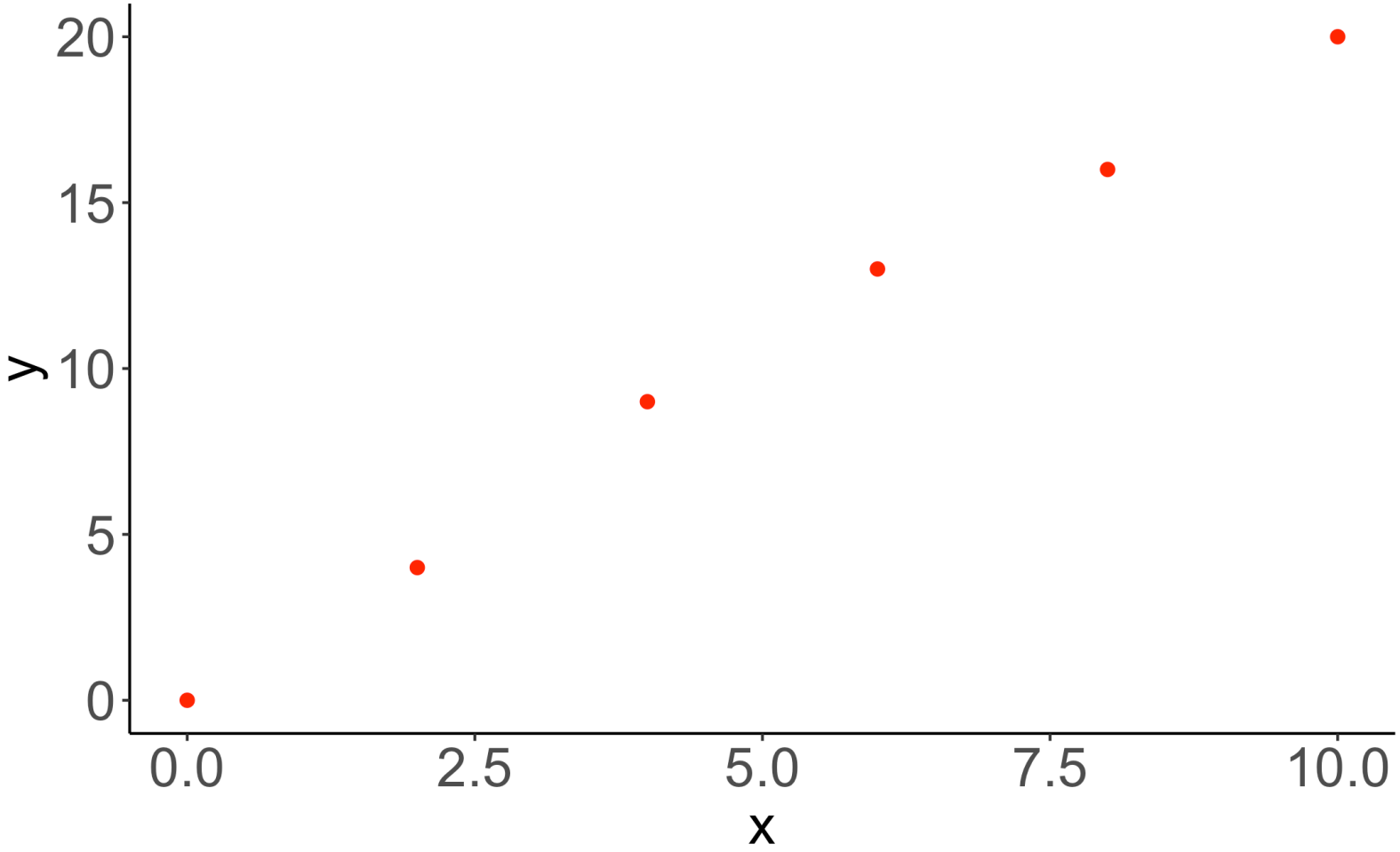
Box and whisker plot - example



Voillet et al (2014, BMC Genomics, https://pubmed.ncbi.nlm.nih.gov/25226791/)

