BGGN 239

Bioinformatics for Immunologists

Barry Grant

UC San Diego

http://thegrantlab.org/bggn239/

HELLO
my name is

BARRY

bjgrant@ucsd.edu

HELLO
HIS my name is

FERHAT

fay@health.ucsd.edu

http://thegrantlab.org/bggn239

Introduce Yourself!

05:00

Introduce Yourself!

[1] Your **neighbor's** name &

[2] Place they identify with most,
Major area of study/research, &

[3] Fun fact or favorite joke!

http://thegrantlab.org/bggn239

ay-lab.github.io

Home Gmail Gcal GitHub BIMM143 BGGN213 GDrive Atmosphere CloudLaunch BIMM194 Blink News

Schedule · BGGN 239

UCSD

BGGN 239

A dedicated course to teach bioinformatics with a specific focus on its applications to important problems in immunology from the Program in Immunology, UCSD

Overview

Schedule

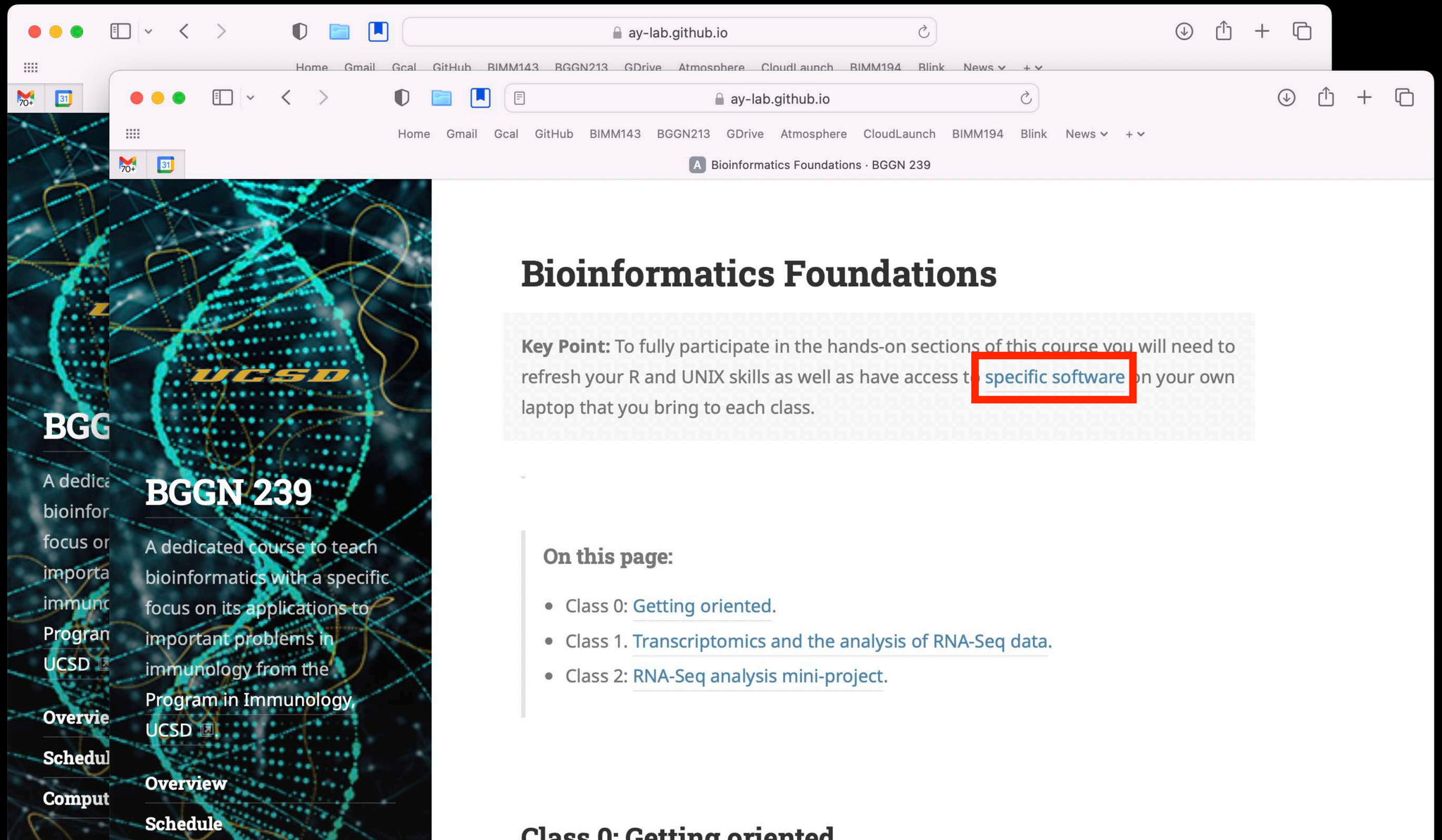
Computer Setup

Schedule

For the Spring 2023 quarter we will meet twice a week on Monday and Wednesday at 4:30-6:20 pm in TATA 2501 (Map). Clicking on the topics below will take you to supporting class content in Google Drive, hands-on "lab session" sheets, walk-through screencasts, required reading material and homework assignments.

#	Date	Topics for Spring 2023
1	Monday 04/03/23 & Wednesday 04/05/23	**Barry Grant - Recap of foundations of bioinformatics**. Topics: - Working with UNIX. - Sequence alignment. - Key online resources. - Data analysis and visualization with R and Bioconductor. - annotation of Gene lists (GO term and pathway enrichment). DriveFolder.
2	Monday 04/10/23 & Wednesday 04/12/23	**Ferhat Ay - Gene expression analysis** Topics: - RNA-seq concepts and basics. - Processing RNA-seq data. - Differential gene expression and relevant statistics. - Gene co-expression analysis. - Visualization of RNA-seq data. - Single-cell RNAseq analysis.

http://thegrantlab.org/bggn239

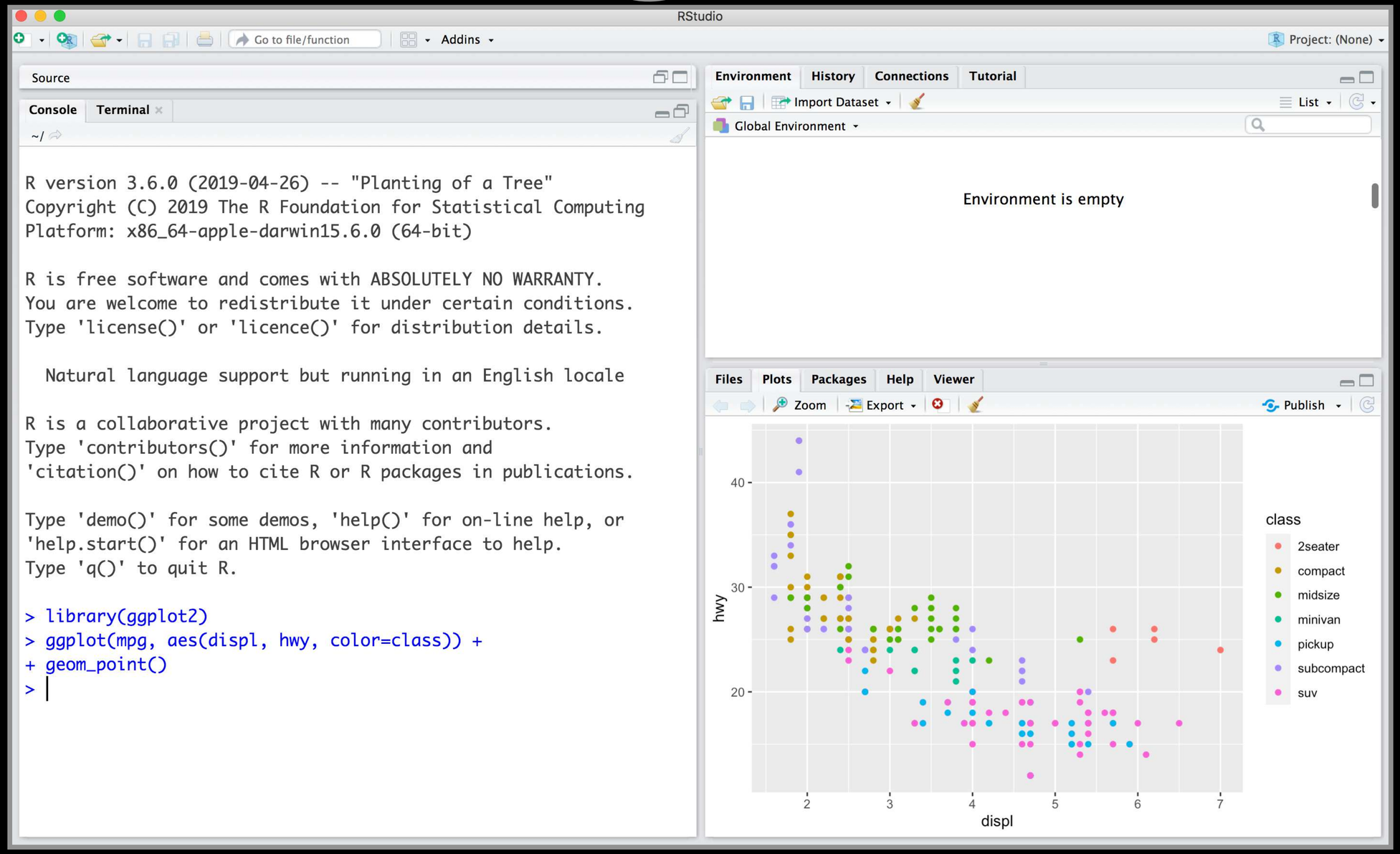


Follow Along!

Please Open RStudio®

Please Open RStudio®

Follow Along!



Please Open RStudio®

Follow Along!

RStudio

Source

Console Terminal

Input

```
R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (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.

  Natural language support but running in an English locale

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.

> library(ggplot2)
> ggplot(mpg, aes(displ, hwy, color=class)) +
+ geom_point()
>
```

Environment History Connections Tutorial

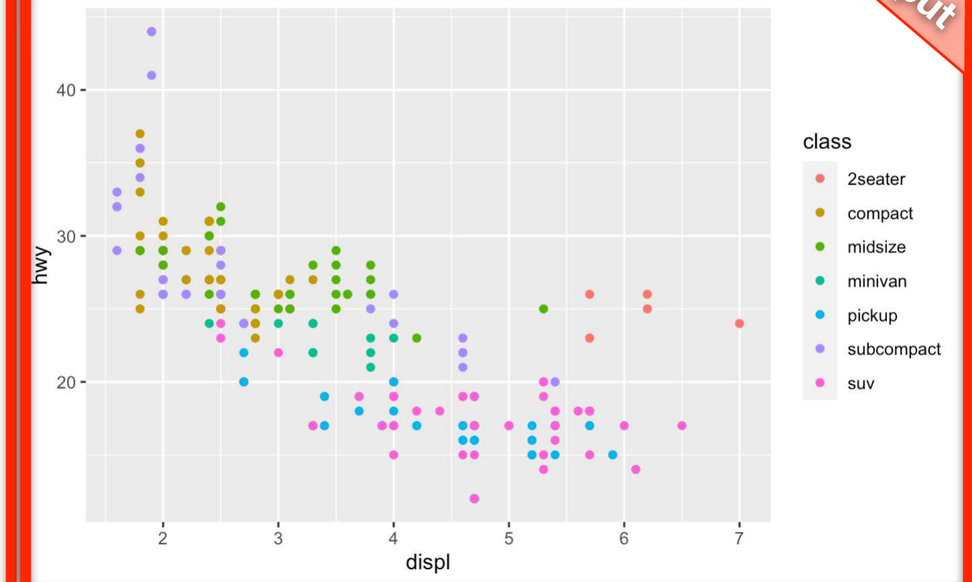
Import Dataset

Global Environment

Environment is empty

Files Plots Packages Help Viewer

Output



Please Open RStudio®

Follow Along!

RStudio

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

Source

Console | Terminal

Input

```
~/

R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (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.

  Natural language support but running in an English locale

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.

> library(ggplot2)
> ggplot(mpg, aes(displ, hwy, color=class)) +
+ geom_point()
>
```

Console

The "R Brain"

Environment | History | Connections | Tutorial

Import Dataset | List

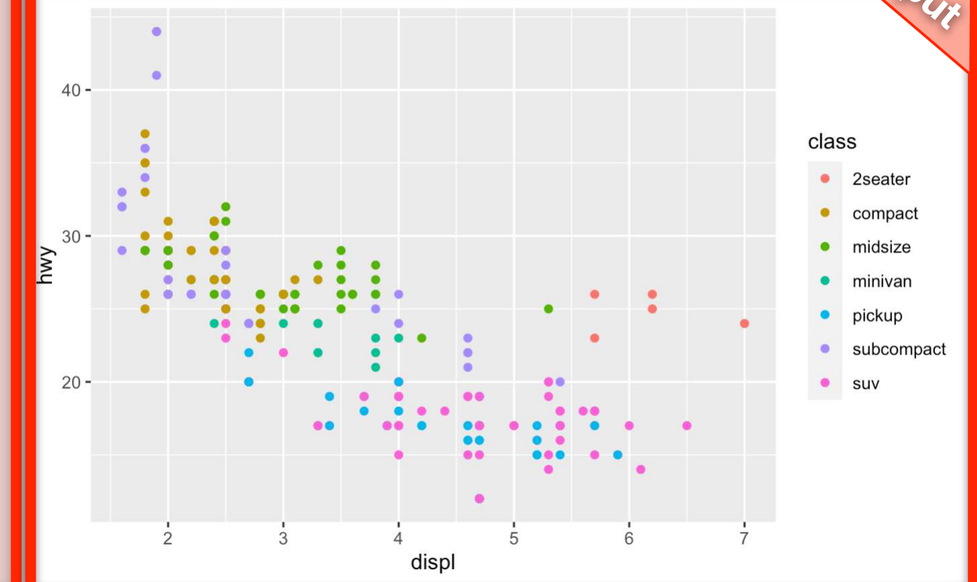
Global Environment

Environment is empty

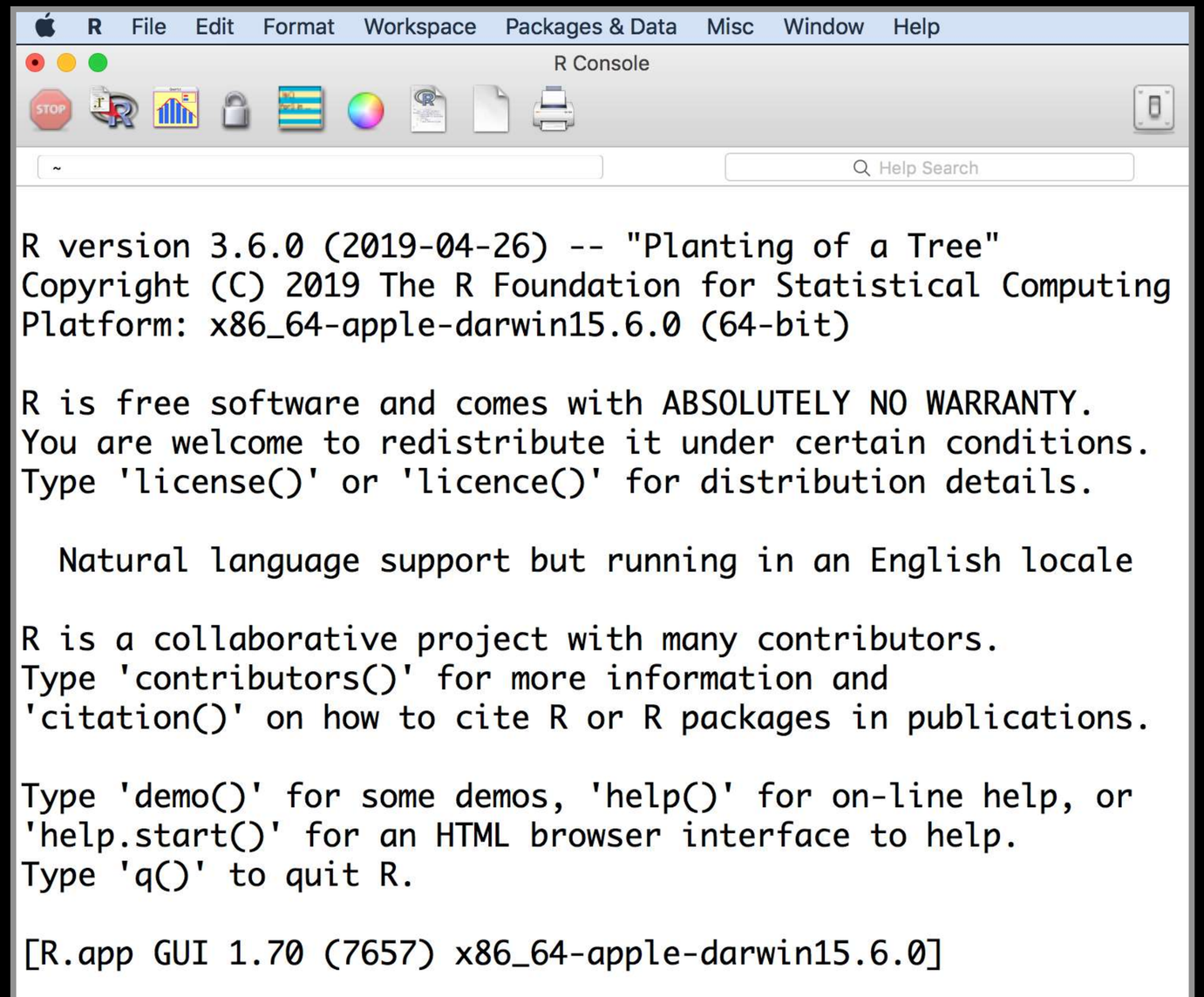
Files | Plots | Packages | Help | Viewer

