

# Getting started with R

## Essential R for data manipulation

JGI Data Science Training  
Wyatt Hartman  
Feb 14, 2018

Slides and R code notebook available at

<https://github.com/JGI-Bioinformatics/JGI-Data-Science/tree/master/R/Getting-started-with-R>

# What is R ?



Open source statistical programming language

Forked from “S” language developed by Bell Labs (1997)

Used by about 50% of data scientists

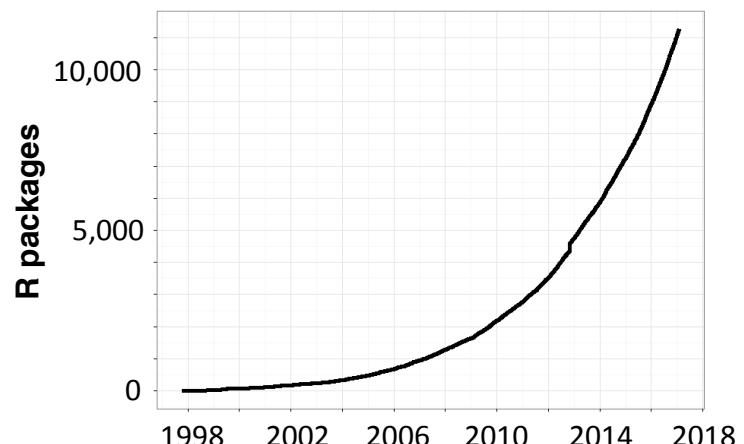
<https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html/>

“Least disliked” programming language

<https://stackoverflow.blog/2017/10/31/disliked-programming-languages/>

“Academic Language”

- New methods often deployed first
- Lots of learning resources
- vs. Python...



# Why use R ?

Stop clicking, start coding

Eliminate repetition, make  
reproducible workflows

Build on top of work  
you've already done

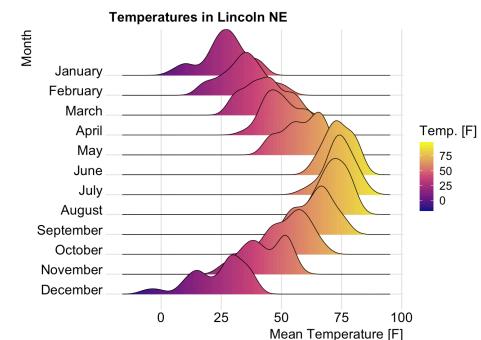
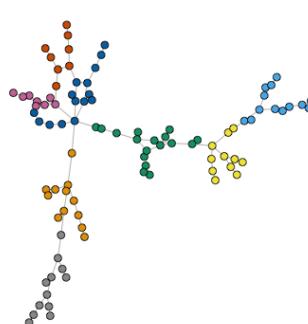
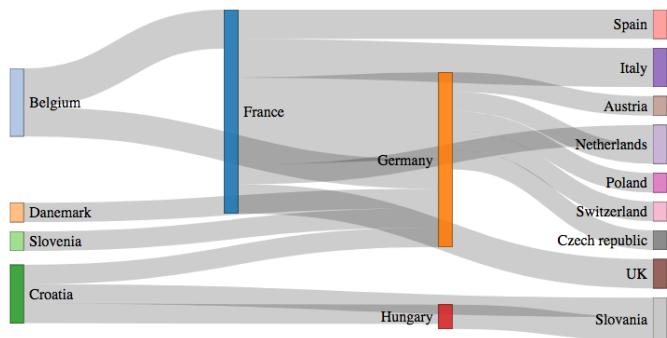
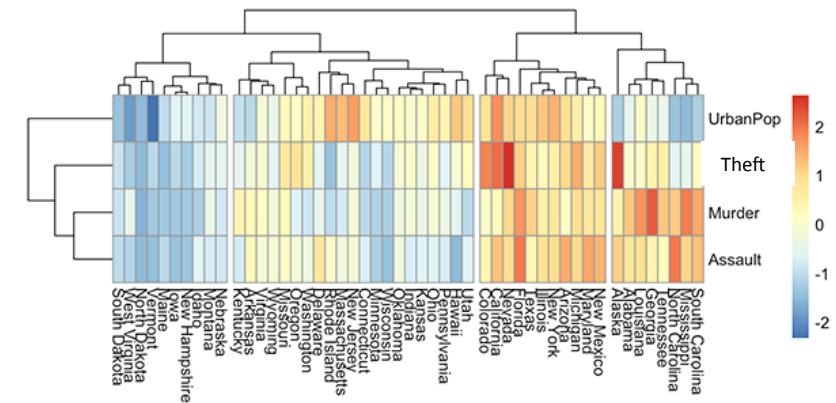
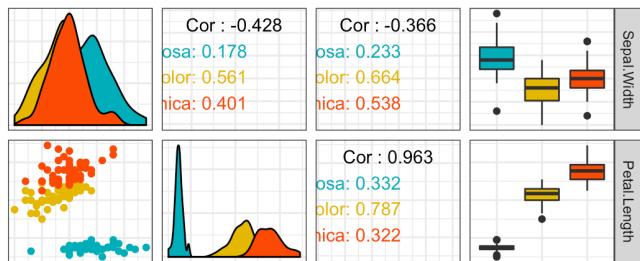
Stand on the shoulders of  
those who made code for you