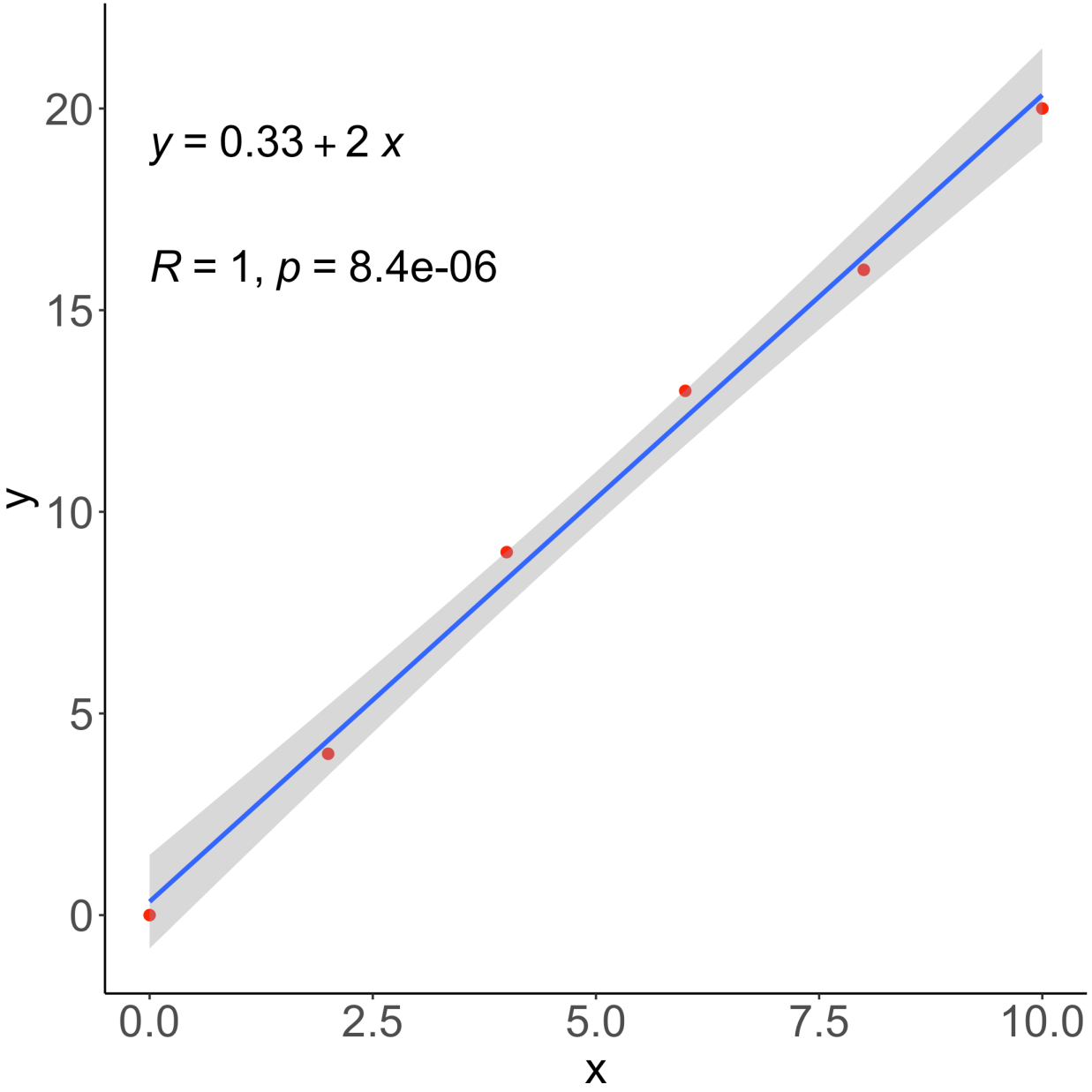
Scatter plot

Shows correlation



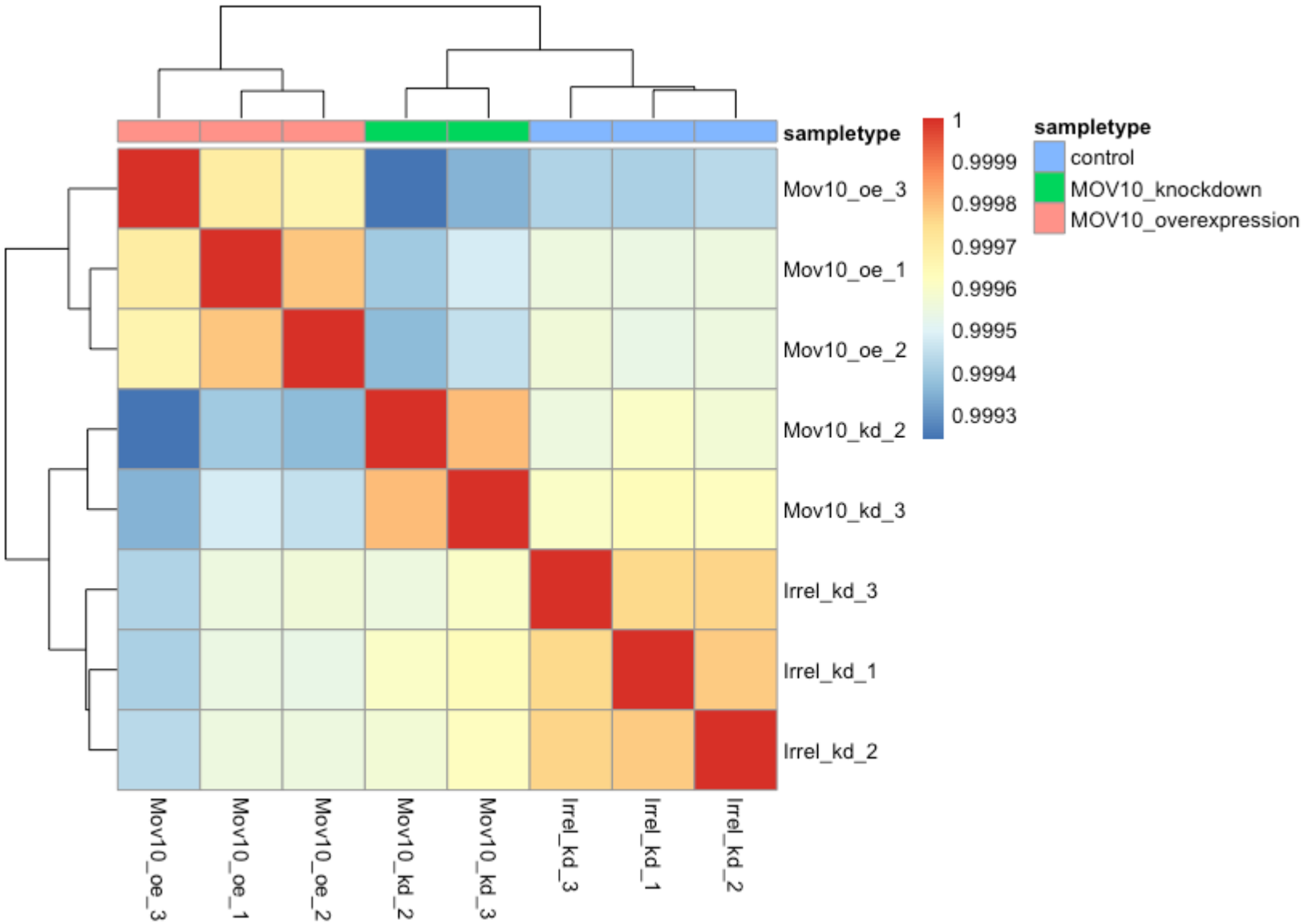
