# Data Science for Biologist

UTSouthwestern Medical Center | BICF

# BICF NanoCourses 2019

- Winter
  - Jan 10-31st (F) Data Science For Biologist
  - Feb 14th Accessing Public Data
  - Feb 27-28 MatLab for Scientific Data Exploration
- Spring/Summer
  - Genomic Analysis (Human Variation)
  - Gene Expression and Regulation
  - Single Cell Analysis

# BICF Help Desk

- Our Help Desk provides drop-in consultations at no charge and with no appointment required.
- Please email us at: bicf@utsouthwestern.edu

## Schedule

| Topic | Instructor |
|---|---|
| 1/10/2020 | Room NB2.100A |
| Introduction to R and R Data Structures | Brandi Cantarel |
| Data Importing and Cleaning with Tidyverse | Brandi Cantarel |
| Data Manipulation and Data Joining with dplyr and tidyverse | Spencer Barnes |
| 1/17/2020 | Room NB2.100A |
| Introduction to Statistical Tests | Jeremey Mathews |
| Correlations and Linear Regression | Jeremey Mathews + Jeon Lee |
| Plotting with GGPlot and plotly | Jeon Lee |
| 1/24/2020 | Room NB2.100A |
| Programming basics | Venkat Malladi |
| Loops and Looping functions with Apply | Chris Bennett |
| Scripting and Markdown | Chris Bennett |
| 1/31/2020 | Room NG3.202 |
| Bioconductor | Gervaise Henry |
| Accessing Public Data Though Bioconductor | Gervaise Henry + Spencer Barnes |
| Student Projects | |

# Student Data Clinic

- Bring your data and your scientific question
- Students group by Data Type and Question
- Instructors and TA are available to help you analyze your data
- Email me your project by Jan 24th

# Introduction to R & R Data Structures

Data Science for Biologist

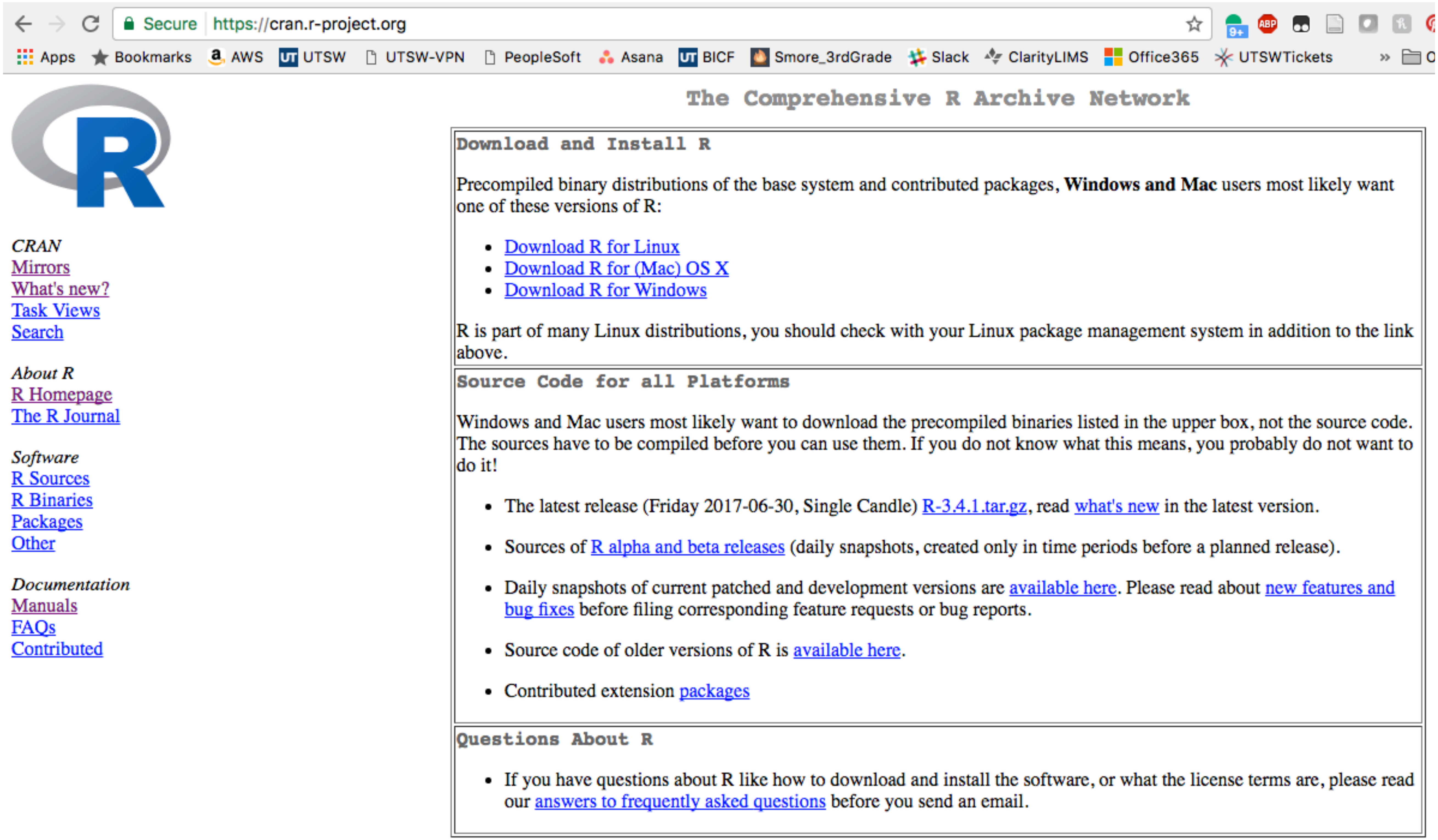
# What is R?