Zoom | Export | Pub

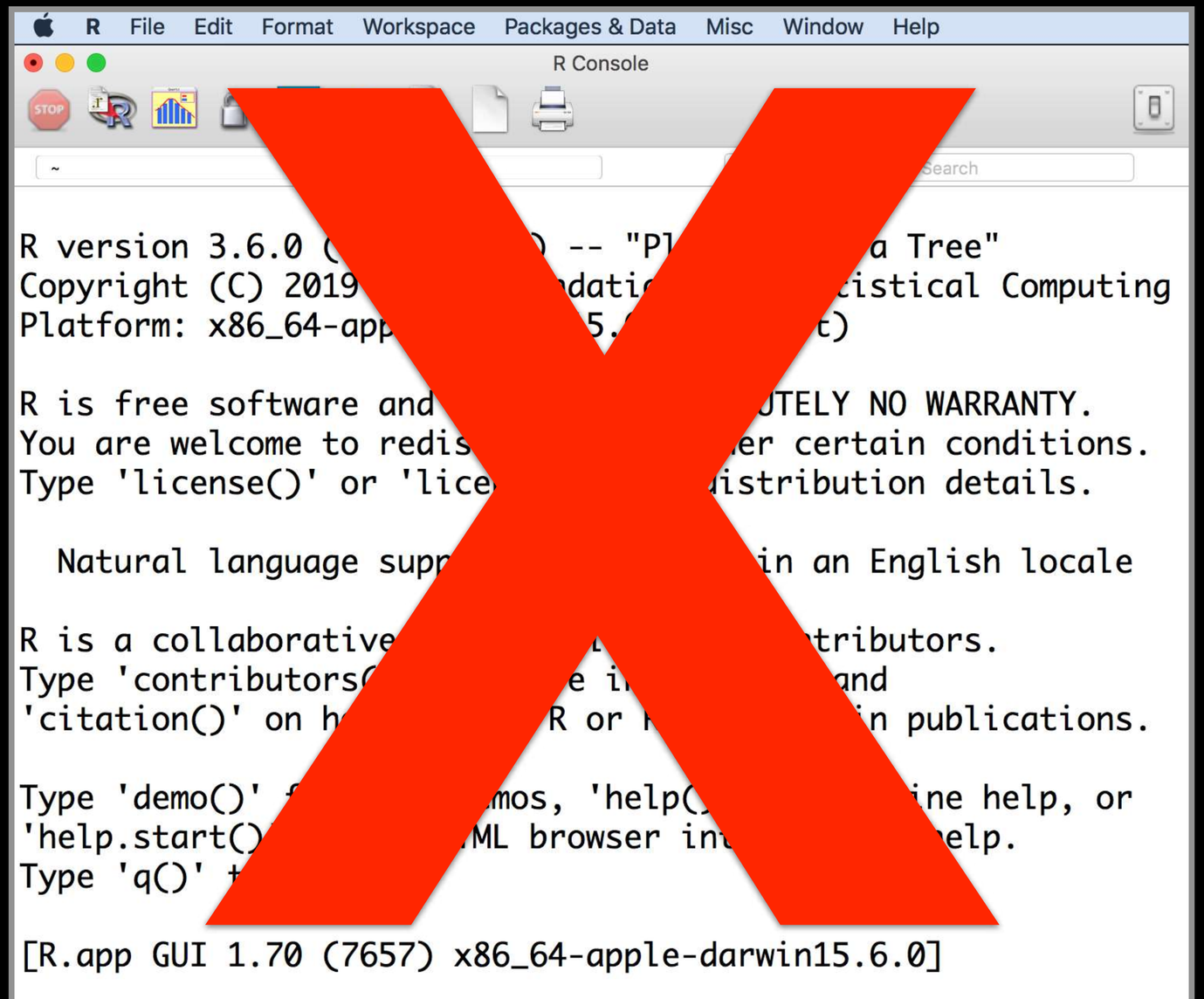
Output



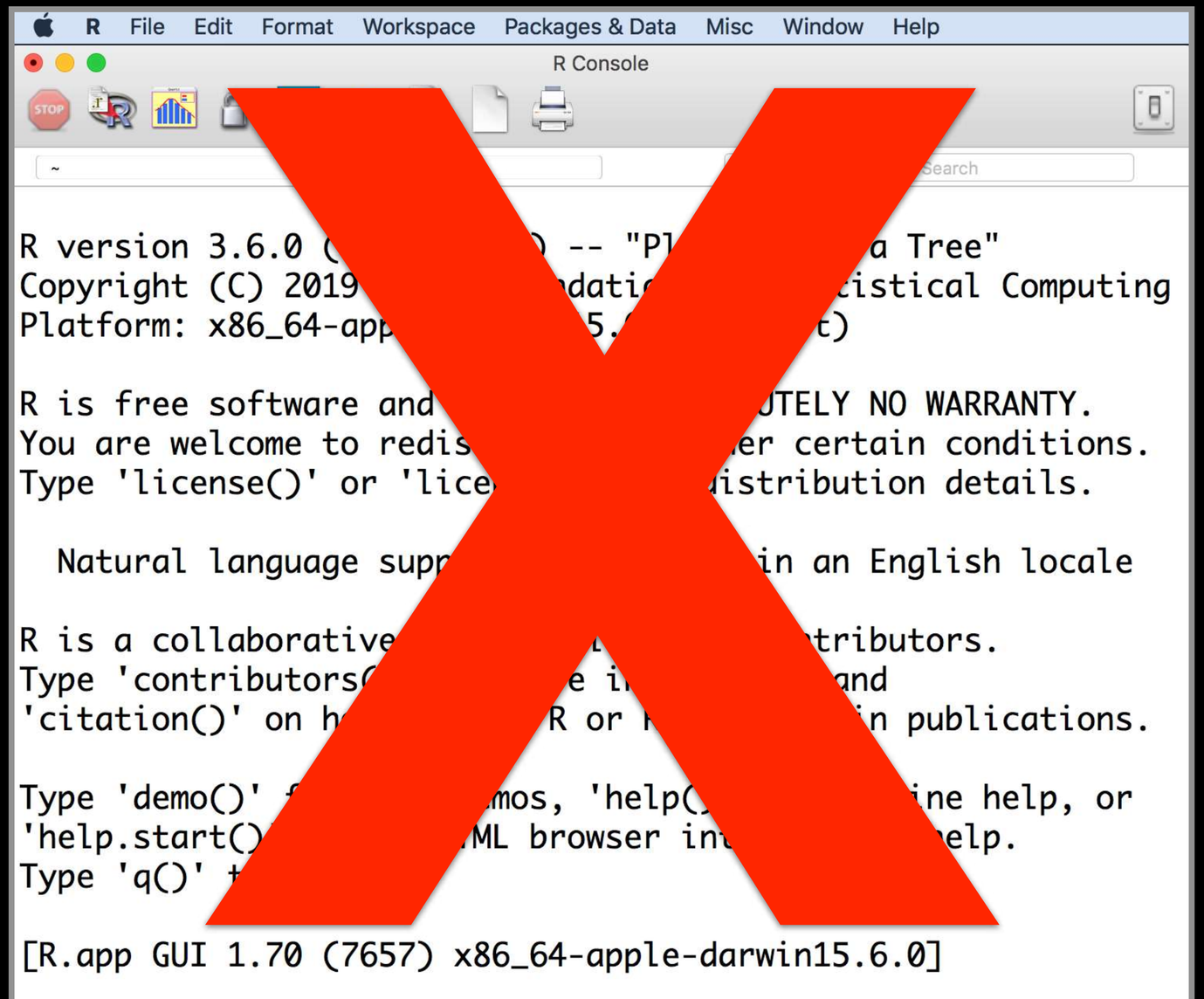
R.app GUI is NOT what we want!

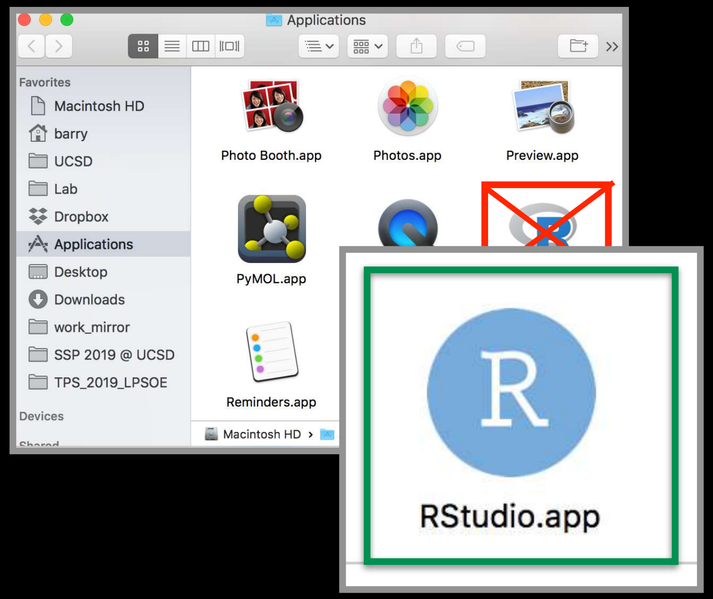


R.app GUI is **NOT** what we want!

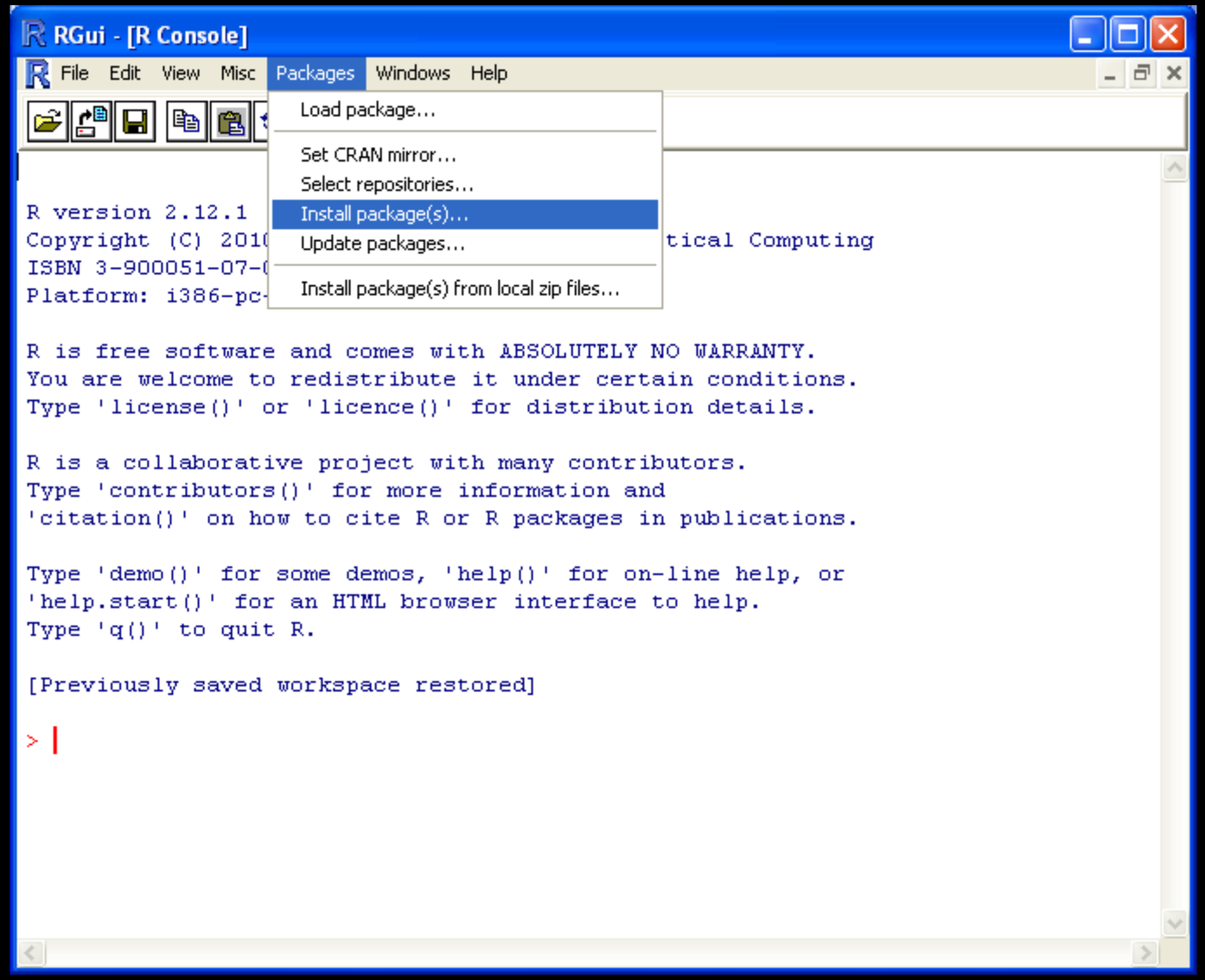


R.app GUI is **NOT** what we want!

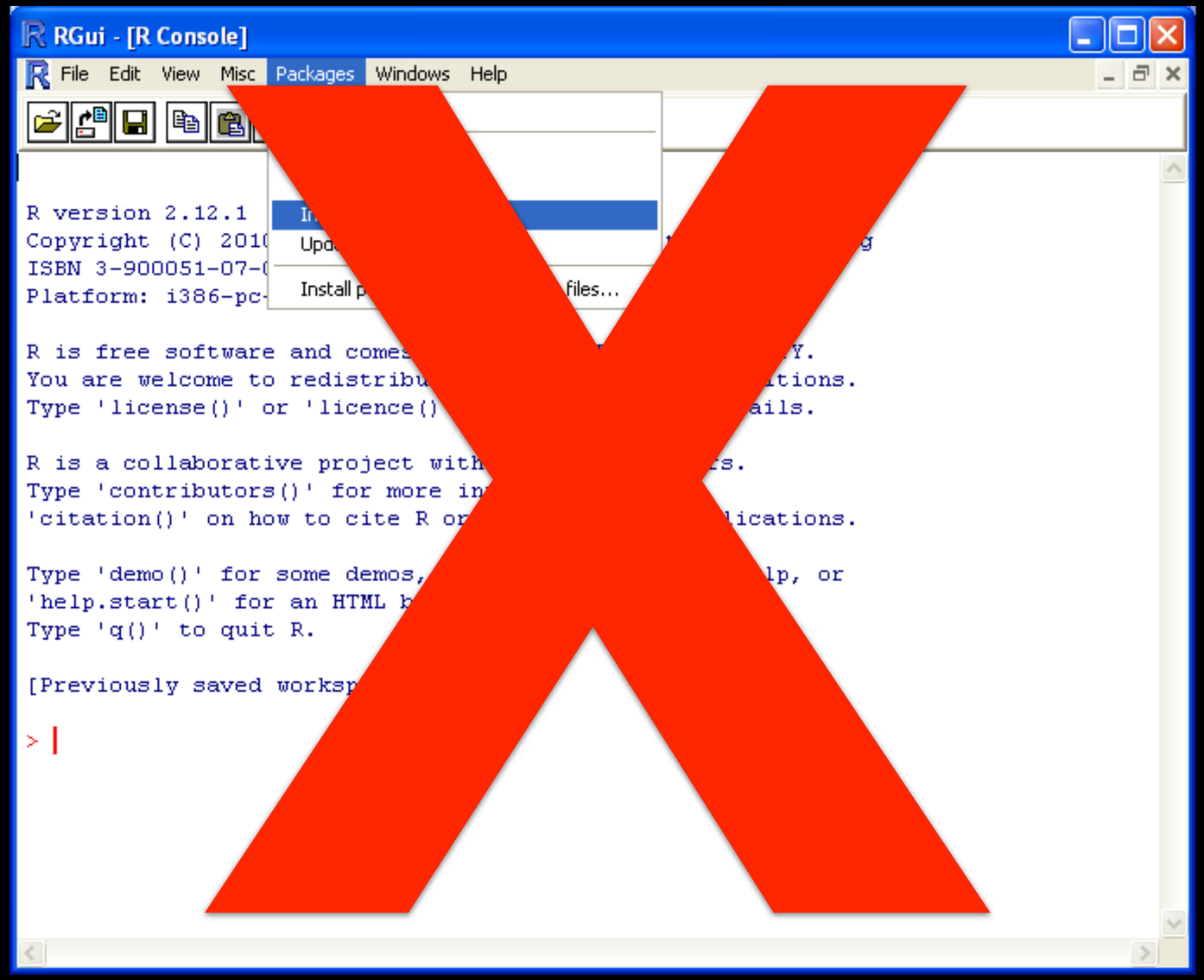




RGui is NOT what we want!



RGui is NOT what we want!



We want: RStudio®

Open RStudio Now!

class04 - RStudio

Console | Terminal | Jobs

~/Desktop/courses/bimm143_S20/class04/

```
R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (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.

  Natural language support but running in an English locale

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.

>
```