Scatter plot with regression line

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

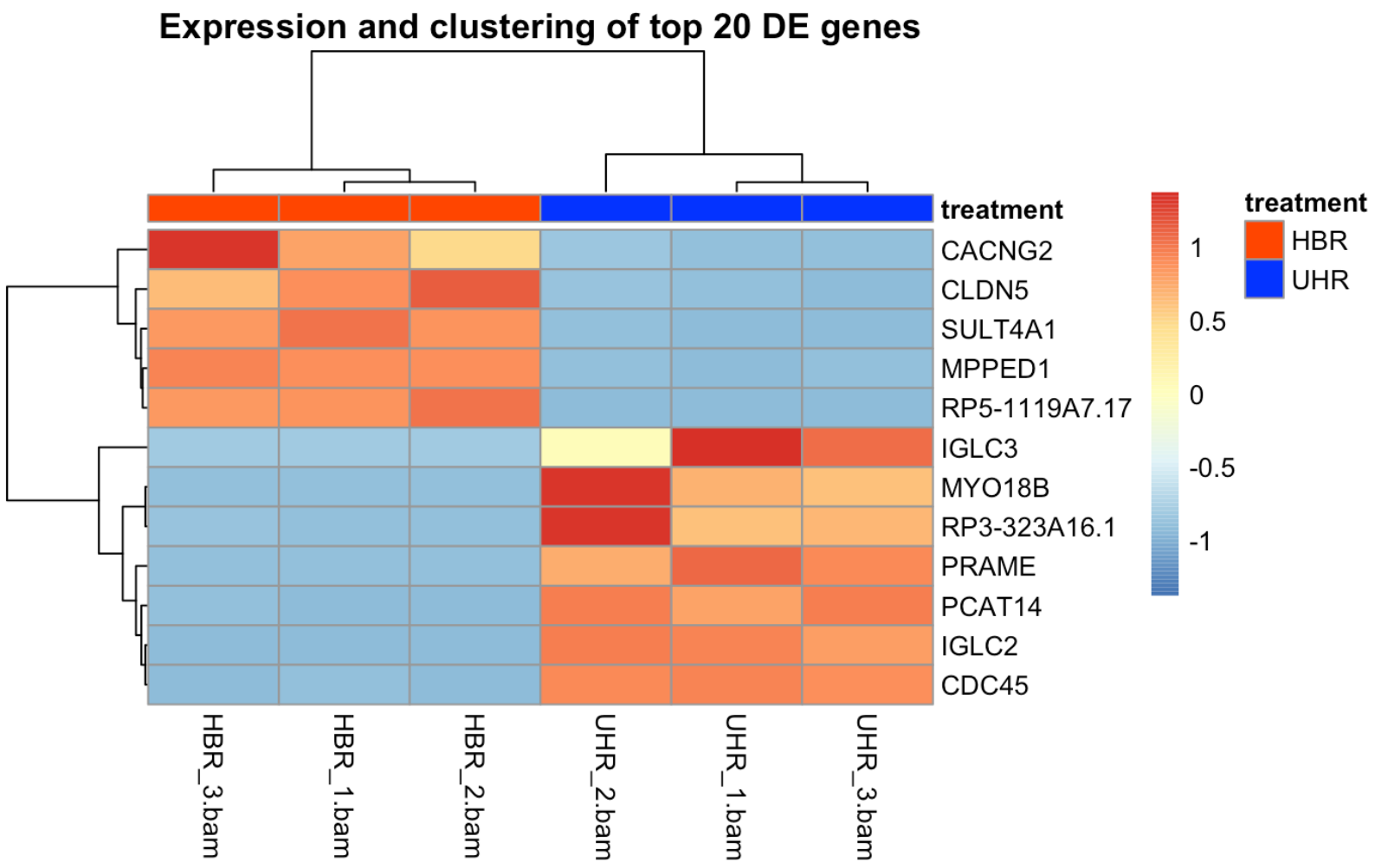
- Allows for visualization of correlations and clustering



Source: https://hbctraining.github.io/DGE_workshop_salmon/lessons/03_DGE_QC_analysis.html