- R is a free software environment for statistical computing and graphics
- Object oriented statistical language
- 2000: R version 1.0.0 was released.
- Quickly became popular for bioinformatics, microarray analysis
- New version released every 6 months
- Versions for Windows (32 and 64bit), UNIX/Linux, MacOS, and RStudio (GUI version)

# What is R?

- Suite of operators for calculations on arrays and matrices
- Sophisticated graphical facilities for display or output files
- Active R community - R-help and R-devel mailing lists
- ~25 base, or standard, packages
- Thousands of contributed packages in repositories:
  - CRAN: http://CRAN.R-project.org
  - Bioconductor: www.bioconductor.org
  - Many more packages available on personal websites

# Downloading R



*CRAN*
Mirrors
What's new?
Task Views
Search

*About R*
R Homepage
The R Journal

*Software*
R Sources
R Binaries
Packages
Other

*Documentation*
Manuals
FAQs
Contributed

## The Comprehensive R Archive Network

### Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

### Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

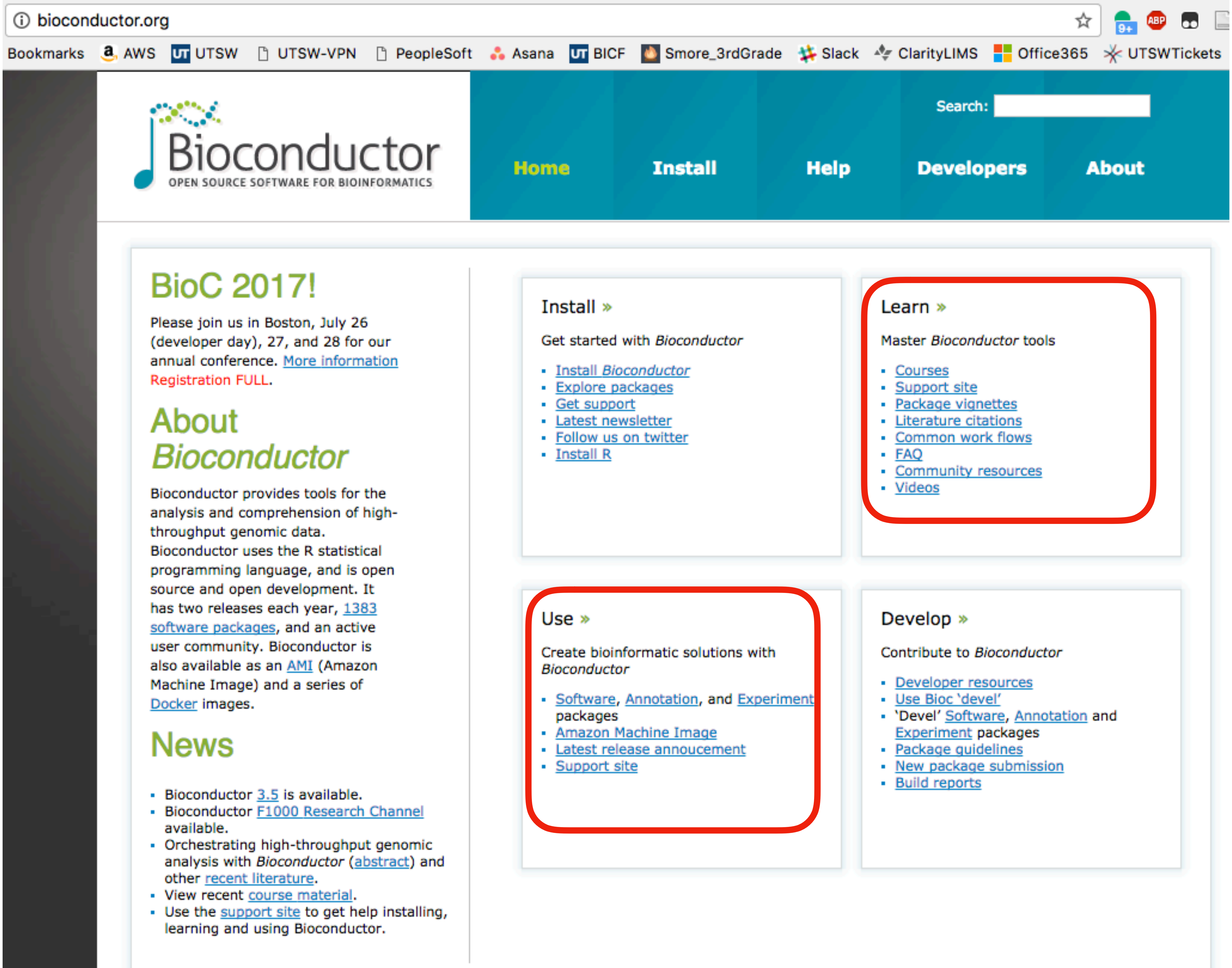
- The latest release (Friday 2017-06-30, Single Candle) R-3.4.1.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

### Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

# Tools for Biologist

- Bioconductor — www.bioconductor.org
- A group of R packages aimed at high-throughput genomic data analysis and genomic annotations
- Open source and open development
- Each Bioconductor package usually has a “vignette” for documentation ie a tutorial for common usage
- Easy to download Bioconductor packages within R:
  - source(“http://www.bioconductor.org/biocLite.R")
  - biocLite()
  - biocLite(“package.name”)



## BioC 2017!

Please join us in Boston, July 26 (developer day), 27, and 28 for our annual conference. More information Registration FULL.

## About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 1383 software packages, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.

## News

- Bioconductor 3.5 is available.
- Bioconductor F1000 Research Channel available.
- Orchestrating high-throughput genomic analysis with *Bioconductor* (abstract) and other recent literature.
- View recent course material.
- Use the support site to get help installing, learning and using Bioconductor.