<http://www.computergeekblog.com/7-excel-tips-tricks-impress-boss/>

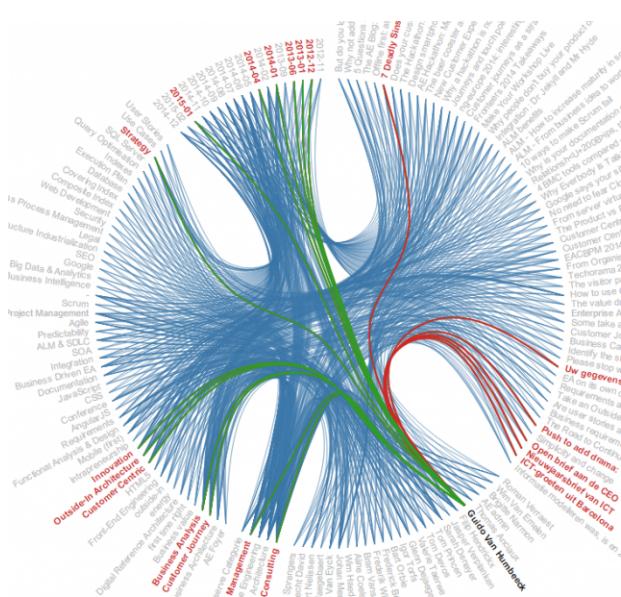
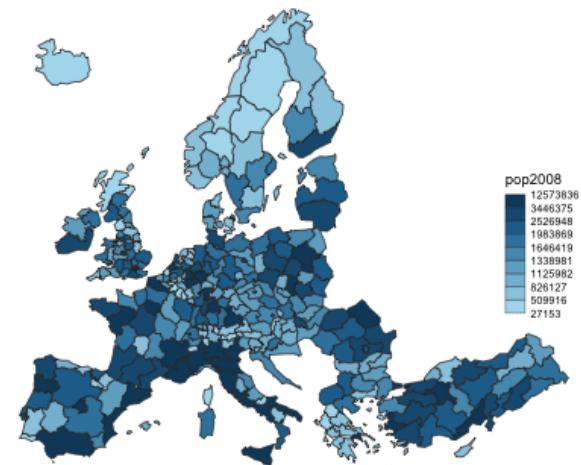
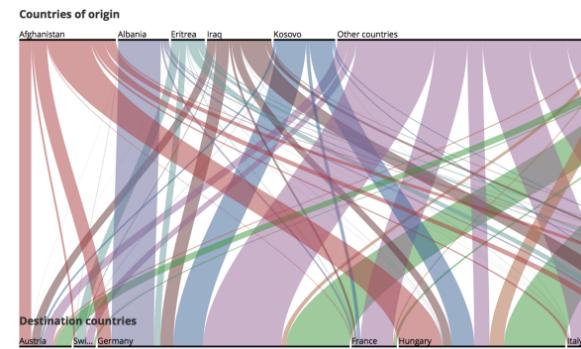
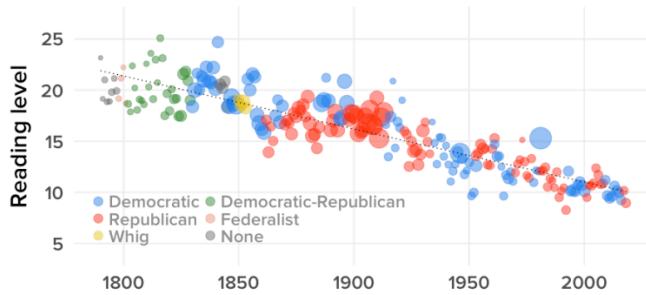
# Why use R ?