Heatmap and dendrogram

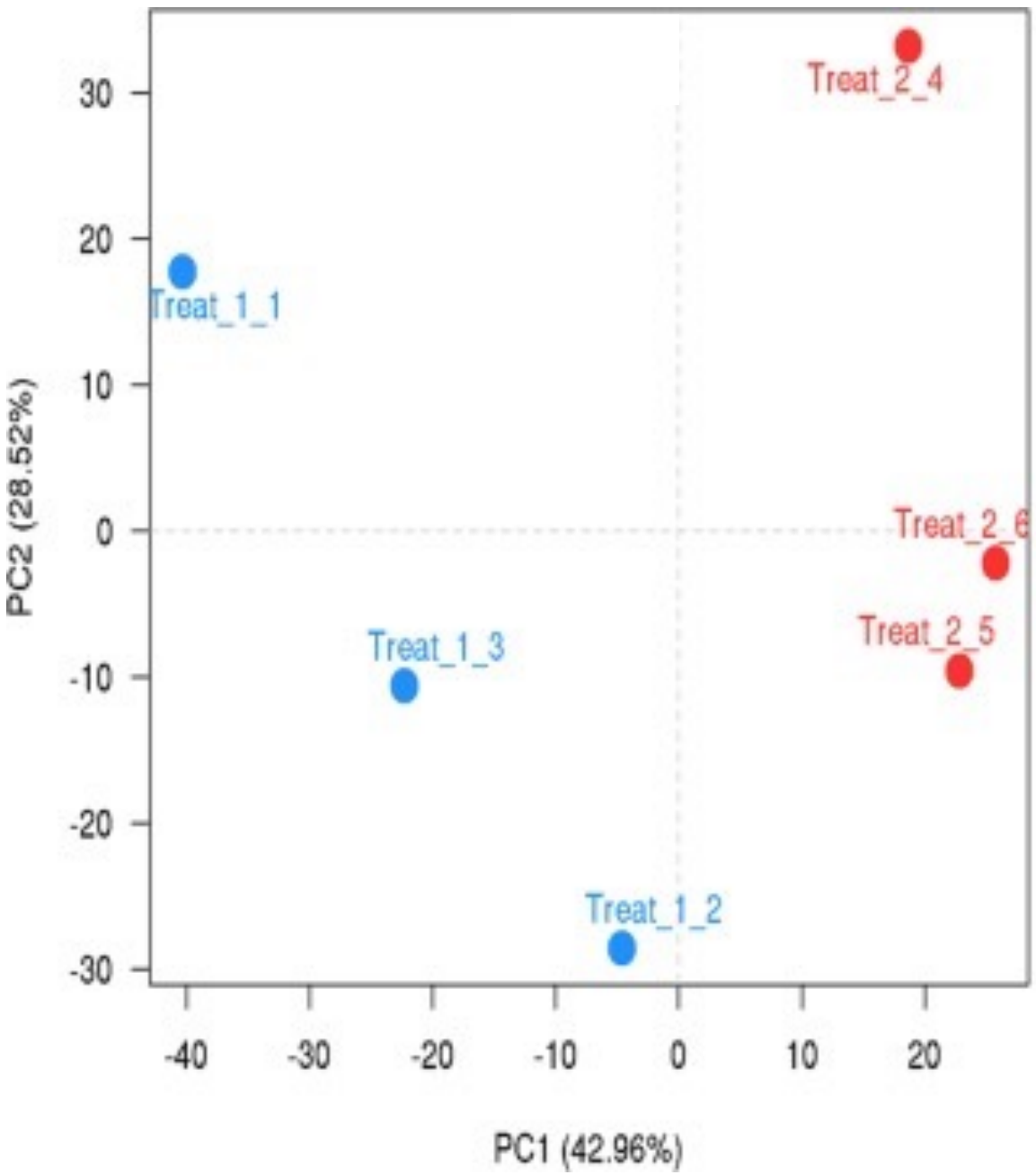
- Allows for visualization of correlations and clusters



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

Principal component analysis (PCA)