Environment | History | Connections | Tutorial

Import Dataset | List

Global Environment

Environment is empty

Files | Plots | Packages | Help | Viewer

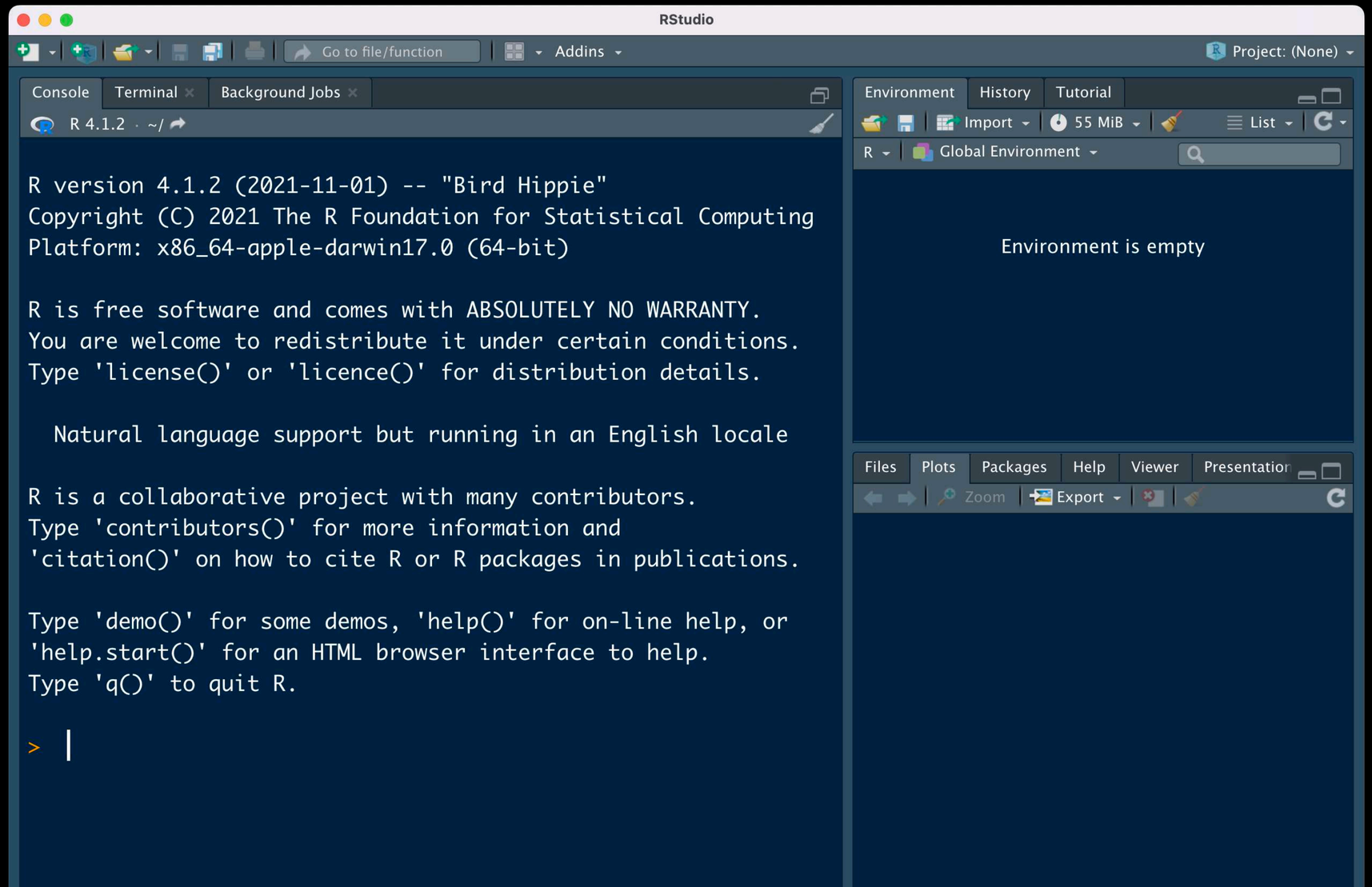
New Folder | Delete | Rename | More

ne > Desktop > courses > bimm143_S20 > class04

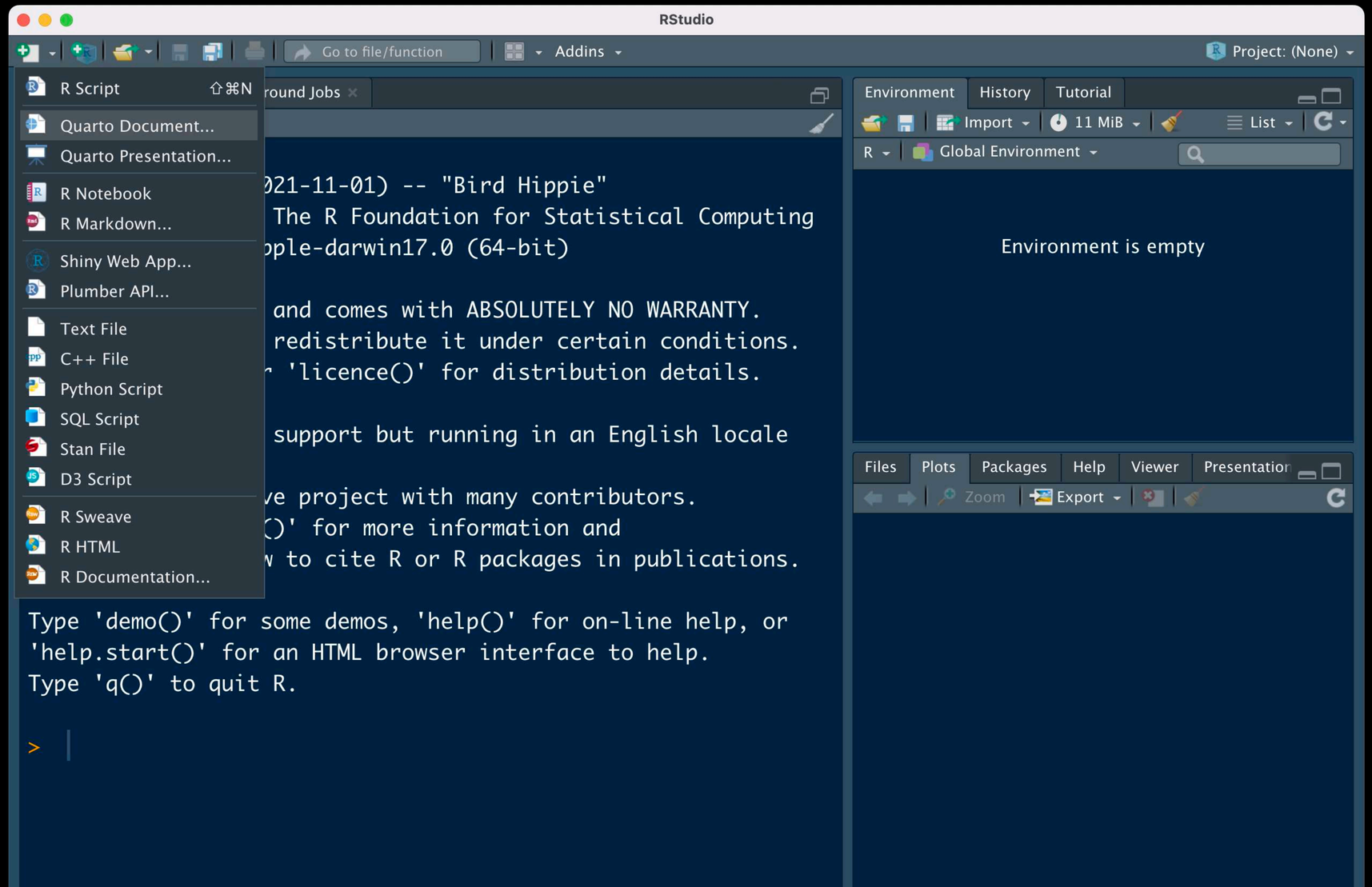
Name	Size
..	
class04.Rproj	205 B

Separate IDE for R!

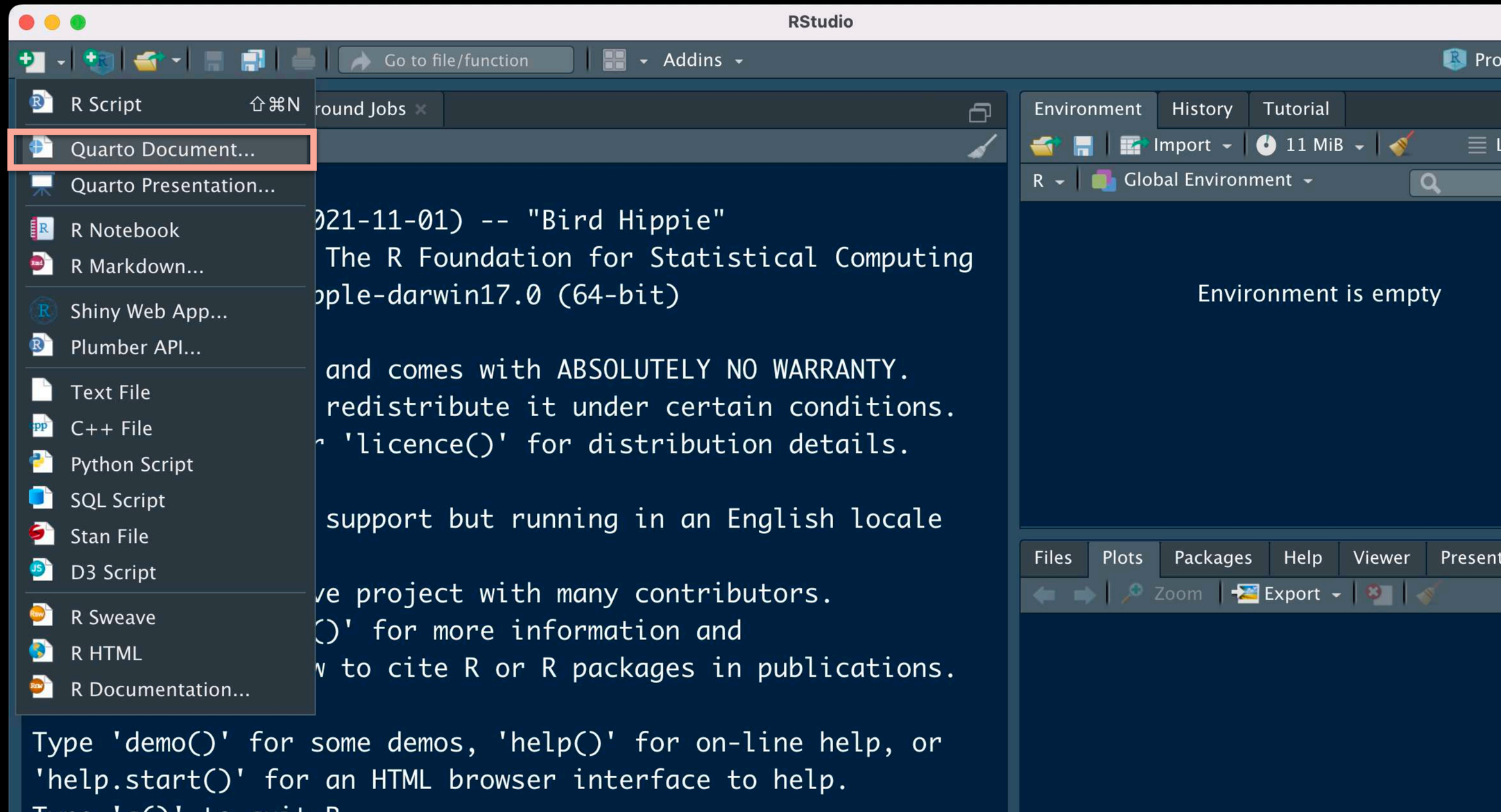
We can customize later…



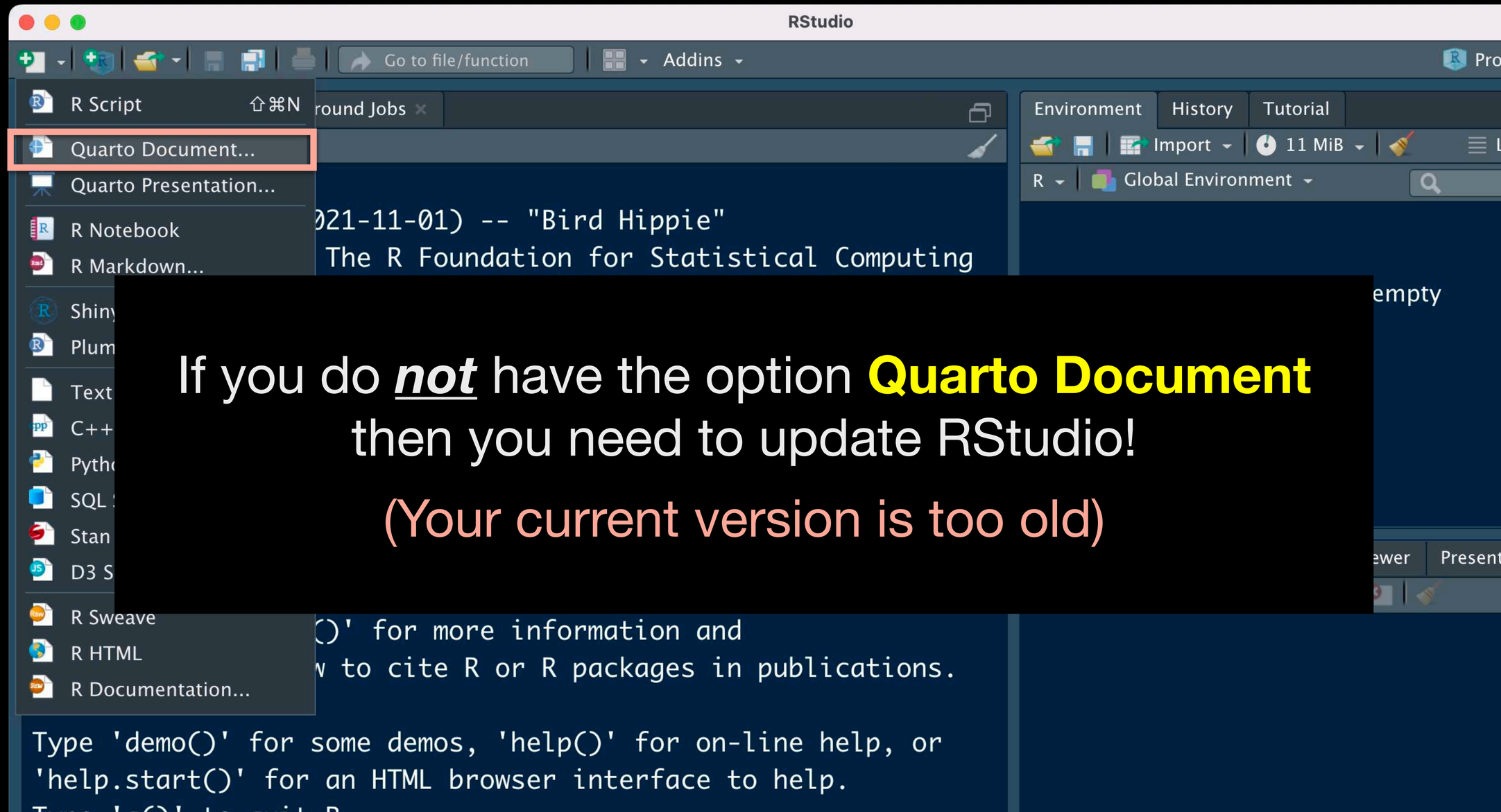
File > New File > Quarto Document



File > New File > Quarto Document



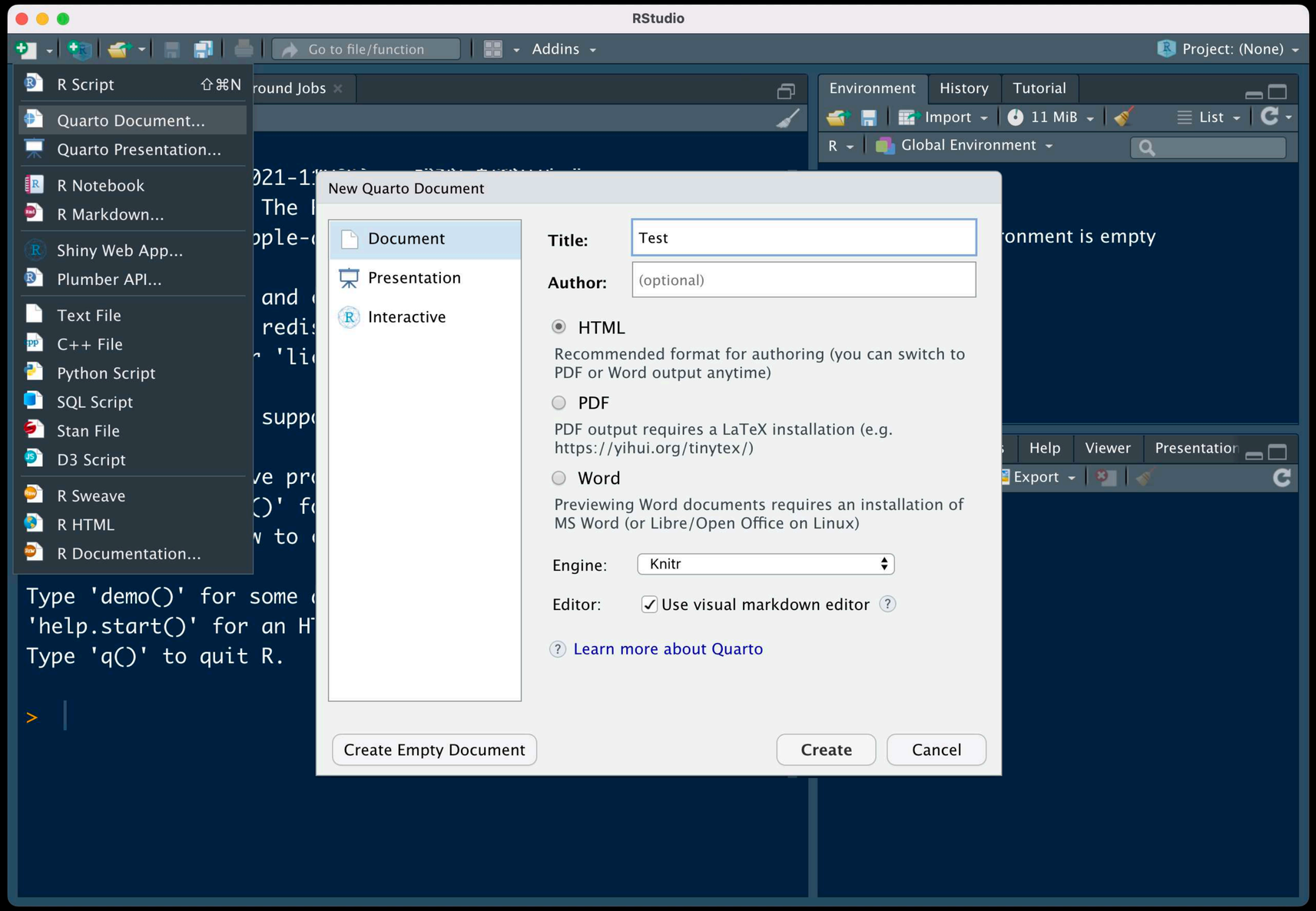
File > New File > Quarto Document



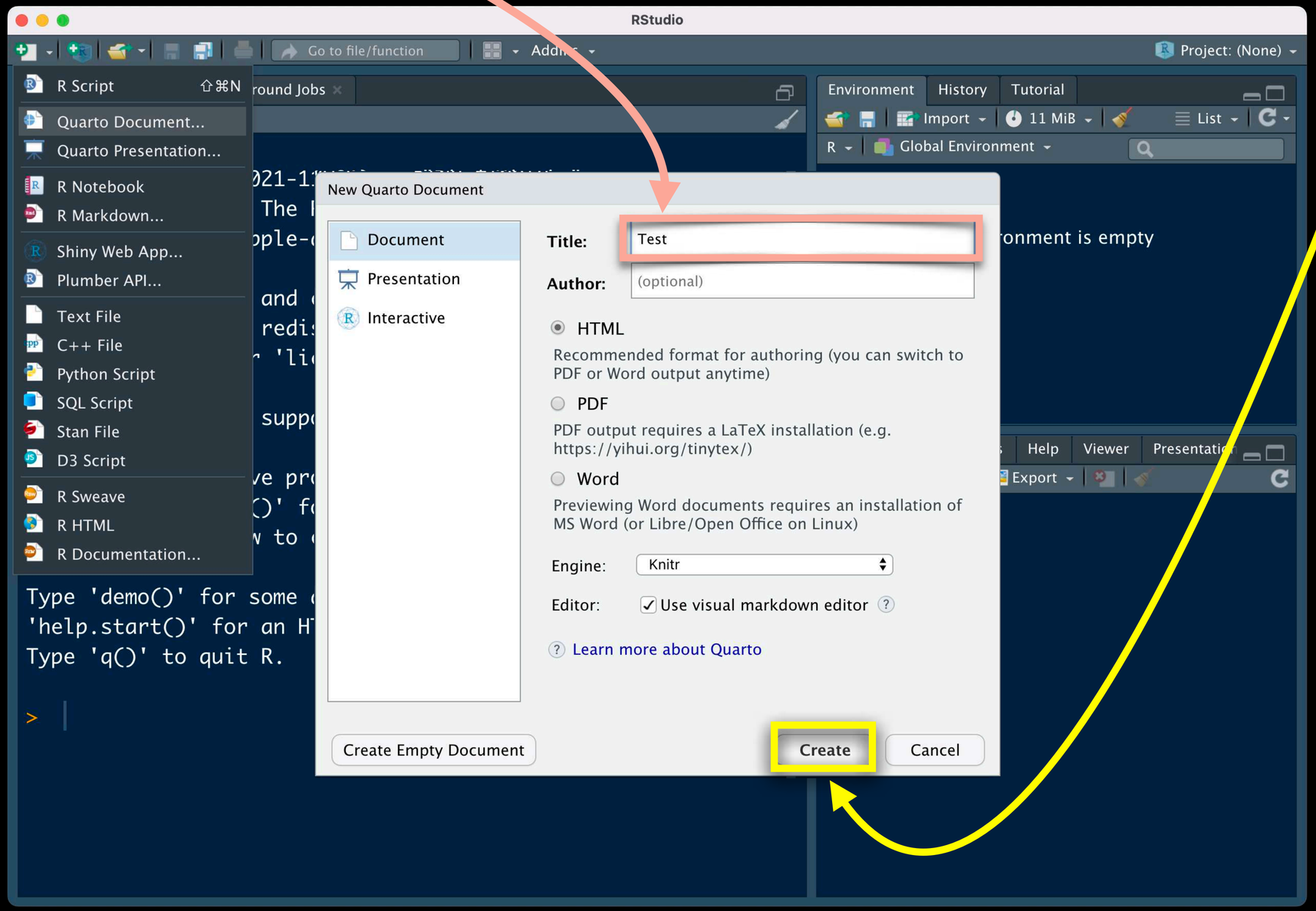
If you do ***not*** have the option **Quarto Document** then you need to update RStudio!