## Striking data visualizations



# Why use R ?

## Polished, publication quality graphics

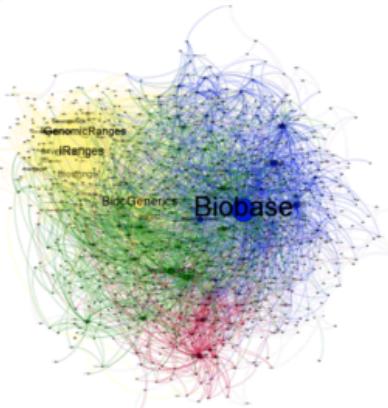
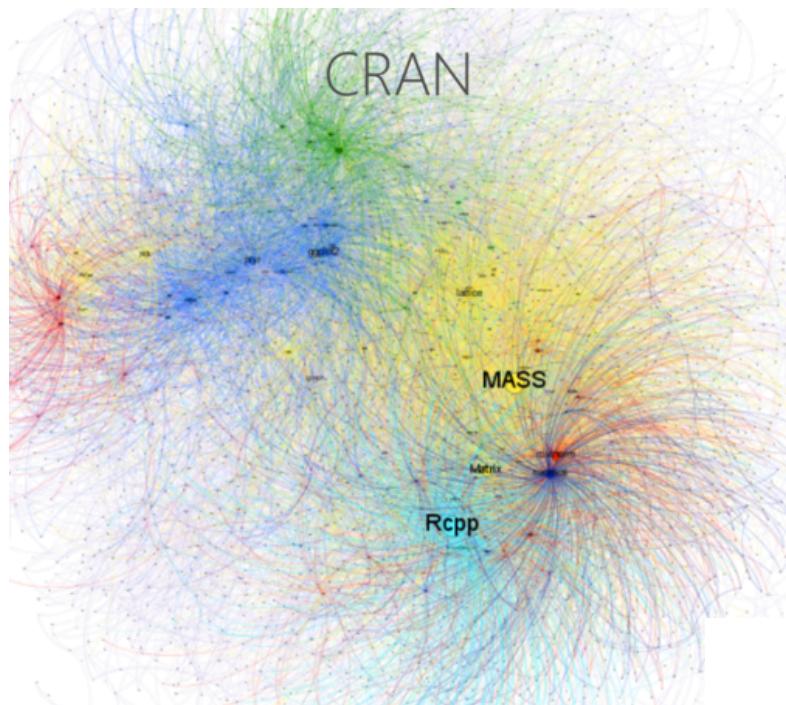


# Why use R ?

# Extensive and powerful statistical libraries

> 11,000 R packages in CRAN  
Comprehensive R Archive Network

> 6,000 Bioconductor packages  
**for analysis of genomic data**



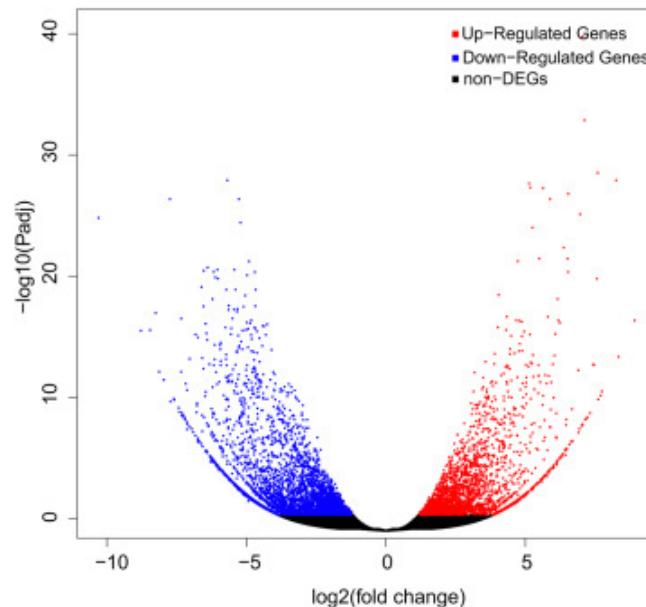
<http://blog.revolutionanalytics.com/2015/08/differences-in-the-network-structure-of-cran-and-bioconductor.html>