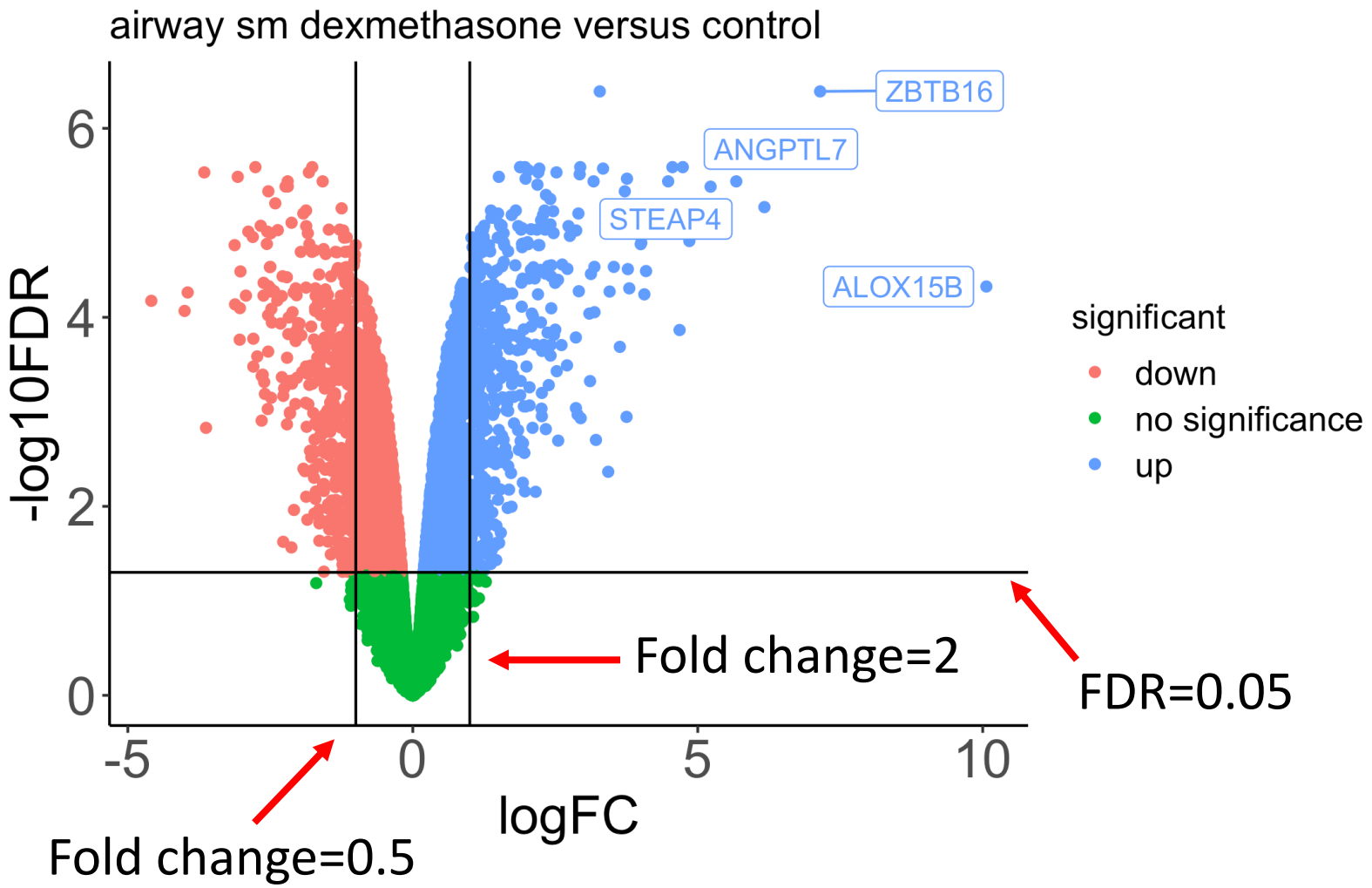
- Allows for visualization of clusters
- If using a differential expression package, we can import PCA results into ggplot2 to make the plot