### Install »

Get started with *Bioconductor*

- Install *Bioconductor*
- Explore packages
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

### Learn »

Master *Bioconductor* tools

- Courses
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos

### Use »

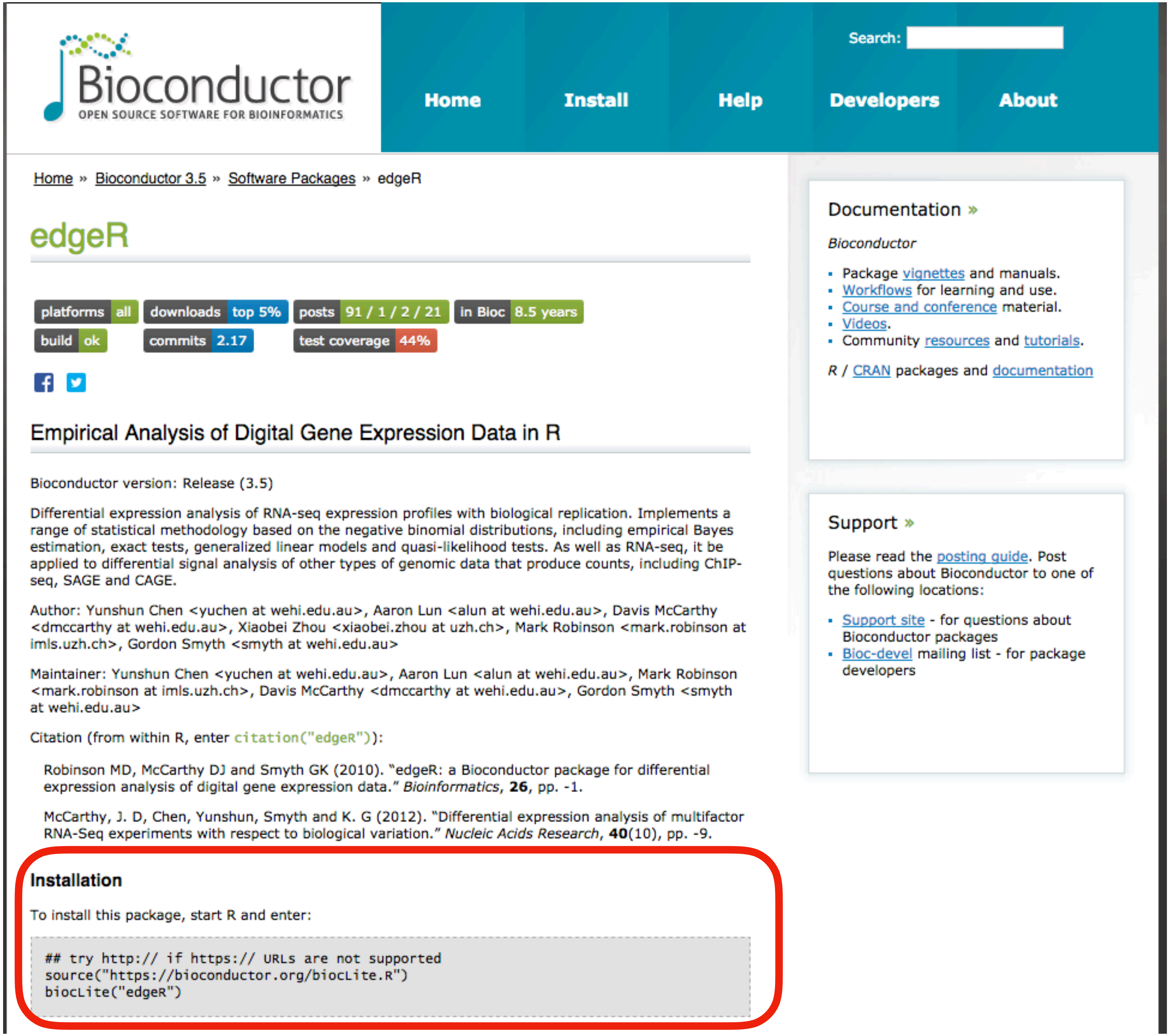
Create bioinformatic solutions with *Bioconductor*

- Software, Annotation, and Experiment packages
- Amazon Machine Image
- Latest release annoucement
- Support site

### Develop »

Contribute to *Bioconductor*

- Developer resources
- Use Bioc 'devel'
- 'Devel' Software, Annotation and Experiment packages
- Package guidelines
- New package submission
- Build reports



Home » Bioconductor 3.5 » Software Packages » edgeR

# edgeR

## Empirical Analysis of Digital Gene Expression Data in R

Bioconductor version: Release (3.5)

Differential expression analysis of RNA-seq expression profiles with biological replication. Implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests. As well as RNA-seq, it be applied to differential signal analysis of other types of genomic data that produce counts, including ChIP-seq, SAGE and CAGE.