# Why use R ?

## Bioconductor: Popular sequence data workflows

### Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

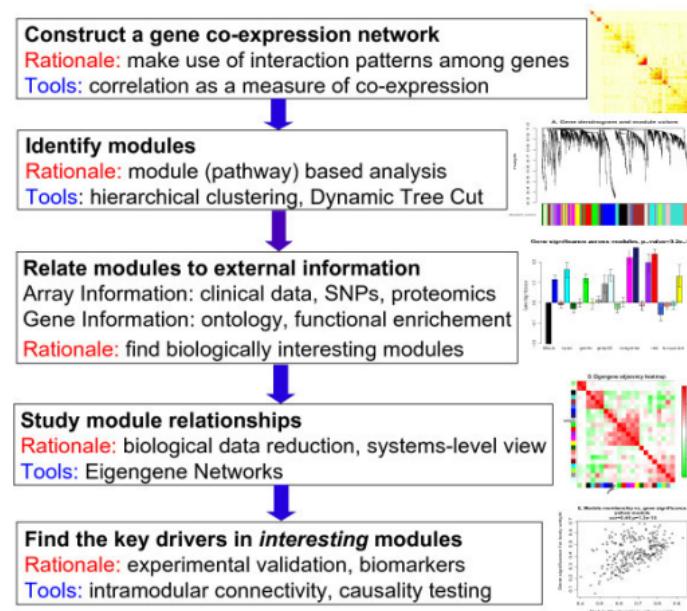
Michael I Love<sup>1,2,3</sup>, Wolfgang Huber<sup>2</sup> and Simon Anders<sup>2\*</sup>