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

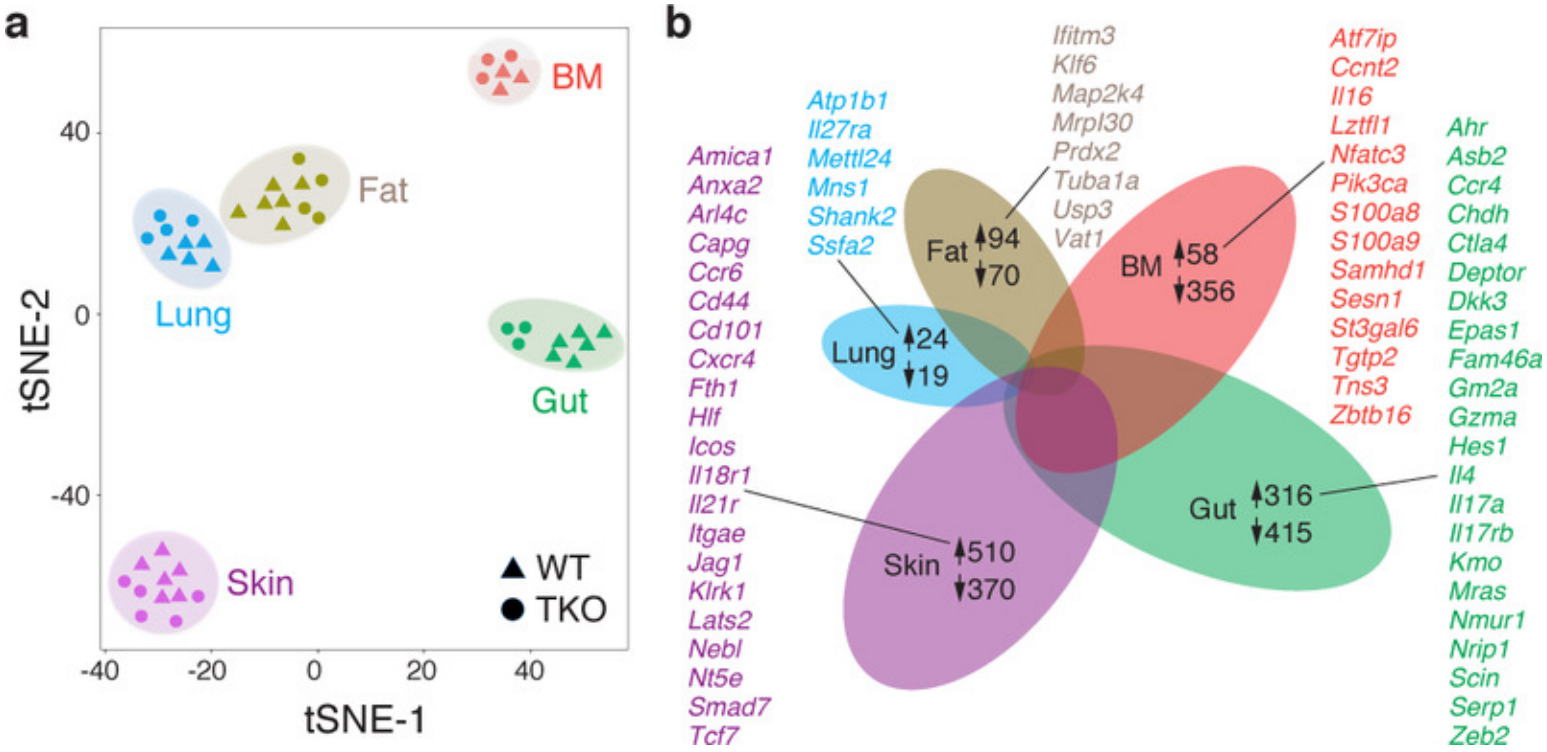
Volcano plot

Differential gene expression results from the airway dataset (https://bioconductor.org/packages/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.