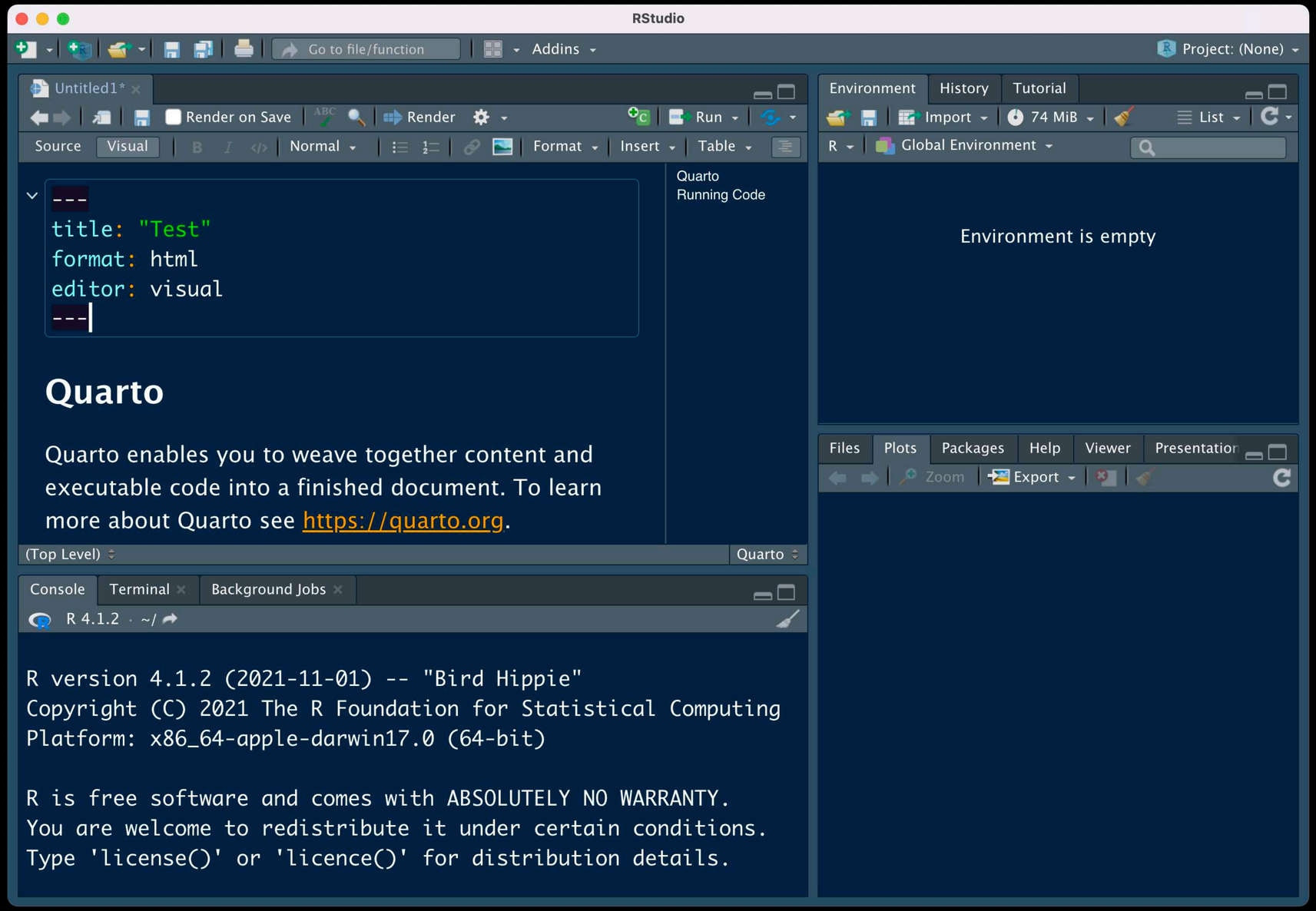
(Your current version is too old)

File > New File > Quarto Document



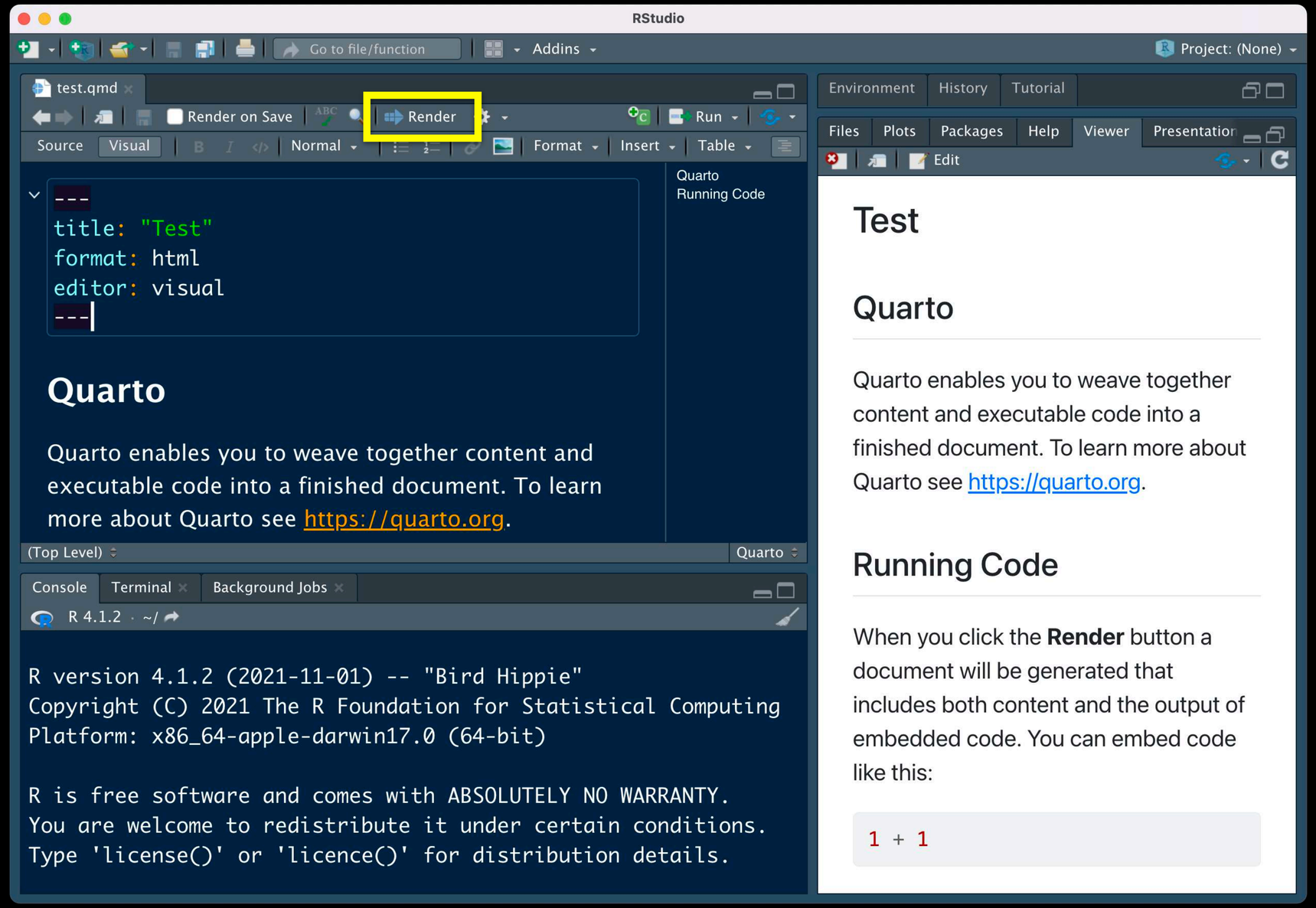
Set Title: Test then click Create




RStudio
Go to file/function
Addins
Project: (None)
Untitled1*
Render on Save
Render
Run
Source
Visual
Normal
Format
Insert
Table
Quarto
Running Code
---
title: "Test"
format: html
editor: visual

Quarto
Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.
(Top Level)
Quarto
Console
Terminal
Background Jobs
R 4.1.2 · ~/
R version 4.1.2 (2021-11-01) -- "Bird Hippie"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (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.
Environment
History
Tutorial
Import
74 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export

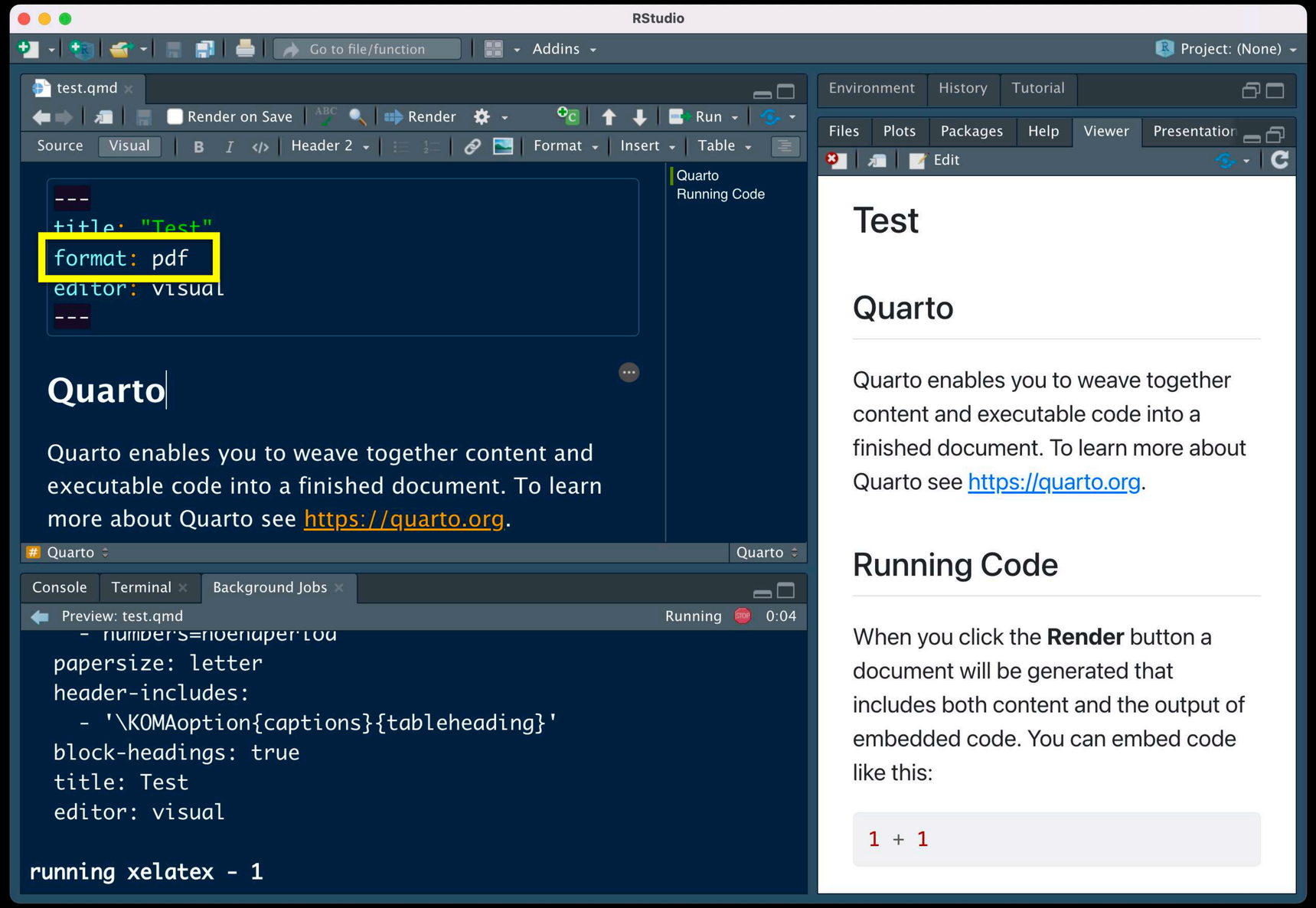
Click “Render”



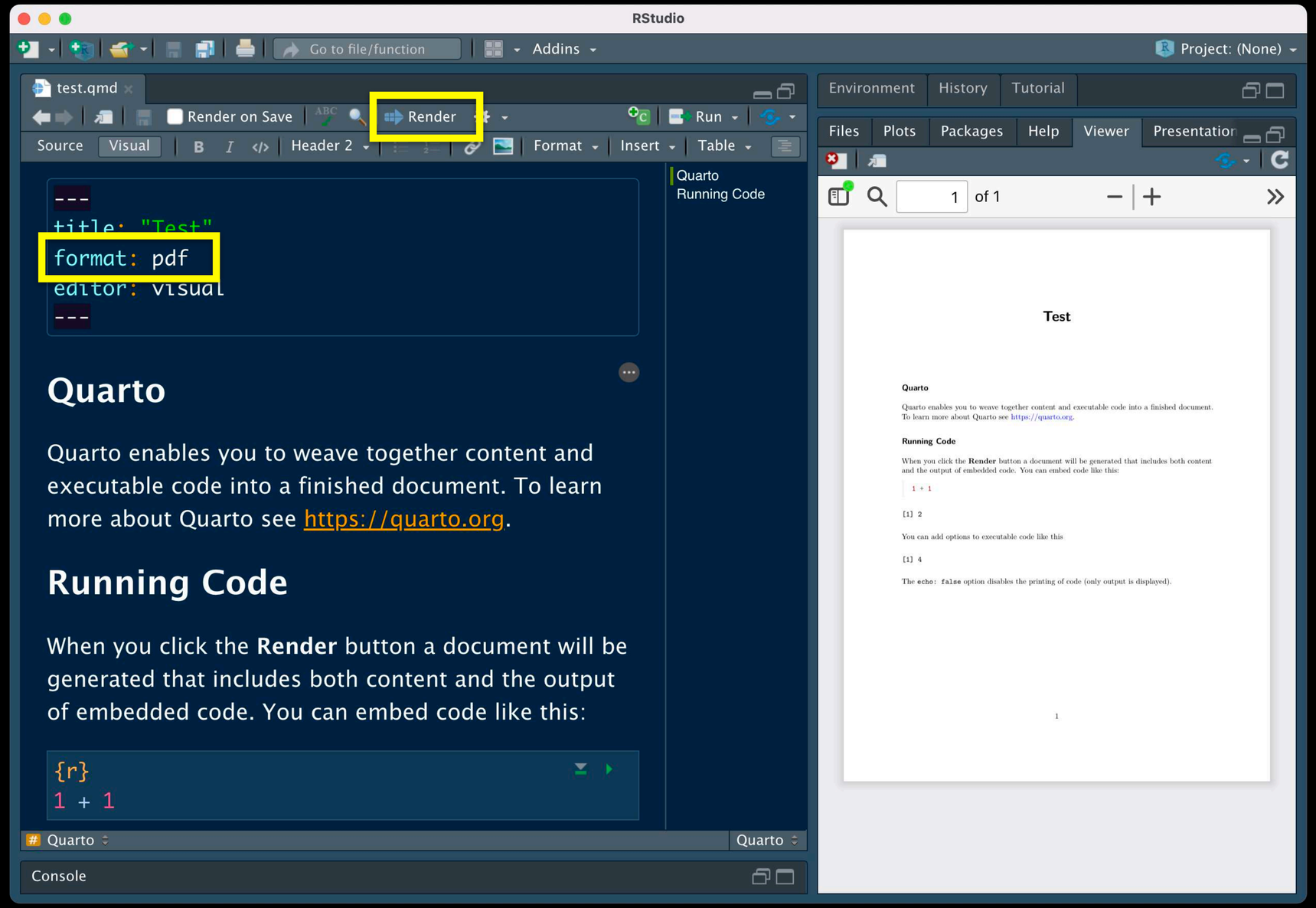
Install missing package

```
install.packages("rmarkdown")
```

Change to format: pdf



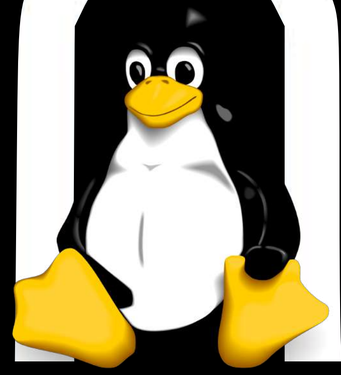
Change to format: pdf



Install missing package

```
install.packages("tinytex")
tinytex::install_tinytex()
```