Author: Yunshun Chen <yuchen at wehi.edu.au>, Aaron Lun <alun at wehi.edu.au>, Davis McCarthy <dmccarthy at wehi.edu.au>, Xiaobei Zhou <xiaobei.zhou at uzh.ch>, Mark Robinson <mark.robinson at imls.uzh.ch>, Gordon Smyth <smyth at wehi.edu.au>

Maintainer: Yunshun Chen <yuchen at wehi.edu.au>, Aaron Lun <alun at wehi.edu.au>, Mark Robinson <mark.robinson at imls.uzh.ch>, Davis McCarthy <dmccarthy at wehi.edu.au>, Gordon Smyth <smyth at wehi.edu.au>

Citation (from within R, enter `citation("edgeR")`):

Robinson MD, McCarthy DJ and Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, **26**, pp. -1.

McCarthy, J. D, Chen, Yunshun, Smyth and K. G (2012). "Differential expression analysis of multifactor RNA-Seq experiments with biological variation." *Nucleic Acids Research*, **40**(10), pp. -9.

### Installation

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("edgeR")
```

### Documentation »

*Bioconductor*

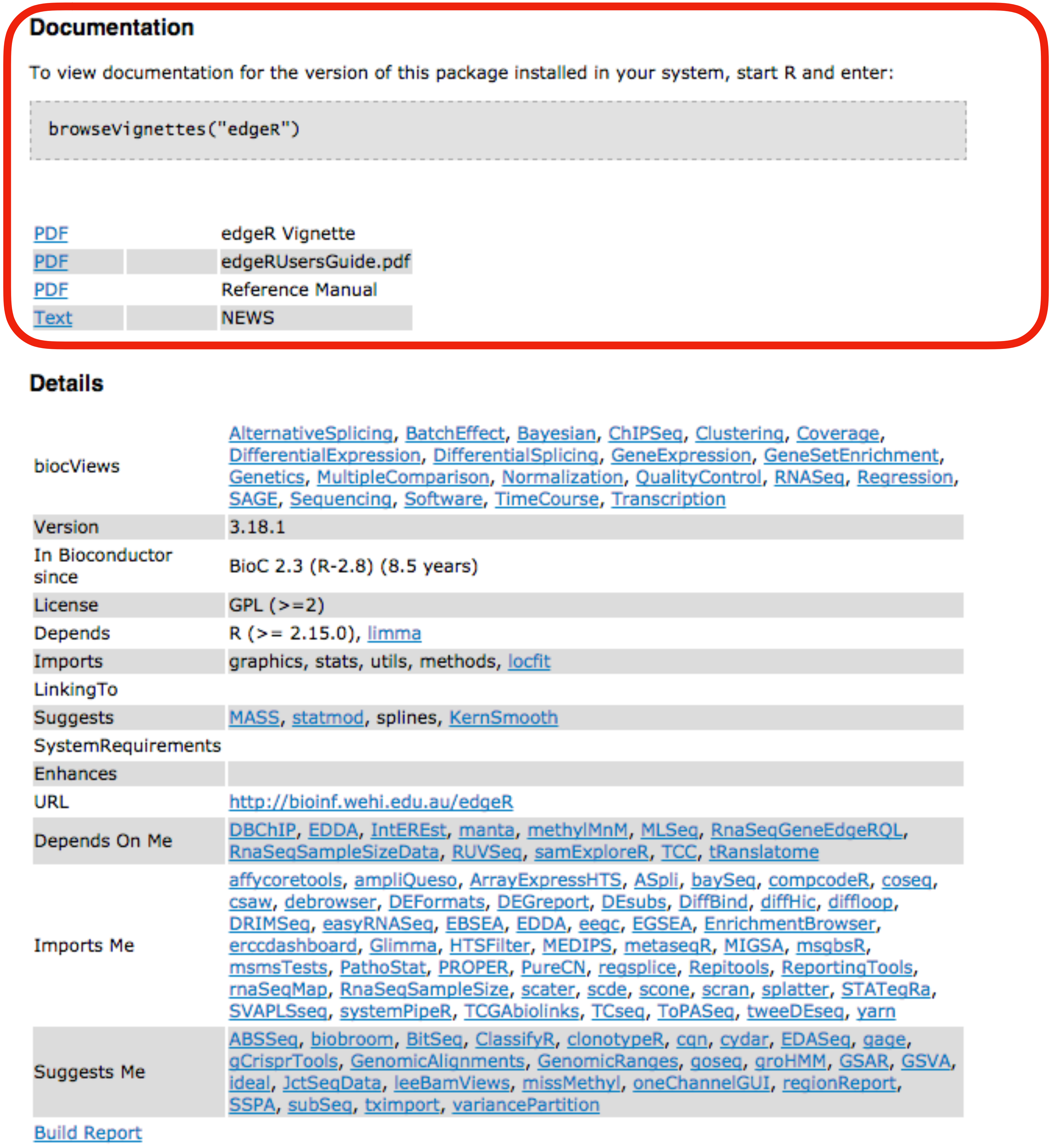
- Package vignettes and manuals.
- Workflows for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

*R* / CRAN packages and documentation

### Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers



## Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("edgeR")
```