(2014) Genome Biol. 15: 550

### WGCNA: an R package for weighted correlation network analysis

Peter Langfelder<sup>1</sup> and Steve Horvath<sup>\*2</sup>

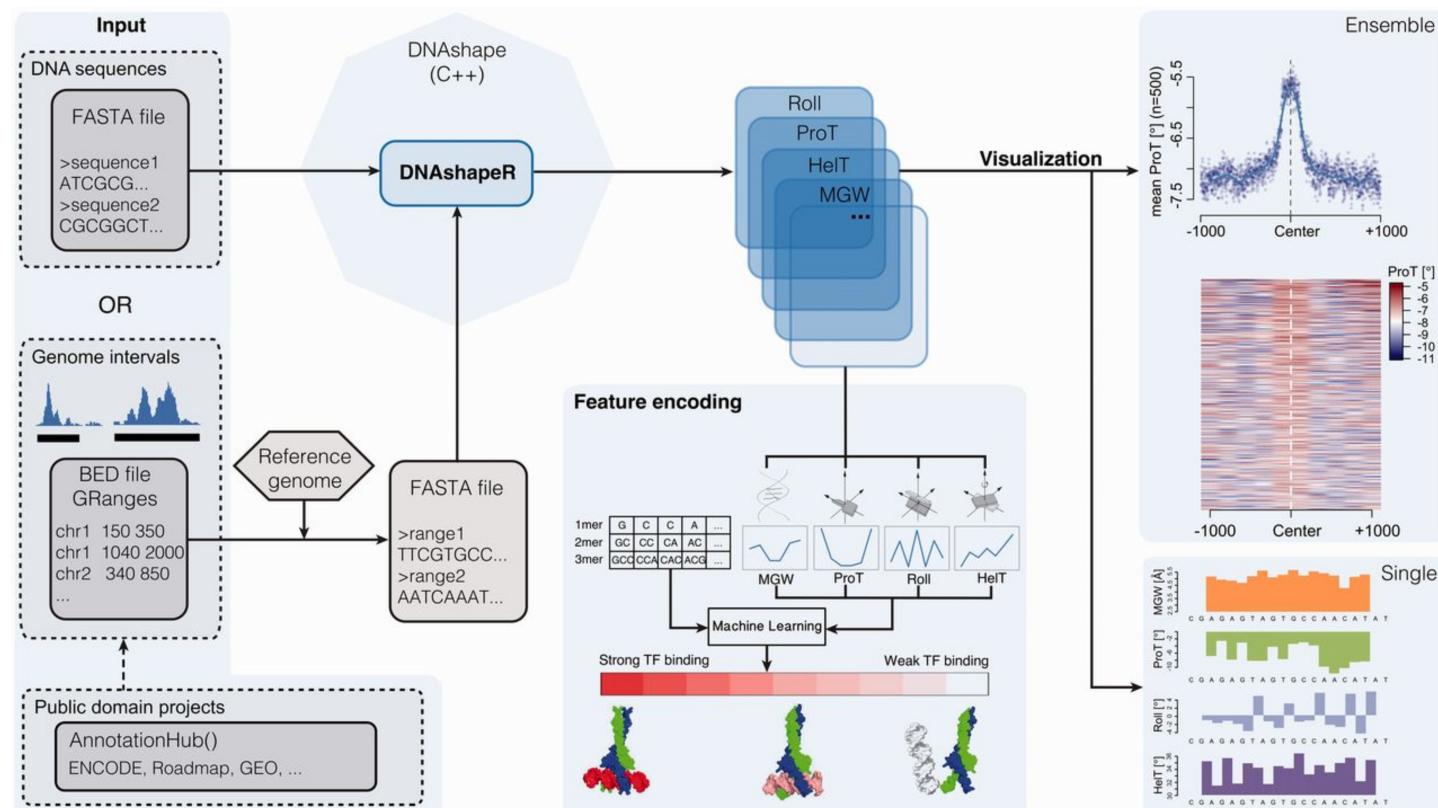


(2008) BMC Bioinf. 9: 559

# Why use R ?

## Intensive genomics analyses in Bioconductor

DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding



# Why use R ?

## Data manipulation and graphing libraries

Modern, well documented libraries  
cover the basics, cleanly



### Data Wrangling with dplyr and tidyverse



#### Syntax - Helpful conventions for wrangling

`dplyr::tbl_df(iris)`

Converts data to `tbl` class. `tbl`'s are easier to examine than data frames. R displays only the data that fits onscreen:

```
Source: local data frame (150 x 5)
  Sepal.Length Sepal.Width Petal.Length
1          5.1         3.5          1.4
2          4.9         3.0          1.4
3          4.7         3.2          1.3
4          4.6         3.1          1.5
5          5.0         3.6          1.4
...           ...
Variables not shown: Petal.Width (dbl), Species (fctr)
```

`dplyr::glimpse(iris)`

Information dense summary of `tbl` data.

`utils::View(iris)`

View data set in spreadsheet-like display (note capital V).

### Tidy Data

In a tidy data set:  
 &   
 Each variable is saved in its own column  
 Each observation saved in its own row

### Reshaping Data

`tidyverse::gather(cases, "year", "n", 2:4)`  
 Gather columns into rows.  
  
`tidyverse::separate(storms, date, c("y", "m", "d"))`  
 Separate one column into several.  


### Subset Observations (Row Selection)

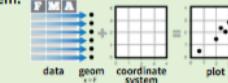


### Data Visualization with ggplot2



#### Basics

`ggplot2` is based on the **grammar of graphics**, the idea that you can build every graph from the same few components: a **data set**, a set of **geoms**—visual marks that represent data points, and a **coordinate system**.



To display data values, map variables in the data set to aesthetic properties of the geom like **size**, **color**, and **x** and **y** locations.



### Geoms - Use a geom to represent data points

#### One Variable

##### Continuous

`a + geom_area(stat = "bin")`  
`x, y, alpha, color, fill, linetype, size`

`a + geom_density(kernel = "gaussian")`  
`x, y, alpha, color, fill, linetype, size, weight`

`a + geom_dotplot()`  
`x, y, alpha, color, fill`

`a + geom_freqpoly()`  
`x, y, alpha, color, linetype, size`

`b + geom_freqpoly(aes(y = ..density..))`

`a + geom_histogram(binwidth = 5)`  
`x, y, alpha, color, fill, linetype, size, weight`