Essential UNIX

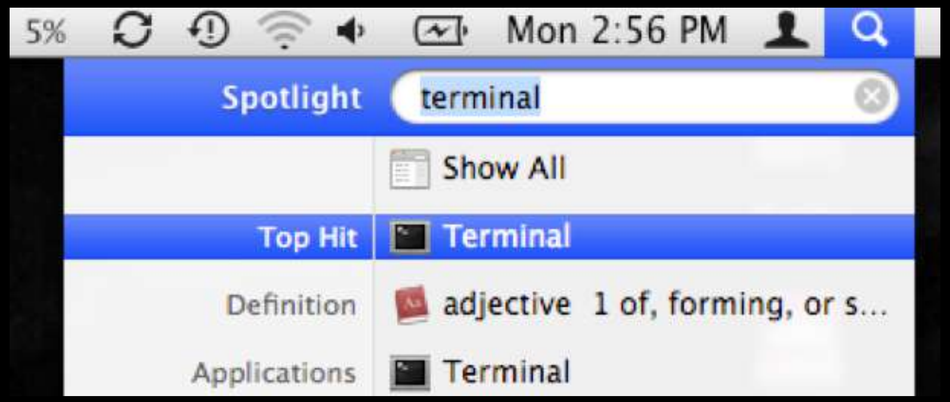


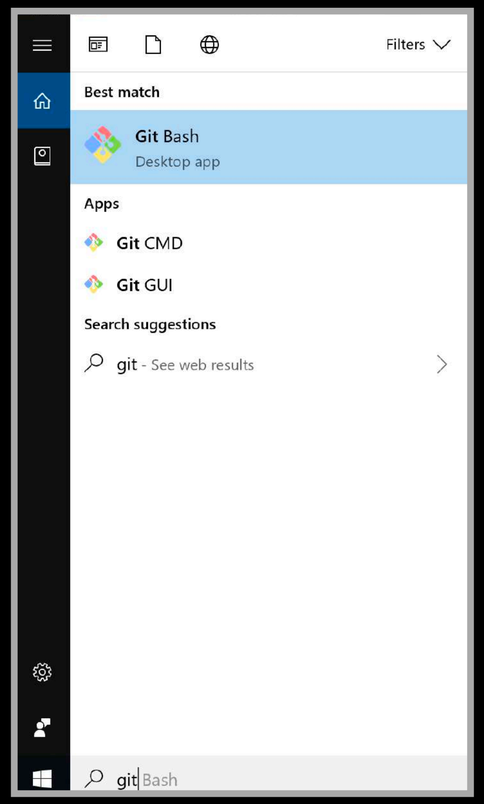
For Bioinformatics

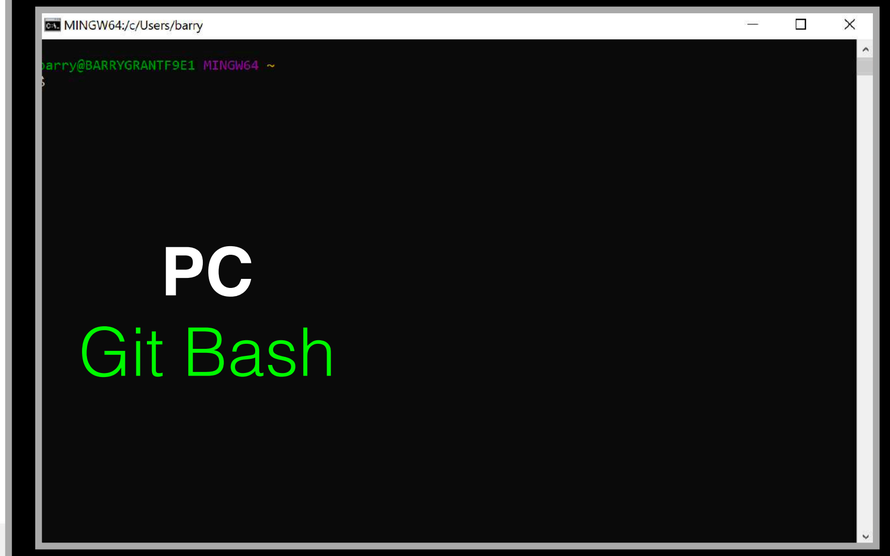
Check if you can use UNIX…

Mac

Terminal



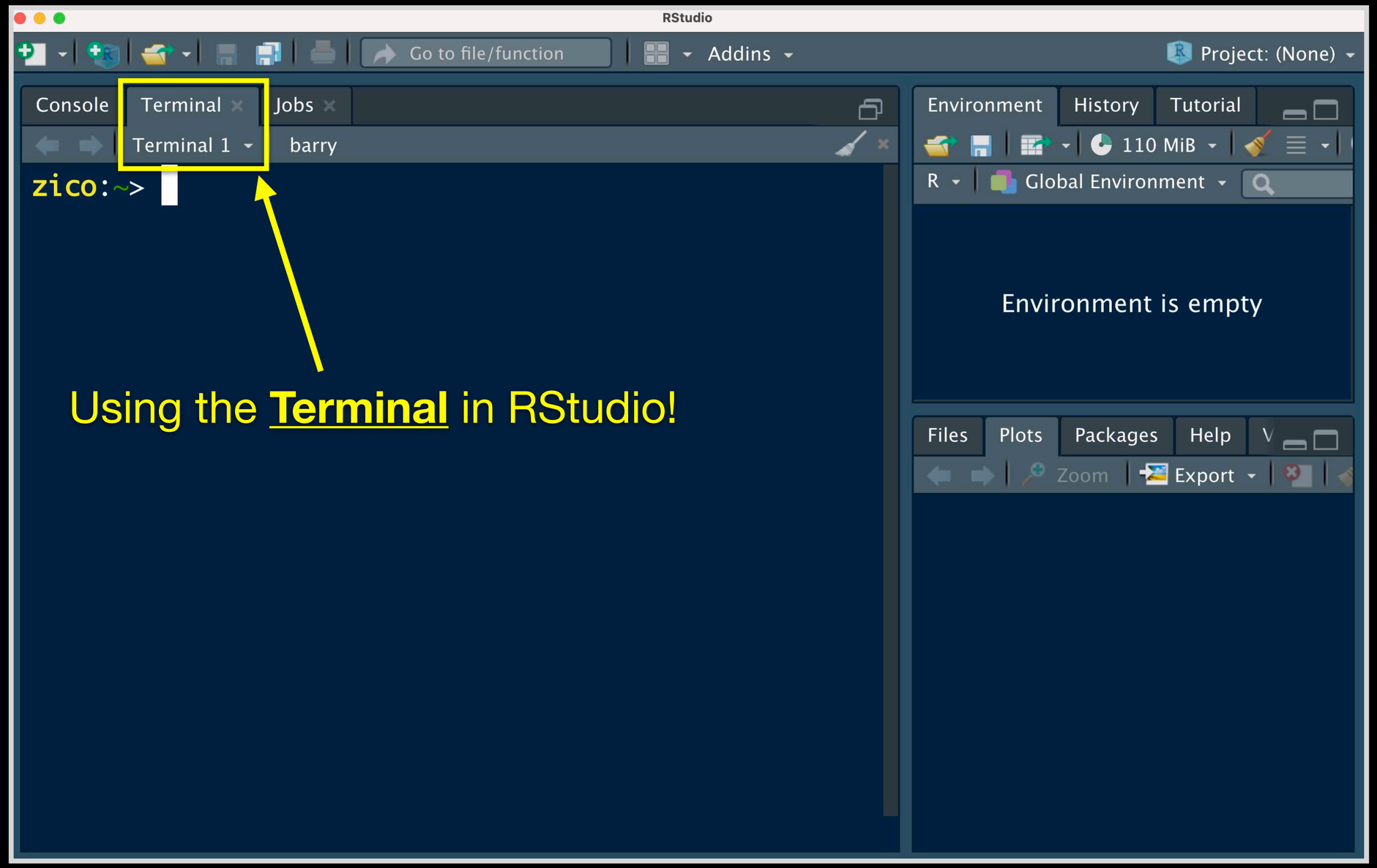




PC

Git Bash

We can also use UNIX in RStudio…



Side-Note:

Being **organized** is key to being successful in this class

Make a new class folder (a.k.a. *Directory*) called **BGGN239** for storing your lab work

Within this folder make sub-folders for each class (e.g. class01)

Review the Computer Setup class page and the optional Recap videos under class00

Review Material

ay-lab.github.io

Home Gmail Gcal GitHub BIMM143 BGGN213 GDrive Atmosphere CloudLaunch BIMM194 Blink News

Bioinformatics Foundations · BGGN 239

BGGN 239

A dedicated course to teach bioinformatics with a specific focus on its applications to important problems in immunology from the Program in Immunology, UCSD

Overview

Schedule

Computer Setup

cycling).

- Be able to install R packages from CRAN and BioConductor.
- Use UNIX command-line tools for file system navigation and text file manipulation.

Supporting material:

- Handout: Class Syllabus,
- Computer Setup Instructions.

Optional Recap Videos from BGGN213:

- 0.1.1 - Introduction to bioinformatics (what, where and why of bioinformatics),
- 0.1.2 - Major bioinformatics resource providers (NCBI and EBI),
- 0.1.3 - A quick tour of the GENE, UniProt, GO, OMIM, PDB and PFAM.
- 0.2.1 - Major R data structures, data types, and using functions,
- 0.2.2 - Introduction to ggplot,
- 0.2.3 - Introduction to CRAN & BioConductor,
- 0.2.4 - Quick introduction to RMarkdown,
- 0.3.1 - Essential UNIX for bioinformatics I,
- 0.3.2 - Essential UNIX for bioinformatics II,
- 0.3.3 - Manipulating files on UNIX machines
- 0.3.4 - UNIX superpowers: using pipes and conecting to remote machines.

Unlimited **DataCamp** Access

Bonus!

app.datacamp.com

Home Gmail Gcal GitHub BIMM143 BGGN213 BGGN239 GDrive Atmosphere CloudLaunch BIMM194 Blink News

Members

datacamp Home Learn Workspace Certification Jobs Groups

Bioinformatics_2023

Invite Members

Dashboard
Members
Teams
Settings

LEARN

Custom Tracks TRY
Assignments
Leaderboard
Insights & Analytics
Reporting
Custom Reports
Skill Matrix

Members / Everyone

Invite Members | Create Invite Link

ALL MEMBERS (1) | PENDING INVITES (18) | WAITING LIST (0) | INVITE LINKS

Search invites

Send Reminders | Withdraw Invites

EMAIL	TEAMS	LEARN	WORKSPACE	ROLE	INVITED BY	INVITED AT
fay@health.ucsd.edu	BGGN239	Classroom	Classroom	MEMBER	Barry Grant	Apr 3, 10:25 PDT
dzangwil@ucsd.edu	BGGN239	Classroom	Classroom	MEMBER	Barry Grant	Apr 3, 10:25 PDT
svandenburgh@ucsd.edu	BGGN239	Classroom	Classroom	MEMBER	Barry Grant	Apr 3, 10:25 PDT
ktakehar@ucsd.edu	BGGN239	Classroom	Classroom	MEMBER	Barry Grant	Apr 3, 10:25 PDT
bstack@ucsd.edu	BGGN239	Classroom	Classroom	MEMBER	Barry Grant	Apr 3, 10:25 PDT

Break!

Class 01

Hands-on Lab Session

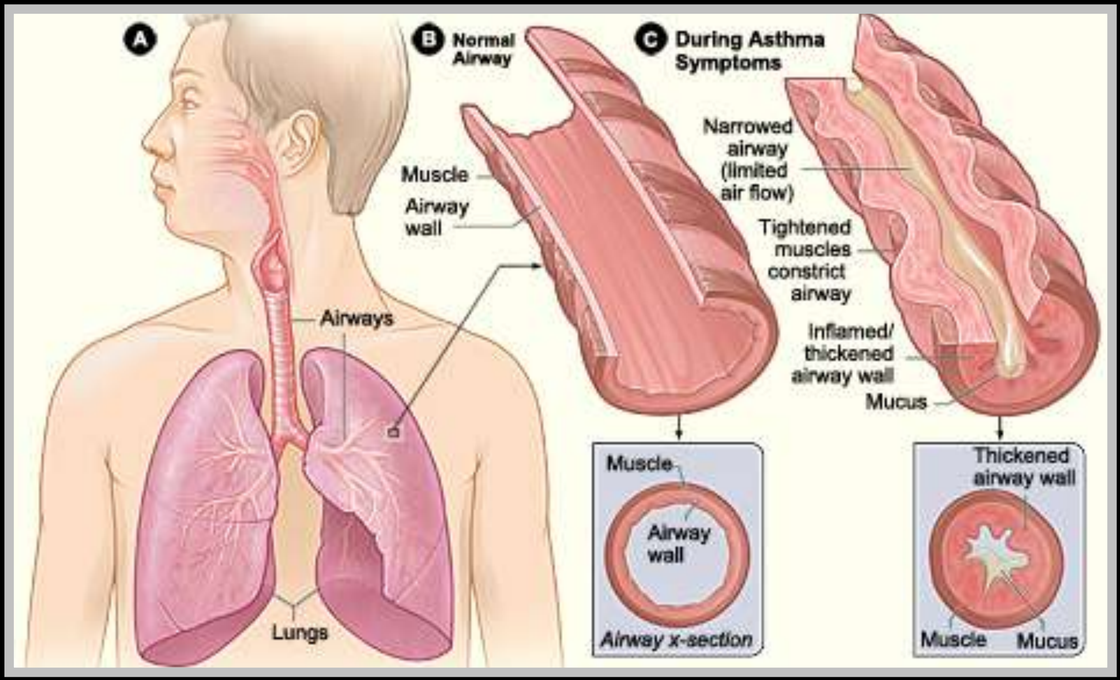
Barry Grant

UC San Diego

http://thegrantlab.org/bggn239/

Background to hands-on data

Glucocorticoids inhibit inflammatory processes and are often used to treat **asthma** because of their anti-inflammatory effects on airway smooth muscle (ASM) cells.



Mechanism?

- Data from: Himes *et al.* "RNA-Seq Transcriptome Profiling Identifies CRISPLD2 as a Glucocorticoid Responsive Gene that Modulates Cytokine Function in Airway Smooth Muscle Cells." PLoS ONE. 2014 Jun 13;9(6):e99625.

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 FEBRUARY 25, 2021 VOL. 384 NO. 8

Dexamethasone in Hospitalized Patients with Covid-19

The RECOVERY Collaborative Group*

ABSTRACT

BACKGROUND

Coronavirus disease 2019 (Covid-19) is associated with diffuse lung damage. Glucocorticoids may modulate inflammation-mediated lung injury and thereby reduce progression to respiratory failure and death.

METHODS

In this controlled, open-label trial comparing a range of possible treatments in patients who were hospitalized with Covid-19, we randomly assigned patients to receive oral or intravenous dexamethasone (at a dose of 6 mg once daily) for up to 10 days or to receive usual care alone. The primary outcome was 28-day mortality. Here, we report the final results of this assessment.

RESULTS

A total of 2104 patients were assigned to receive dexamethasone and 4321 to receive usual care. Overall, 482 patients (22.9%) in the dexamethasone group and 1110 patients (25.7%) in the usual care group died within 28 days after randomization (age-adjusted rate ratio, 0.83; 95% confidence interval [CI], 0.75 to 0.93; P<0.001). The proportional and absolute between-group differences in mortality varied considerably according to the level of respiratory support that the patients were receiving at the time of randomization. In the dexamethasone group, the incidence of death was lower than that in the usual care group among patients receiving invasive mechanical ventilation (29.3% vs. 41.4%; rate ratio, 0.64; 95% CI, 0.51 to 0.81) and among those receiving oxygen without invasive mechanical ventilation (23.3% vs. 26.2%; rate ratio, 0.82; 95% CI, 0.72 to 0.94) but not among those who were receiving no respiratory support at randomization (17.8% vs. 14.0%; rate ratio, 1.19; 95% CI, 0.92 to 1.55).

CONCLUSIONS

In patients hospitalized with Covid-19, the use of dexamethasone resulted in lower 28-day mortality among those who were receiving either invasive mechanical ventilation or oxygen alone at randomization but not among those receiving no respiratory support. (Funded by the Medical Research Council and National Institute for Health Research and others; RECOVERY ClinicalTrials.gov number, NCT04381936; ISRCTN number, 50189673.)

The members of the writing committee (Peter Horby, F.R.C.P., Wei Shen Lim, F.R.C.P., Jonathan R. Emberson, Ph.D., Marion Mafham, M.D., Jennifer L. Bell, M.Sc., Louise Linsell, D.Phil., Natalie Staplin, Ph.D., Christopher Brightling, F.Med.Sci., Andrew Ustianowski, Ph.D., Einas Elmahi, M.Phil., Benjamin Prudon, F.R.C.P., Christopher Green, D.Phil., Timothy Felton, Ph.D., David Chadwick, Ph.D., Kanchan Rege, F.R.C.Path., Christopher Fegan, M.D., Lucy C. Chappell, Ph.D., Saul N. Faust, F.R.C.P.C.H., Thomas Jaki, Ph.D., Katie Jeffery, Ph.D., Alan Montgomery, Ph.D., Kathryn Rowan, Ph.D., Edmund Juszczak, M.Sc., J. Kenneth Baillie, M.D., Ph.D., Richard Haynes, D.M., and Martin J. Landray, F.R.C.P.) assume responsibility for the overall content and integrity of this article.

The affiliations of the members of the writing committee are listed in the Appendix. Address reprint requests to Drs. Horby and Landray at RECOVERY Central Coordinating Office, Richard Doll Bldg., Old Road Campus, Roosevelt Dr., Oxford OX3 7LF, United Kingdom, or at recoverytrial@ndph.ox.ac.uk.

*A complete list of collaborators in the RECOVERY trial is provided in the Supplementary Appendix, available at NEJM.org.

Drs. Horby, Lim, and Emberson and Drs. Haynes and Landray contributed equally to this article.

A preliminary version of this article was published on July 17, 2020, at NEJM.org.

N Engl J Med 2021;384:693-704.
DOI: 10.1056/NEJMoa2021436

For COVID-19 patients on ventilators, dexamethasone treatment was shown to reduce mortality by about one third

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 FEBRUARY 25, 2021 VOL. 384 NO. 8

Dexamethasone in Hos

The RECOV

BACKGROUND

Coronavirus disease 2019 (Covid-19) is associated w
corticoids may modulate inflammation-mediated
progression to respiratory failure and death.

METHODS

In this controlled, open-label trial comparing a
in patients who were hospitalized with Covid-19, w
receive oral or intravenous dexamethasone (at a d
to 10 days or to receive usual care alone. The primar
Here, we report the final results of this assessmer

RESULTS

A total of 2104 patients were assigned to receive
ceive usual care. Overall, 482 patients (22.9%) in
1110 patients (25.7%) in the usual care group died
tion (age-adjusted rate ratio, 0.83; 95% confider
P<0.001). The proportional and absolute between
varied considerably according to the level of respi
were receiving at the time of randomization. In
incidence of death was lower than that in the us
receiving invasive mechanical ventilation (29.3% v
CI, 0.51 to 0.81) and among those receiving oxyg
ventilation (23.3% vs. 26.2%; rate ratio, 0.82; 95%
those who were receiving no respiratory support at
rate ratio, 1.19; 95% CI, 0.92 to 1.55).

CONCLUSIONS

In patients hospitalized with Covid-19, the use
lower 28-day mortality among those who were rece
ventilation or oxygen alone at randomization but
respiratory support. (Funded by the Medical Rese
stitute for Health Research and others; RECOVI
NCT04381936; ISRCTN number, 50189673.)