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, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6202223/)

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=<DATA>) + <GEOM_FUNCTION> (mapping = aes(<MAPPING>))
```

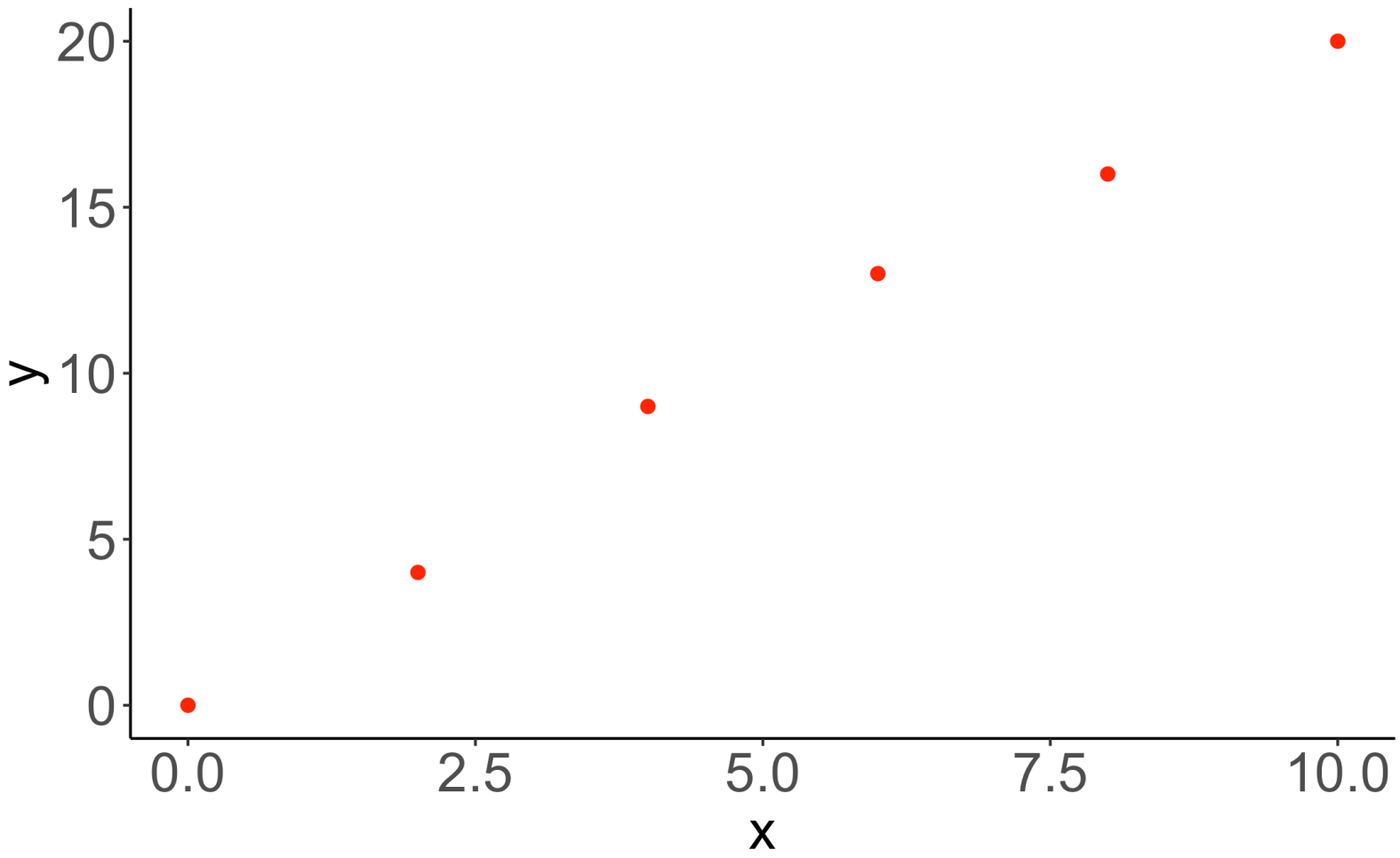
Data that we want to plot

Specify what type of plot we want

Specify things like what we want for x and y axes

```
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 https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r
- 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
  - Importing dataset
  - Generating plots from the dataset – so we will have plenty of opportunity to become familiar with ggplot2
- Lesson 2 will highlight
  - Basic ggplot2 syntax
  - Geoms
  - Faceting
- For detail course outline, visit the course page at https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r