`b + geom_histogram(aes(y = ..density..))`

##### Discrete

`b + ggplot(mpg, aes(fct))`

`b + geom_bar()`  
`x, alpha, color, fill, linetype, size, weight`

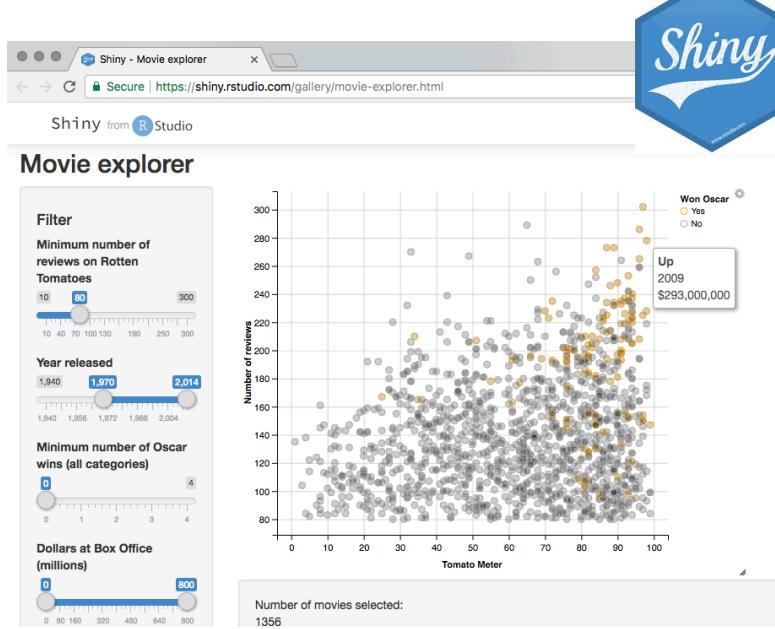
<https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf>

<https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>

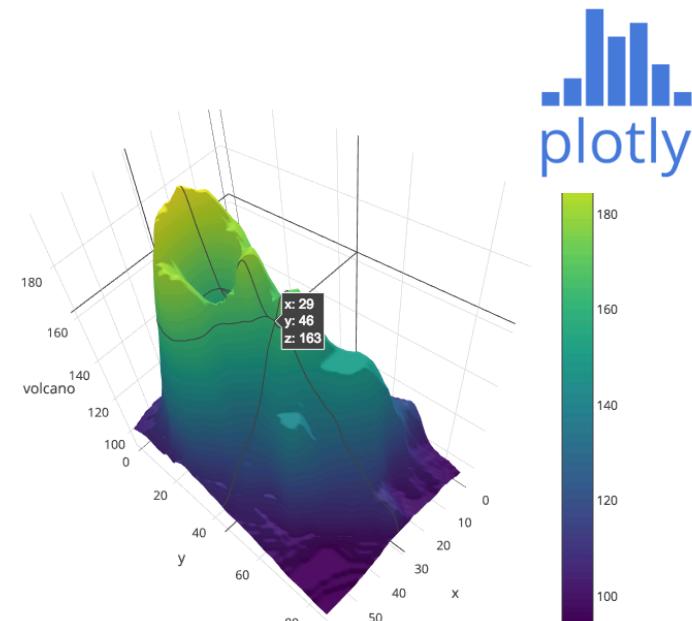
# Why use R ?

## Interactive online graphics

Build dashboards and live graphics on the web



<http://shiny.rstudio.com/tutorial/>



<https://moderndata.plot.ly/interactive-r-visualizations-with-d3-ggplot2-rstudio/>