Corticosteroids for COVID-19: the search for an optimum duration of therapy

Michael A Matthay and B Taylor Thompson[1] have very nicely summarised the evidence-based role of dexamethasone in hospitalised patients with COVID-19. Their pertinent analysis is based on the background of the RECOVERY trial,[2] which concluded that therapy with dexamethasone at a dose of 6 mg once daily for up to 10 days decreased 28-day mortality in patients with COVID-19 on respiratory support. Patients not requiring oxygen showed no benefit but had a possibility of harm with corticosteroid therapy.[2]

One crucial feature of corticosteroid therapy is its duration, particularly in patients with COVID-19 with sustained persistence of ground-glass opacities. Currently, an extended course of corticosteroids beyond 10 days is considered only in select cases of thrombi and microthrombi were seen.[4] Dexamethasone (6 mg per day) tends to increase clotting factor and fibrinogen concentrations. Thus, it is plausible for exogenous glucocorticoids to precipitate clinical thrombosis.[5] In addition, protracted corticosteroid therapy might contribute to the so-called long COVID syndrome that manifests with fatigue and psychological symptoms, in which steroid-related adverse drug reactions such as myopathy, neuromuscular weakness, and psychiatric symptoms might have a part to play.[6,7]

Late in the disease course, corticosteroids do not appear to have a role in managing acute respiratory distress syndrome (ARDS). Routine use of methylprednisolone for persistent ARDS is not recommended despite improving cardiopulmonary physiology. Even initiating methylprednisolone therapy more than 2 weeks after the onset of ARDS might increase the risk of death.[7]

A meta-analysis of 21 350 patients such an extended course of steroids could be detrimental.

We declare no competing interests.

*Gyanshankar P Mishra, Jasmin Mulani
gpmishra81@gmail.com

Department of Respiratory Medicine, Indira Gandhi Government Medical College, Nagpur, Maharashtra 440018, India (GPM) and Department of Biochemistry, Government Medical College, Nagpur, Maharashtra, India (JM)

1 Matthay MA, Thompson BT. Dexamethasone in hospitalised patients with COVID-19: addressing uncertainties. *Lancet Respir Med* 2020; **8:** 1170–72.

2 The RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with Covid-19—preliminary report. *N Engl J Med* 2020; published online Jul 17. https://www.nejm.org/doi/full/10.1056/NEJMoa2021436.

3 Villar J, Confalonieri M, Pastores SM, Meduri GU. Rationale for prolonged corticosteroid treatment in the acute respiratory distress syndrome caused by coronavirus disease 2019. *Crit Care Explor* 2020; **2:** e0111.

4 Maiese A, Manetti AC, La Russa R, et al. Autopsy findings in COVID-19-related deaths: a literature review. *Forensic Sci Med Pathol* 2020; published online Oct 7. https://doi.org/10.1007/s12024-020-00310-8.

5 Brotman DJ, Girod JP, Posch A, et al. Effects of short-term glucocorticoids on hemostatic factors in healthy volunteers. *Thromb Res* 2006; **118:** 247–52.

6 Warrington TP, Bostwick JM. Psychiatric

Published Online
November 26, 2020
https://doi.org/10.1016/S2213-2600(20)30530-0

Background to hands-on data

- Himes *et al.* used **RNA-seq** to profile gene expression changes in 4 ASM cell lines treated with **dexamethasone** (a common synthetic glucocorticoid).

Background to hands-on data

- Himes *et al.* used **RNA-seq** to profile gene expression changes in 4 ASM cell lines treated with **dexamethasone** (a common synthetic glucocorticoid).

- Used Tophat and Cufflinks to quantify transcript abundances

Background to hands-on data

- Himes *et al.* used **RNA-seq** to profile gene expression changes in 4 ASM cell lines treated with **dexamethasone** (a common synthetic glucocorticoid).
- Used Tophat and Cufflinks to quantify transcript abundances
- They found many differentially expressed genes and focused on CRISPLD2 that encodes a secreted protein involved in lung development

Background to hands-on data

- Himes *et al.* used **RNA-seq** to profile gene expression changes in 4 ASM cell lines treated with **dexamethasone** (a common synthetic glucocorticoid).
- Used Tophat and Cufflinks to quantify transcript abundances
- They found many differentially expressed genes and focused on CRISPLD2 that encodes a secreted protein involved in lung development
- SNPs in CRISPLD2 in previous GWAS associated with inhaled corticosteroid resistance and bronchodilator response in asthma patients.

Background to hands-on data

- Himes *et al.* used **RNA-seq** to profile gene expression changes in 4 ASM cell lines treated with **dexamethasone** (a common synthetic glucocorticoid).
- Used Tophat and Cufflinks to quantify transcript abundances
- They found many differentially expressed genes and focused on CRISPLD2 that encodes a secreted protein involved in lung development
- SNPs in CRISPLD2 in previous GWAS associated with inhaled corticosteroid resistance and bronchodilator response in asthma patients.
- Confirmed the upregulated CRISPLD2 with qPCR and increased protein expression with Western blotting.

Background to hands-on data

- Himes *et al.* used **RNA-seq** to profile gene expression changes in 4 ASM cell lines treated with **dexamethasone** (a common synthetic glucocorticoid).

- Used Tophat and Cufflinks to quantify transcript abundances

- They found many differentially expressed genes and focused on CRISPLD2 that encodes a secreted protein involved in lung development

- SNPs in CRISPLD2 in previous GWAS associated with inhaled corticosteroid resistance and bronchodilator response in asthma patients.

- Confirmed the upregulated CRISPLD2 with qPCR and increased protein expression with Western blotting.

Background to hands-on data

- Himes *et al.* used **RNA-seq** to profile gene expression changes in 4 ASM cell lines treated with **dexamethasone** (a common synthetic glucocorticoid).
- Used Tophat and Cufflinks to quantify transcript abundances

- Our starting point is a count matrix: each cell indicates the number of reads originating from a particular gene (in rows) for each sample (in columns).

counts

countData

gene	ctrl_1	ctrl_2	exp_1	exp_2
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
...	...	...	...	...

countData is the count matrix
(Number of reads coming from each gene for each sample)

counts + metadata

1 **countData**

gene	ctrl_1	ctrl_2	exp_1	exp_2
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
...	...	...	...	...

countData is the count matrix
(Number of reads coming from each gene for each sample)

2 **colData**

counts + metadata

1 **countData**

gene	ctrl_1	ctrl_2	exp_1	exp_2
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
...	...	...	...	...

countData is the count matrix
(Number of reads coming from each gene for each sample)

2 **colData**

id	treatment	sex	...
ctrl_1	control	male	...
ctrl_2	control	female	...
exp_1	treated	male	...
exp_2	treated	female	...

colData describes metadata about the *columns* of countData

counts + metadata

1 **countData**

gene	ctrl_1	ctrl_2	exp_1	exp_2
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
...	...	...	...	...

countData is the count matrix (Number of reads coming from each gene for each sample)

2 **colData**

id	treatment	sex	...
ctrl_1	control	male	...
ctrl_2	control	female	...
exp_1	treated	male	...
exp_2	treated	female	...

colData describes metadata about the *columns* of countData

N.B. First column of **colData** must match column names (i.e. sample names) of **countData**

Do it Yourself!

Install DESeq2

Bioconductor Setup Link

```
install.packages("BiocManager")
BiocManager::install()

# For this class, you'll also need DESeq2:
BiocManager::install("DESeq2")
```

Note: Answer **NO** to prompts to install from source or update...

Do it Yourself!

Do this in your CONSOLE not an Qmd document!

Install DESeq2

Bioconductor Setup Link

```
install.packages("BiocManager")
BiocManager::install()

# For this class, you'll also need DESeq2:
BiocManager::install("DESeq2")
```

Note: Answer **NO** to prompts to install from source or update...

```
Old packages: 'devtools', 'dplyr', 'DT', 'ggplot2', 'ggpubr',
  'lattice', 'MASS', 'Matrix', 'mclust', 'mgcv', 'openssl',
  'packrat', 'pkgload', 'ps', 'psych', 'raster', 'rcmdcheck',
  'Rcpp', 'remotes', 'rsconnect', 'sessioninfo', 'shiny',
  'shinythemes', 'survival', 'tidyr', 'tinytex', 'xfun'
Update all/some/none? [a/s/n]:
n
```

Do it Yourself!

Install DESeq2

Bioconductor Setup Link

```r
install.packages("BiocManager")
BiocManager::install()

# For this class, you'll also need DESeq2:
BiocManager::install("DESeq2")
```

Note: Answer **NO** to prompts to install from source or update...

```r
metaFile <- "data/GSE37704_metadata.csv"
countFile <- "data/GSE37704_featurecounts.csv"
```

Input file names

```r
colData = read.csv(metaFile, row.names=1)
countData = read.csv(countFile, row.names=1)
```

Read files

```r
dds = DESeqDataSetFromMatrix(countData=countData,
                             colData=colData,
                             design=~condition)
dds = DESeq(dds)
```

Setup required DESeq object

```r
res <- results(dds)
res
```

Run the DESeq pipeline

X	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
ENSG00000152583	954.77093	4.3683590	0.23713648	18.421286	8.867079e-76	1.342919e-71	SPARCL1
ENSG00000179094	743.25269	2.8638885	0.17555825	16.313039	7.972621e-60	6.037267e-56	PER1
ENSG00000116584	2277.91345	-1.0347000	0.06505273	-15.905557	5.798513e-57	2.927283e-53	ARHGEF2
ENSG00000189221	2383.75371	3.3415441	0.21241508	15.731200	9.244206e-56	3.500088e-52	MAOA
ENSG00000120129	3440.70375	2.9652108	0.20370277	14.556557	5.306416e-48	1.607313e-44	DUSP1
ENSG00000148175	13493.92037	1.4271683	0.10036663	14.219550	6.929711e-46	1.749175e-42	STOM
ENSG00000178695	2685.40974	-2.4890689	0.17806407	-13.978501	2.108817e-44	4.562576e-41	KCTD12
ENSG00000109906	439.54152	5.9275950	0.42819442	13.843233	1.397758e-43	2.646131e-40	ZBTB16
ENSG00000134686	2933.64246	1.4394898	0.10582729	13.602255	3.882769e-42	6.533838e-39	PHC2
ENSG00000101347	14134.99177	3.8504143	0.28490701	13.514635	1.281894e-41	1.941428e-38	SAMHD1
ENSG00000096060	2630.23049	3.9450524	0.29291821	13.468102	2.409807e-41	3.317866e-38	FKBP5
ENSG00000166741	7542.25287	2.2195906	0.16673544	13.312050	1.970000e-40	2.486304e-37	NNMT
ENSG00000125148	3695.87946	2.1985636	0.16700546	13.164621	1.402400e-39	1.633797e-36	MT2A
ENSG00000162614	5646.18314	1.9711402	0.15020631	13.122885	2.434854e-39	2.633990e-36	NEXN
ENSG00000106976	989.04683	-1.8501713	0.14778657	-12.519211	5.861471e-36	5.918132e-33	DNM1
ENSG00000187193	199.07694	3.2551424	0.26090711	12.476250	1.006146e-35	9.523804e-33	MT1X
ENSG00000256235	1123.47954	1.2801193	0.10547438	12.136779	6.742862e-34	6.007096e-31	SMIM3
ENSG00000177666	2639.57020	1.1399947	0.09606884	11.866436	1.768422e-32	1.487930e-29	PNPLA2
ENSG00000164125	7257.00808	1.0248523	0.08657600	11.837603	2.494830e-32	1.988642e-29	FAM198B
ENSG00000198624	2020.04495	2.8141014	0.24063429	11.694515	1.359615e-31	1.029569e-28	CCDC69
ENSG00000123562	5008.55294	1.0045453	0.08901501	11.285123	1.554241e-29	1.120904e-26	MORF4L2
ENSG00000144369	1283.77980	-1.3090041	0.11714863	-11.173875	5.473974e-29	3.768333e-26	FAM171B
ENSG00000196517	241.91536	-2.3456877	0.21047366	-11.144804	7.591120e-29	4.998588e-26	SLC6A9
ENSG00000135821	19973.40000	3.0413943	0.27601796	11.018828	3.100706e-28	1.956675e-25	GLUL

X	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
ENSG00000152583	954.77093	4.3683590	0.23713648	18.421286	8.867079e-76	1.342919e-71	SPARCL1
ENSG00000179	[illegible]	2.8638885	0.17555825	16.313039	7.972621e-60	6.037267e-56	PER1
ENSG00000116	[illegible]	-1.0347000	0.06505273	-15.905557	5.798513e-57	2.927283e-53	ARHGEF2
ENSG00000189	[illegible]	3.3415441	0.21241508	15.731200	9.244206e-56	3.500088e-52	MAOA
ENSG00000120129	3440.70375	2.9652108	0.20370277	14.556557	5.306416e-48	1.607313e-44	DUSP1
ENSG00000148175	13493.92037	[illegible]	0.10036663	14.219550	6.929711e-46	1.749175e-42	STOM
ENSG00000178695	2685.40974	[illegible]	0.17806407	-13.978501	2.108817e-44	4.562576e-41	KCTD12
ENSG00000109906	439.54152	5.9275950	0.42819442	13.843233	1.397758e-43	2.646131e-40	ZBTB16
ENSG00000134686	2933.64246	1.4394898	[illegible]	13.602255	3.882769e-42	6.533838e-39	PHC2
ENSG00000101347	14134.99177	3.8504143	[illegible]	13.514635	1.281894e-41	1.941428e-38	SAMHD1
ENSG00000096060	2630.23049	3.9450524	0.29291821	13.468102	2.409807e-41	3.317866e-38	FKBP5
ENSG00000166741	7542.25287	2.2195906	0.16673544	[illegible]	.970000e-40	2.486304e-37	NNMT
ENSG00000125148	3695.87946	2.1985636	0.16700546	[illegible]	.402400e-39	1.633797e-36	MT2A
ENSG00000162614	5646.18314	1.9711402	0.15020631	[illegible]	.434854e-39	2.633990e-36	NEXN
ENSG00000106976	989.04683	-1.8501713	0.14778657	-12.519211	[illegible]	5.918132e-33	DNM1
ENSG00000187193	199.07694	3.2551424	0.26090711	12.476250	[illegible]	9.523804e-33	MT1X
ENSG00000256235	1123.47954	1.2801193	0.10547438	12.136779	[illegible]	6.007096e-31	SMIM3
ENSG00000177666	2639.57020	1.1399947	0.09606884	11.866436	1.768422e-[illegible]	[illegible]	[illegible]PLA2
ENSG00000164125	7257.00808	1.0248523	0.08657600	11.837603	2.494830e-[illegible]	[illegible]	[illegible]M198B
ENSG00000198624	2020.04495	2.8141014	0.24063429	11.694515	1.359615e-[illegible]	[illegible]	[illegible]DC69
ENSG00000123562	5008.55294	1.0045453	0.08901501	11.285123	1.554241e-29	1.120904e-26	MORF4L2
ENSG00000144369	1283.77980	-1.3090041	0.11714863	-11.173875	5.473974e-29	3.768333e-26	FAM171B
ENSG00000196517	241.91536	-2.3456877	0.21047366	-11.144804	7.591120e-29	4.998588e-26	SLC6A9
ENSG00000135821	19973.40000	3.0413943	0.27601796	11.018828	3.100706e-28	1.956675e-25	GLUL

mean counts from all samples

log2 fold change

standard error

Wald statistic

Wald p-value

BH adjusted p-values

X	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
ENSG00000152583	954.77093	4.3683590	0.23713648	18.421286	8.867079e-76	1.342919e-71	SPARCL1
ENSG00000179094	[illegible]	2.8638885	0.17555825	16.313039	7.972621e-60	6.037267e-56	PER1
ENSG00000116[illegible]	[illegible]	-1.0347000	0.06505273	-15.905557	5.798513e-57	2.927283e-53	ARHGEF2
ENSG00000189[illegible]	[illegible]	3.3415441	0.21241508	15.731200	9.244206e-56	3.500088e-52	MAOA
ENSG00000120129	3440.70375	2.9652108	0.20370277	14.556557	5.306416e-48	1.607313e-44	DUSP1
ENSG00000148175	13493.92037	[illegible]	0.10036663	14.219550	6.929711e-46	1.749175e-42	STOM
ENSG00000178695	2685.40974	[illegible]	0.17806407	-13.978501	2.108817e-44	4.562576e-41	KCTD12
ENSG00000109906	439.54152	5.9275950	0.42819442	13.843233	1.397758e-43	2.646131e-40	ZBTB16
ENSG00000134686	2933.64246	1.4394898	[illegible]	13.602255	3.882769e-42	6.533838e-39	PHC2
ENSG00000101347	14134.99177	3.8504143	[illegible]	13.514635	1.281894e-41	1.941428e-38	SAMHD1
ENSG00000096060	2630.23049	3.9450524	0.29291821	13.468102	2.409807e-41	3.317866e-38	FKBP5
ENSG00000166741	7542.25287	2.2195906	0.16673544	[illegible]	[illegible].970000e-40	2.486304e-37	NNMT
ENSG00000125148	3695.87946	2.1985636	0.16700546	[illegible]	[illegible].402400e-39	1.633797e-36	MT2A
ENSG00000162614	5646.18314	1.9711402	0.15020631	[illegible]	[illegible].434854e-39	2.633990e-36	NEXN
ENSG00000106976	989.04683	-1.8501713	0.14778657	-12.519211	[illegible]	5.918132e-33	DNM1
ENSG00000187193	199.07694	3.2551424	0.26090711	12.476250	[illegible]	9.523804e-33	MT1X
ENSG00000256235	1123.47954	1.2801193	0.10547438	12.136779	[illegible]	6.007096e-31	SMIM3
ENSG00000177666	2639.57020	1.1399947	0.09606884	11.866436	1.768422e-[illegible]	[illegible]	[illegible]PLA2
ENSG00000164125	7257.00808	1.0248523	0.08657600	11.837603	2.494830e-[illegible]	[illegible]	[illegible]M198B
ENSG00000198624	2020.04495	2.8141014	0.24063429	11.694515	1.359615e-[illegible]	[illegible]	[illegible]DC69
ENSG00000123562	5008.55294	1.0045453	0.08901501	11.285123	1.554241e-29	1.120904e-26	MORF4L2
ENSG00000144369	1283.77980	-1.3090041	0.11714863	-11.173875	5.473974e-29	3.768333e-26	FAM171B
ENSG00000196517	241.91536	-2.3456877	0.21047366	-11.144804	7.591120e-29	4.998588e-26	SLC6A9
ENSG00000135821	19973.40000	3.0413943	0.27601796	11.018828	3.100706e-28	1.956675e-25	GLUL

mean counts from all samples

log2 fold change

standard error

Wald statistic

Wald p-value

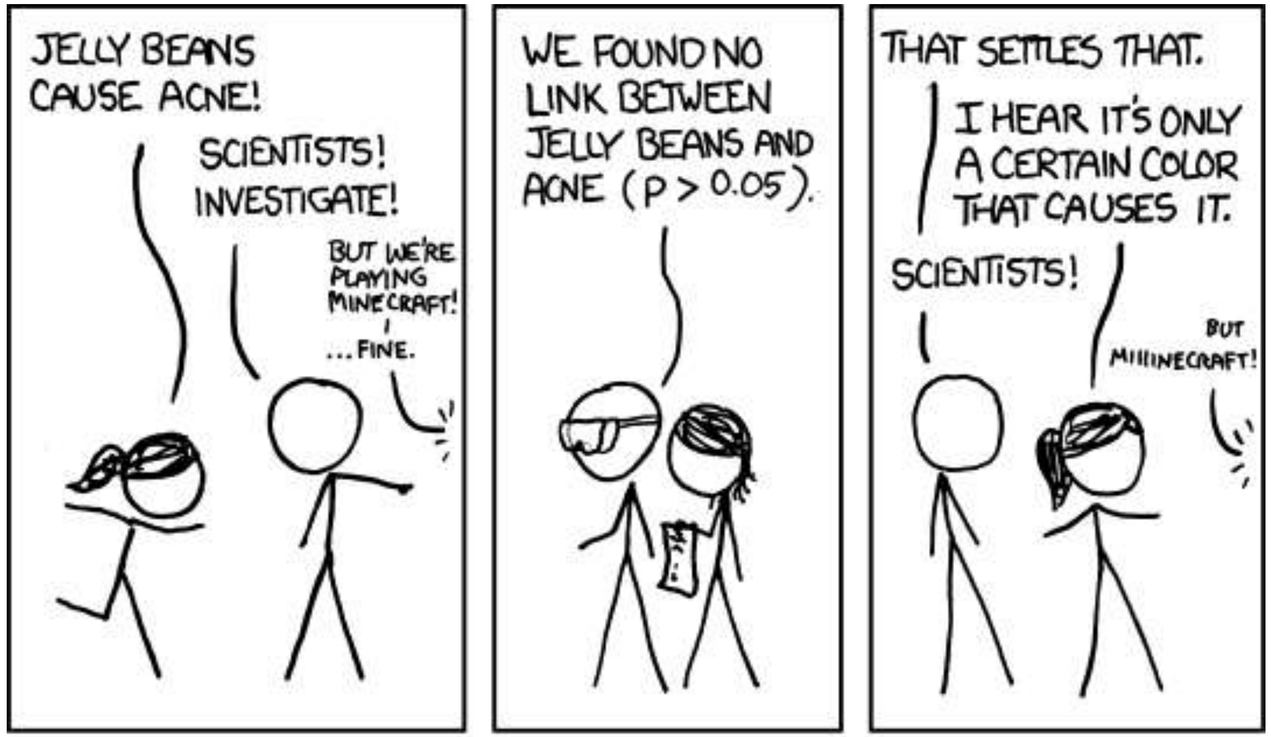
BH adjusted p-values

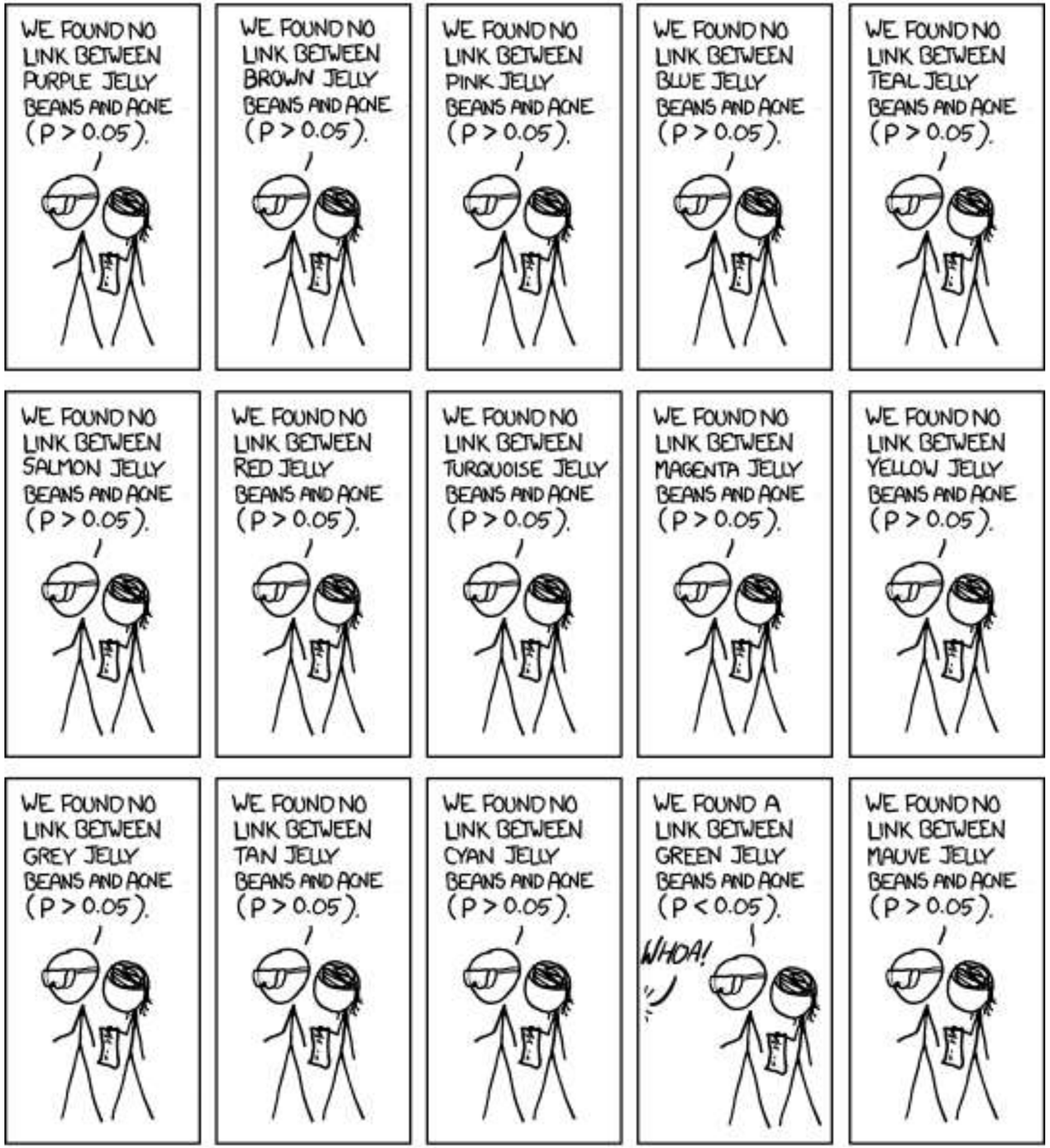
We need to add gene names (a.k.a. gene symbols) and other database identifiers

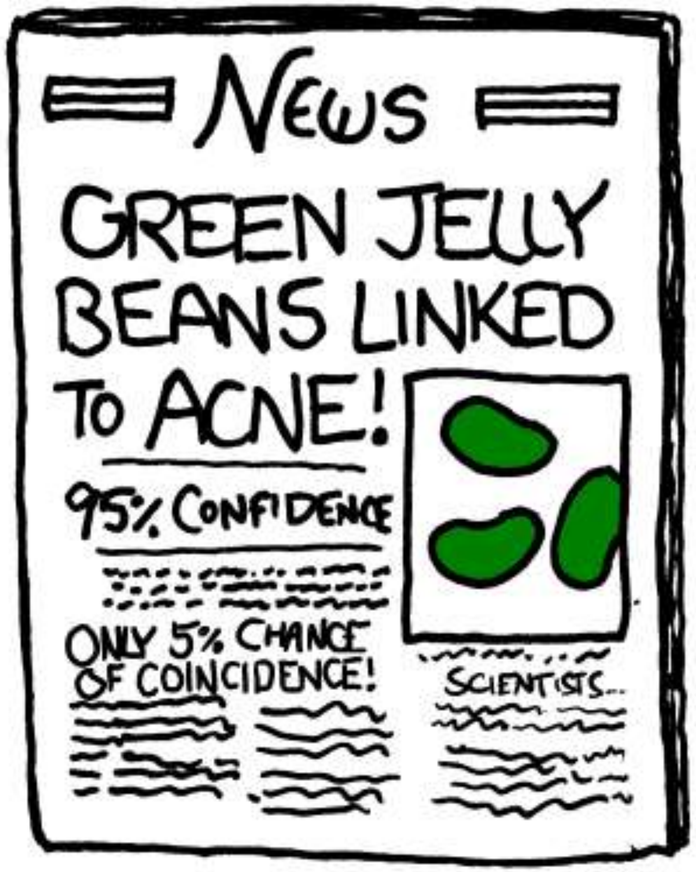
X	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
ENSG00000152583	954.77093	4.3683590	0.23713648	18.421286	8.867079e-76	1.342919e-71	SPARCL1
ENSG00000179094	743.25269	2.8638885	0.17555825	16.313039	7.972621e-60	6.037267e-56	PER1
ENSG00000116584	2277.91345	-1.0347000	0.06505273	-15.905557	5.798513e-57	2.927283e-53	ARHGEF2
ENSG00000189221	2383.75371	3.3415441	0.21241508	15.731200	9.244206e-56	3.500088e-52	MAOA
ENSG00000120129	3440.70375	2.9652108	0.20370277	14.556557	5.306416e-48	1.607313e-44	DUSP1
ENSG00000148175	13493.92037	1.4271683	0.10036663	14.219550	6.929711e-46	1.749175e-42	STOM
ENSG00000178695	2685.40974	-2.4890689	0.17806407	-13.978501	2.108817e-44	4.562576e-41	KCTD12
ENSG00000109906	439.54152	5.9275950	0.42819442	13.843233	1.397758e-43	2.646131e-40	ZBTB16
ENSG00000134686	2933.64246	1.4394898	0.10582729	13.602255	3.882769e-42	6.533838e-39	PHC2
ENSG00000101347	14134.99177	3.8504143	0.28490701	13.514635	1.281894e-41	1.941428e-38	SAMHD1
ENSG00000256235	1123.47954	1.2801193	0.10547438	12.136779	6.742862e-34	6.007096e-31	SMIM3
ENSG00000177666	2639.57020	1.1399947	0.09606884	11.866436	1.768422e-32	1.487930e-29	PNPLA2
ENSG00000164125	7257.00808	1.0248523	0.08657600	11.837603	2.494830e-32	1.988642e-29	FAM198B
ENSG00000198624	2020.04495	2.8141014	0.24063429	11.694515	1.359615e-31	1.029569e-28	CCDC69
ENSG00000123562	5008.55294	1.0045453	0.08901501	11.285123	1.554241e-29	1.120904e-26	MORF4L2
ENSG00000144369	1283.77980	-1.3090041	0.11714863	-11.173875	5.473974e-29	3.768333e-26	FAM171B
ENSG00000196517	241.91536	-2.3456877	0.21047366	-11.144804	7.591120e-29	4.998588e-26	SLC6A9
ENSG00000135821	19973.40000	3.0413943	0.27601796	11.018828	3.100706e-28	1.956675e-25	GLUL

Genomics = Lots of Data = Lots of Hypothesis Tests

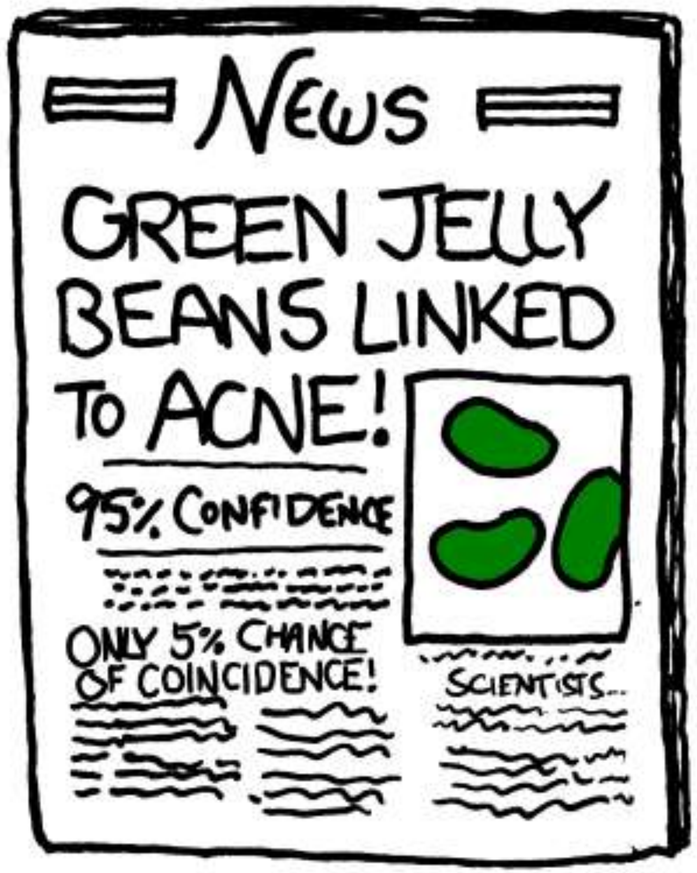
20,000 separate hypothesis tests with a standard p-value cut-off of 0.05, we'd expect 1,000 genes to be deemed "significant" by chance!


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Key Point: Torture the data long enough, and it will confess



padj: Adjustment of p-values for doing multiple tests

- ***"Torture the data long enough, and it will confess"***
  - With each *question* you are increasing the chance of being fooled by chance (20,000 tests @ alpha=0.05; 20,000 x 0.05 = 1,000).
  - You increase your *type 1 errors* mistakenly concluding that an effect is statistically significant.

- **In DESeq2, the p-values are corrected for multiple testing using the *Benjamini and Hochberg* method:**
  - First, rank the genes by p-value. Then multiply each p-value by (total number of tests)/rank.
  - Alternative *Bonferroni method*: p-value x (total number of tests)

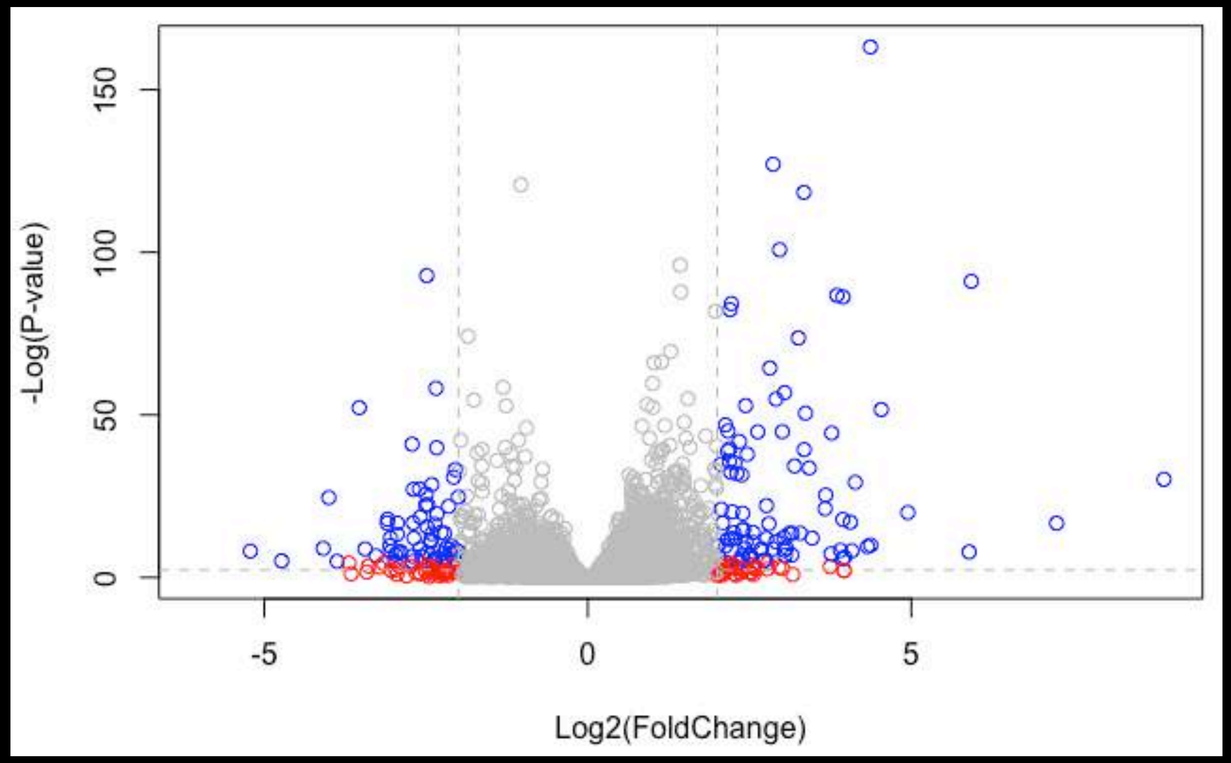
Fold change (log ratios)

- **To a statistician fold change is sometimes considered meaningless**.
  - Fold change can be large (e.g. >>two-fold up- or down-regulation) without being statistically significant (e.g. based on **p-values**).

- **To a biologist fold change is almost always considered important** for two main reasons.
  - First, a very small but statistically significant fold change might not be relevant to a cell's function.
  - Second, it is of interest to know which genes are most dramatically regulated, as these are often thought to reflect changes in biologically meaningful transcripts and/or pathways.