<https://help.plot.ly/tutorials/>

# How do I use R ?

Rstudio is easiest to install yourself



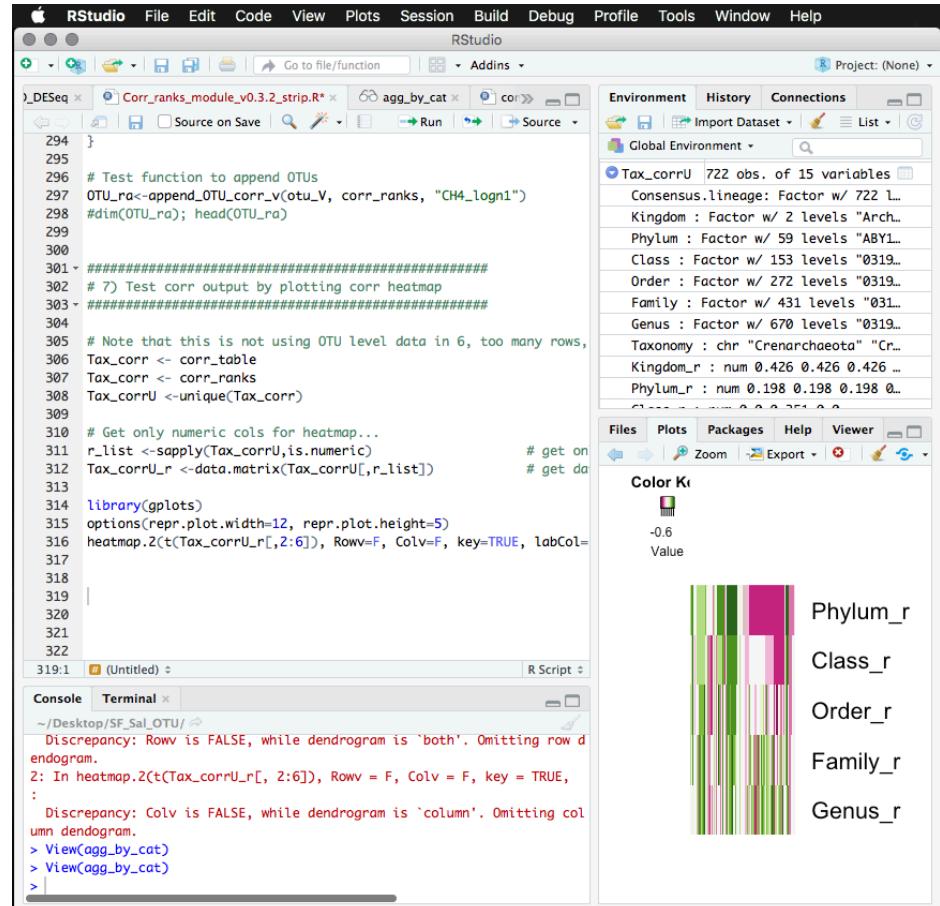
Easy to install software  
Mac – Win - Linux

Helps track code, bugs,  
objects and graphics

Use Rstudio on NERSC:  
<https://rstudio.nersc.gov/>

Install R studio:

<https://www.rstudio.com/products/rstudio/download/>



A screenshot of the RStudio interface. The top menu bar includes RStudio, File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Window, and Help. A toolbar below the menu has icons for file operations like Open, Save, and Run. The main workspace shows an R script titled '\_DESeq' with several lines of R code. To the right of the script is the Global Environment pane, listing variables like 'Tax\_corrU' (722 obs. of 15 variables) and their types and levels. Below the environment is a 'Color Key' for a heatmap, ranging from -0.6 to Value. On the far right, there's a vertical bar chart titled 'Phylum\_r' with categories 'Class\_r', 'Order\_r', 'Family\_r', and 'Genus\_r'.

```
294 }
295
296 # Test function to append OTUs
297 OTU_ra<-append_OTU_corr_V(OTU_V, corr_ranks, "CH4_login1")
298 #dim(OTU_ra); head(OTU_ra)
299
300
301 #####7#####
302 # 7) Test corr output by plotting corr heatmap
303 #####
304
305 # Note that this is not using OTU level data in 6, too many rows,
306 Tax_corr <- corr_table
307 Tax_corr <- corr_ranks
308 Tax_corrU <-unique(Tax_corr)
309
310 # Get only numeric cols for heatmap...
311 r_list <-sapply(Tax_corrU,is.numeric)
312 Tax_corrU_r <-data.matrix(Tax_corrU[,r_list])
313
314 library(gplots)
315 options(repr.plot.width=12, repr.plot.height=5)
316 heatmap.2(t(Tax_corrU_r[, 2:6]), Rowv=F, Colv=F, key=TRUE, labCol=
317
318 |
319 |
320
321
322
319:1 (Untitled) R Script
Console Terminal ~ /Desktop/SF_Sal_OTU/ Discrepancy: Rowv is FALSE, while dendrogram is 'both'. Omitting row dendrogram.
2: In heatmap.2(t(Tax_corrU_r[, 2:6]), Rowv = F, Colv = F, key = TRUE, :
:   Discrepancy: Colv is FALSE, while dendrogram is 'column'. Omitting column dendrogram.
> View(agg_by_cat)
> View(agg_by_cat)
>
```