Volcano plot

A common summary figure used to highlight genes that are both significantly regulated and display a high fold change

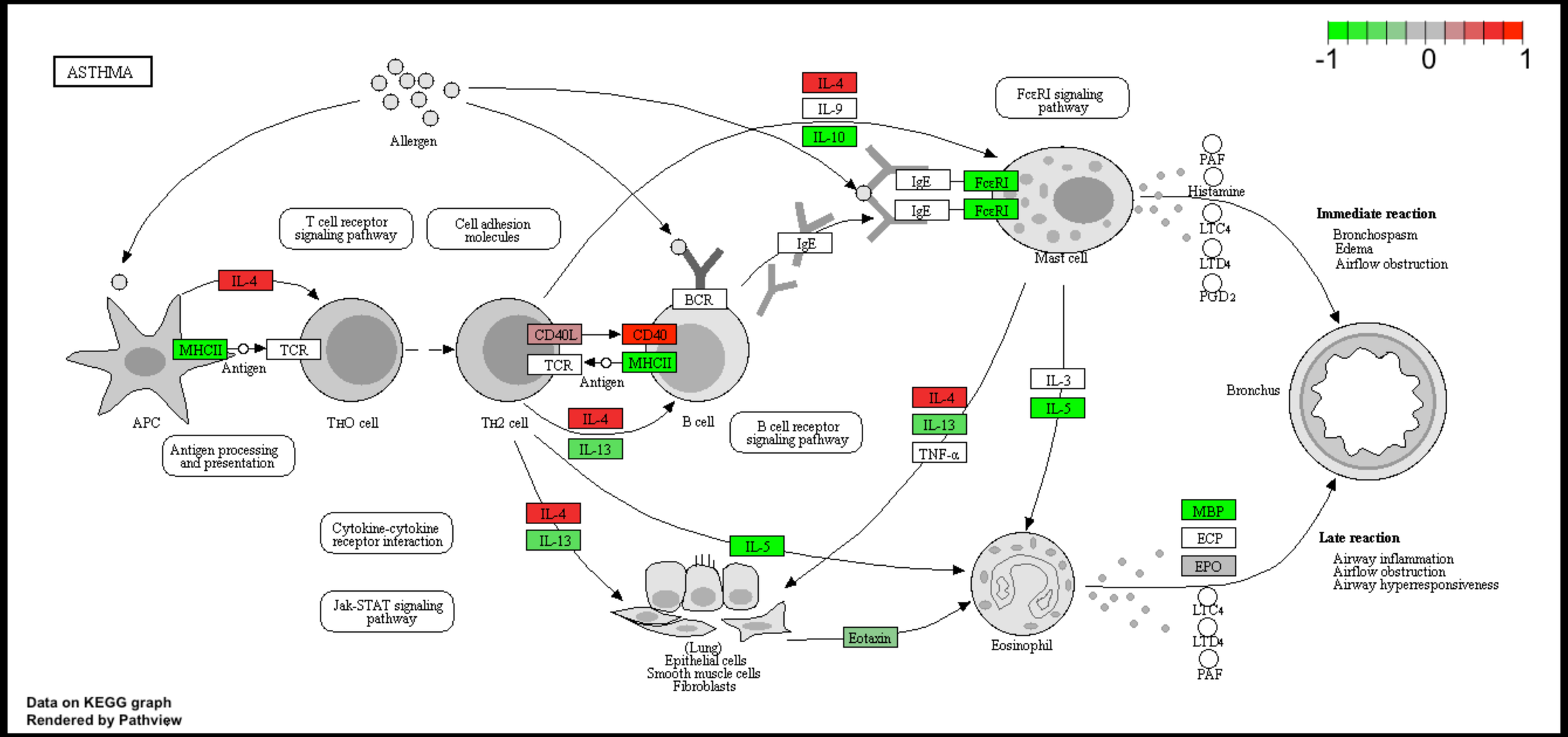


A volcano plot shows fold change (x-axis) versus -log of the *p-value* (y-axis) for a given transcript. The more significant the *p-value*, the larger the -log of that value will be. Therefore we often focus on 'higher up' points.

OPTIONAL: Next steps

Annotation and gene set enrichment
(a.k.a. pathway analysis)

Pathway Analysis



N.B. Render your lab report to PDF and upload to GradeScope

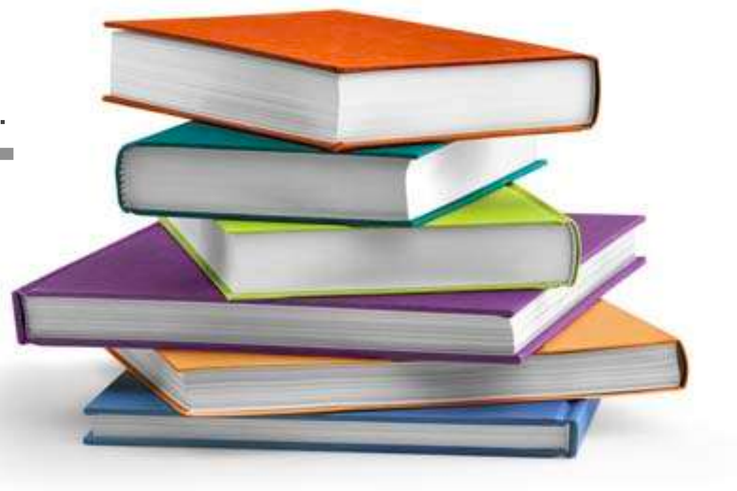
Basic idea: **Pathway analysis**

Differentially Expressed Genes (**DEGs**)

X	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
ENSG00000152583	954.77093	4.3683590	0.23713648	18.421286	8.867079e-76	1.342919e-71	SPARCL1
ENSG00000179094	743.25269	2.8638885	0.17555825	16.313039	7.972621e-60	6.037267e-56	PER1
ENSG00000116584	2277.91345	-1.0347000	0.06505273	-15.905557	5.798513e-57	2.927283e-53	ARHGEF2
ENSG00000189221	2383.75371	3.3415441	0.21241508	15.731200	9.244206e-56	3.500088e-52	MAOA
ENSG00000120129	3440.70375	2.9652108	0.20370277	14.556557	5.306416e-48	1.607313e-44	DUSP1
ENSG00000148175	13493.92037	1.4271683	0.10036663	14.219550	6.929711e-46	1.749175e-42	STOM
ENSG00000178695	2685.40974	-2.4890689	0.17806407	-13.978501	2.108817e-44	4.562576e-41	KCTD12
ENSG00000109906	439.54152	5.9275950	0.42819442	13.843233	1.397758e-43	2.646131e-40	ZBTB16
ENSG00000134686	2933.64246	1.4394898	0.10582729	13.602255	3.882769e-42	6.533838e-39	PHC2
ENSG00000101347	14134.99177	3.8504143	0.28490701	13.514635	1.281894e-41	1.941428e-38	SAMHD1
ENSG00000096060	2630.23049	3.9450524	0.29291821	13.468102	2.409807e-41	3.317866e-38	FKBP5
ENSG00000166741	7542.25287	2.2195906	0.16673544	13.312050	1.970000e-40	2.486304e-37	NNMT
ENSG00000125148	3695.87946	2.1985636	0.16700546	13.164621	1.402400e-39	1.633797e-36	MT2A
ENSG00000162614	5646.18314	1.9711402	0.15020631	13.122885	2.434854e-39	2.633990e-36	NEXN
ENSG00000106976	989.04683	-1.8501713	0.14778657	-12.519211	5.861471e-36	5.918132e-33	DNM1
ENSG00000187193	199.07694	3.2551424	0.26090711	12.476250	1.006146e-35	9.523804e-33	MT1X
ENSG00000256235	1123.47954	1.2801193	0.10547438	12.136779	6.742862e-34	6.007096e-31	SMIM3
ENSG00000177666	2639.57020	1.1399947	0.09606884	11.866436	1.768422e-32	1.487930e-29	PNPLA2
ENSG00000164125	7257.00808	1.0248523	0.08657600	11.837603	2.494830e-32	1.988642e-29	FAM198B
ENSG00000198624	2020.04495	2.8141014	0.24063429	11.694515	1.359615e-31	1.029569e-28	CCDC69
ENSG00000123562	5008.55294	1.0045453	0.08901501	11.285123	1.554241e-29	1.120904e-26	MORF4L2
ENSG00000144369	1283.77980	-1.3090041	0.11714863	-11.173875	5.473974e-29	3.768333e-26	FAM171B
ENSG00000196517	241.91536	-2.3456877	0.21047366	-11.144804	7.591120e-29	4.998588e-26	SLC6A9
ENSG00000135821	19973.40000	3.0413943	0.27601796	11.018828	3.100706e-28	1.956675e-25	GLUL

Annotate...

Gene-sets (Pathways, annotations, etc...)



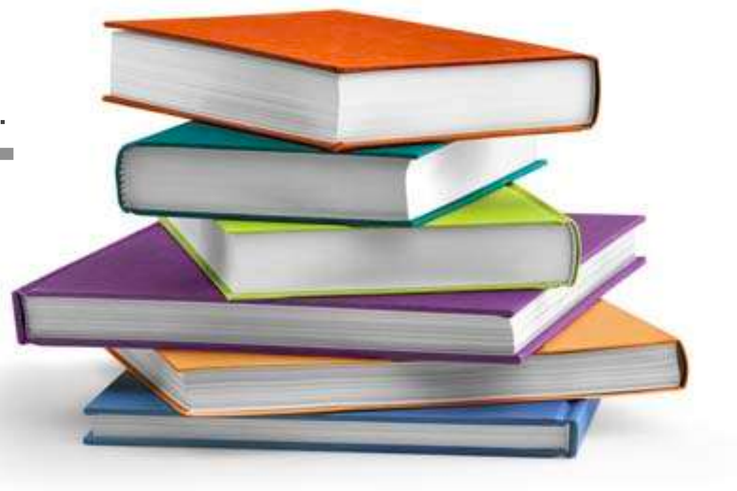
Basic idea: **Pathway analysis**

Differentially Expressed Genes (**DEGs**)

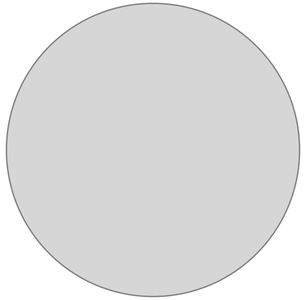
X	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
ENSG00000152583	954.77093	4.3683590	0.23713648	18.421286	8.867079e-76	1.342919e-71	SPARCL1
ENSG00000179094	743.25269	2.8638885	0.17555825	16.313039	7.972621e-60	6.037267e-56	PER1
ENSG00000116584	2277.91345	-1.0347000	0.06505273	-15.905557	5.798513e-57	2.927283e-53	ARHGEF2
ENSG00000189221	2383.75371	3.3415441	0.21241508	15.731200	9.244206e-56	3.500088e-52	MAOA
ENSG00000120129	3440.70375	2.9652108	0.20370277	14.556557	5.306416e-48	1.607313e-44	DUSP1
ENSG00000148175	13493.92037	1.4271683	0.10036663	14.219550	6.929711e-46	1.749175e-42	STOM
ENSG00000178695	2685.40974	-2.4890689	0.17806407	-13.978501	2.108817e-44	4.562576e-41	KCTD12
ENSG00000109906	439.54152	5.9275950	0.42819442	13.843233	1.397758e-43	2.646131e-40	ZBTB16
ENSG00000134686	2933.64246	1.4394898	0.10582729	13.602255	3.882769e-42	6.533838e-39	PHC2
ENSG00000101347	14134.99177	3.8504143	0.28490701	13.514635	1.281894e-41	1.941428e-38	SAMHD1
ENSG00000096060	2630.23049	3.9450524	0.29291821	13.468102	2.409807e-41	3.317866e-38	FKBP5
ENSG00000166741	7542.25287	2.2195906	0.16673544	13.312050	1.970000e-40	2.486304e-37	NNMT
ENSG00000125148	3695.87946	2.1985636	0.16700546	13.164621	1.402400e-39	1.633797e-36	MT2A
ENSG00000162614	5646.18314	1.9711402	0.15020631	13.122885	2.434854e-39	2.633990e-36	NEXN
ENSG00000106976	989.04683	-1.8501713	0.14778657	-12.519211	5.861471e-36	5.918132e-33	DNM1
ENSG00000187193	199.07694	3.2551424	0.26090711	12.476250	1.006146e-35	9.523804e-33	MT1X
ENSG00000256235	1123.47954	1.2801193	0.10547438	12.136779	6.742862e-34	6.007096e-31	SMIM3
ENSG00000177666	2639.57020	1.1399947	0.09606884	11.866436	1.768422e-32	1.487930e-29	PNPLA2
ENSG00000164125	7257.00808	1.0248523	0.08657600	11.837603	2.494830e-32	1.988642e-29	FAM198B
ENSG00000198624	2020.04495	2.8141014	0.24063429	11.694515	1.359615e-31	1.029569e-28	CCDC69
ENSG00000123562	5008.55294	1.0045453	0.08901501	11.285123	1.554241e-29	1.120904e-26	MORF4L2
ENSG00000144369	1283.77980	-1.3090041	0.11714863	-11.173875	5.473974e-29	3.768333e-26	FAM171B
ENSG00000196517	241.91536	-2.3456877	0.21047366	-11.144804	7.591120e-29	4.998588e-26	SLC6A9
ENSG00000135821	19973.40000	3.0413943	0.27601796	11.018828	3.100706e-28	1.956675e-25	GLUL

Annotate...

Gene-sets (Pathways, annotations, etc...)



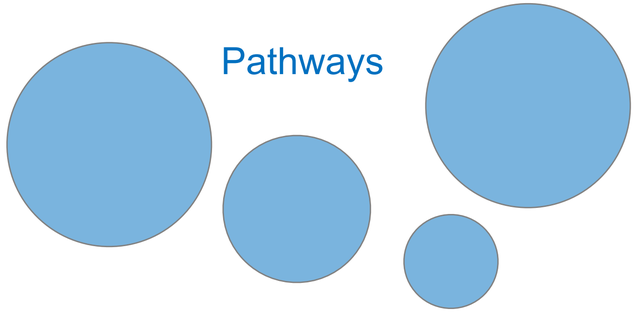
Differentially Expressed Genes (DEGs)



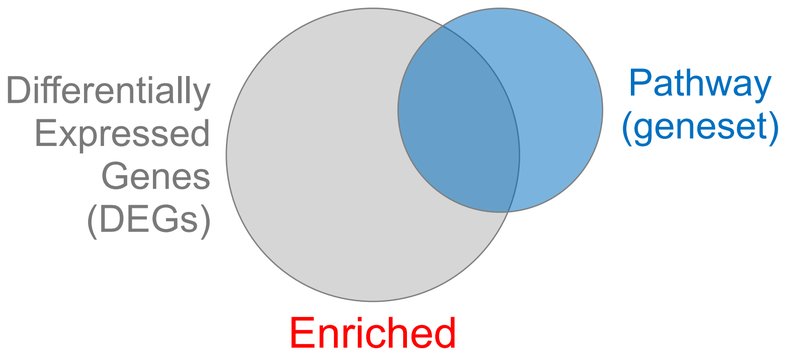
Overlap...

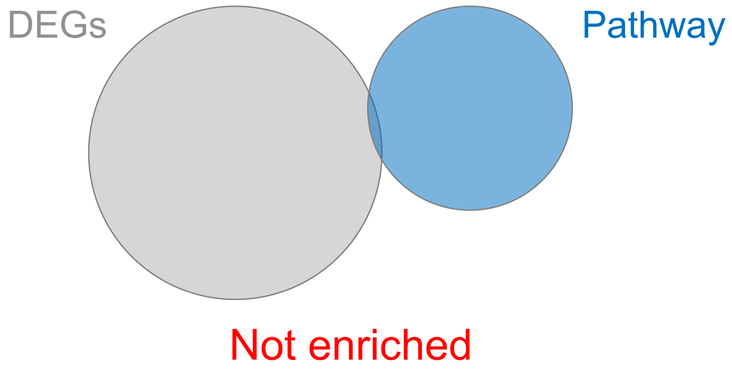
Pathway analysis (geneset enrichment)

Pathways



Principle: **Pathway analysis**

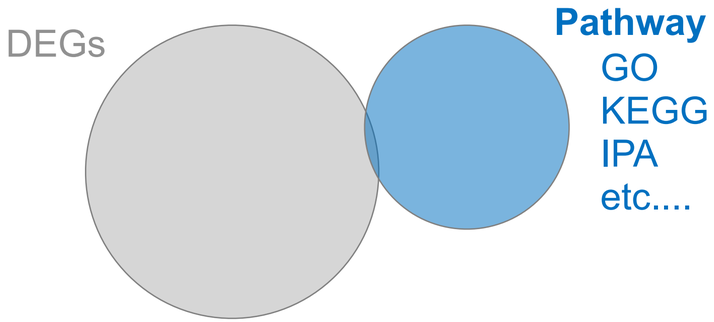




- DEGs come from your experiment ➢ *Critical, needs to be as clean as possible*
- Pathway genes ("geneset") come from annotations ➢ *Important, but typically not a competitive advantage*
- Variations of the math: overlap, ranking, networks... ➢ *Not critical, different algorithms show similar performances*

What functional set databases do you want?

- **Most commonly used:**
  - **Gene Ontology (GO)**
  - **KEGG Pathways** (mostly metabolic)
  - **GeneGO MetaBase**
  - **Ingenuity Pathway Analysis (IPA)**



- Many others...
  - **Enzyme Classification**, **PFAM**, **Reactome**,
  - Disease Ontology, MSigDB, Chemical Entities of Biological Interest, Network of Cancer Genes etc…
  - See: Open Biomedical Ontologies (www.obofoundry.org)

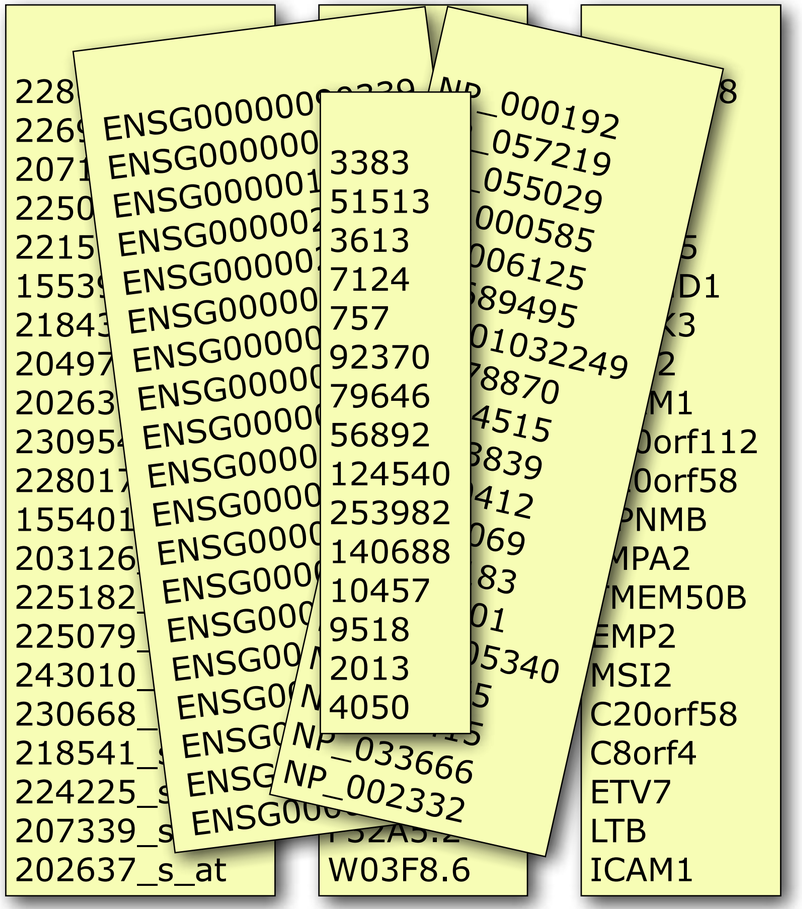
Pathway analysis (a.k.a. geneset enrichment)
Limitations

Side-note:

- **Geneset annotation bias**: can only discover what is already known
- **Non-model organisms**: no high-quality genesets available
- **Post-transcriptional regulation** is neglected
- **Tissue-specific** variations of pathways are not annotated
    - e.g. NF-κB regulates metabolism, not inflammation, in adipocytes
- **Size bias**: stats are influenced by the size of the pathway
    - Many pathways/receptors **converge** to few regulators
      e.g. Tens of innate immune receptors activate four TFs:
      NF-kB, AP-1, IRF3/7, NFAT

Starting point for pathway analysis:
Your gene list

- You have a list of genes/proteins of interest
- You have quantitative data for each gene/protein
  - Fold change
  - p-value
  - Spectral counts
  - Presence/absence



Pathway Analysis