# How do I use R ?

Jupyter notebooks: Portable R code



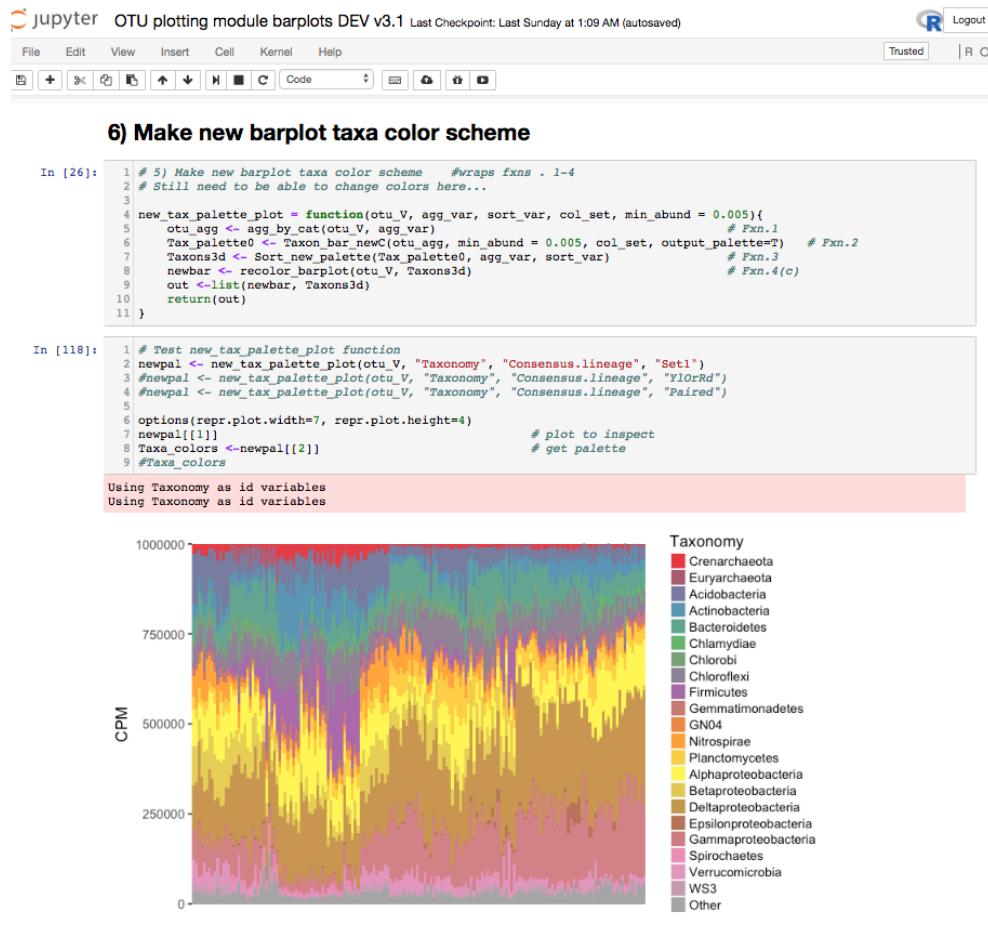
Runs in web browser,  
Github compatibility

Simplified interface,  
helps organize code

R in Jupyter on NERSC:  
<https://jupyter.nersc.gov/>

Installation more complicated,  
requires anaconda Python

<https://anaconda.org/chdoig/jupyter-and-conda-for-r/notebook>  
<https://www.datacamp.com/community/blog/jupyter-notebook-r>



# Getting started with R

## Installing packages

R studio:

```
> install.packages("ggplot2")
```

Tools -> Install Packages

Bioconductor:

<https://www.bioconductor.org/install/>

```
> source('http://bioconductor.org/biocLite.R')  
> biocLite('phyloseq')
```

## Getting help

? Before a function or package gets help

```
> ?read.table > ?ggplot2
```

- Search for tutorials, else a Google rabbit hole...
  - Look for package **vignettes**
- Excellent tutorials for beginners here:

<http://www.sthda.com/english/>

Cutting corners to meet arbitrary management deadlines



Essential

Copying and Pasting  
from Stack Overflow

O'REILLY®

The Practical Developer  
@ThePracticalDev

Statistical Tools for High-throughput Data Analysis

# Data lab objectives

Our R notebook exercise will help you learn to:

- Install and get help with R packages
- Get your data into (and out of) R
- Access and manipulate data
- Make simple data summaries and plots
- Massage data for plotting

Our next workshop covers graphics with the `ggplot2` library

- This one gives you the tools you need to be ready, and a few examples