| | | |
|---|---|---|
| PDF | | edgeR Vignette |
| PDF | | edgeRUsersGuide.pdf |
| PDF | | Reference Manual |
| Text | | NEWS |

## Details

| | |
|---|---|
| biocViews | AlternativeSplicing, BatchEffect, Bayesian, ChIPSeq, Clustering, Coverage, DifferentialExpression, DifferentialSplicing, GeneExpression, GeneSetEnrichment, Genetics, MultipleComparison, Normalization, QualityControl, RNASeq, Regression, SAGE, Sequencing, Software, TimeCourse, Transcription |
| Version | 3.18.1 |
| In Bioconductor since | BioC 2.3 (R-2.8) (8.5 years) |
| License | GPL (>=2) |
| Depends | R (>= 2.15.0), limma |
| Imports | graphics, stats, utils, methods, locfit |
| LinkingTo | |
| Suggests | MASS, statmod, splines, KernSmooth |
| SystemRequirements | |
| Enhances | |
| URL | http://bioinf.wehi.edu.au/edgeR |
| Depends On Me | DBChIP, EDDA, IntEREst, manta, methylMnM, MLSeq, RnaSeqGeneEdgeRQL, RnaSeqSampleSizeData, RUVSeq, samExploreR, TCC, tRanslatome |
| Imports Me | affycoretools, ampliQueso, ArrayExpressHTS, ASpli, baySeq, compcodeR, coseq, csaw, debrowser, DEFormats, DEGreport, DEsubs, DiffBind, diffHic, diffloop, DRIMSeq, easyRNASeq, EBSEA, EDDA, eegc, EGSEA, EnrichmentBrowser, erccdashboard, Glimma, HTSFilter, MEDIPS, metaseqR, MIGSA, msgbsR, msmsTests, PathoStat, PROPER, PureCN, regsplice, Repitools, ReportingTools, rnaSeqMap, RnaSeqSampleSize, scater, scde, scone, scran, splatter, STATegRa, SVAPLSseq, systemPipeR, TCGAbiolinks, TCseq, ToPASeq, tweeDEseq, yarn |
| Suggests Me | ABSSeq, biobroom, BitSeq, ClassifyR, clonotypeR, cqn, cydar, EDASeq, gage, gCrisprTools, GenomicAlignments, GenomicRanges, goseq, groHMM, GSAR, GSVA, ideal, JctSeqData, leeBamViews, missMethyl, oneChannelGUI, regionReport, SSPA, subSeq, tximport, variancePartition |

Build Report

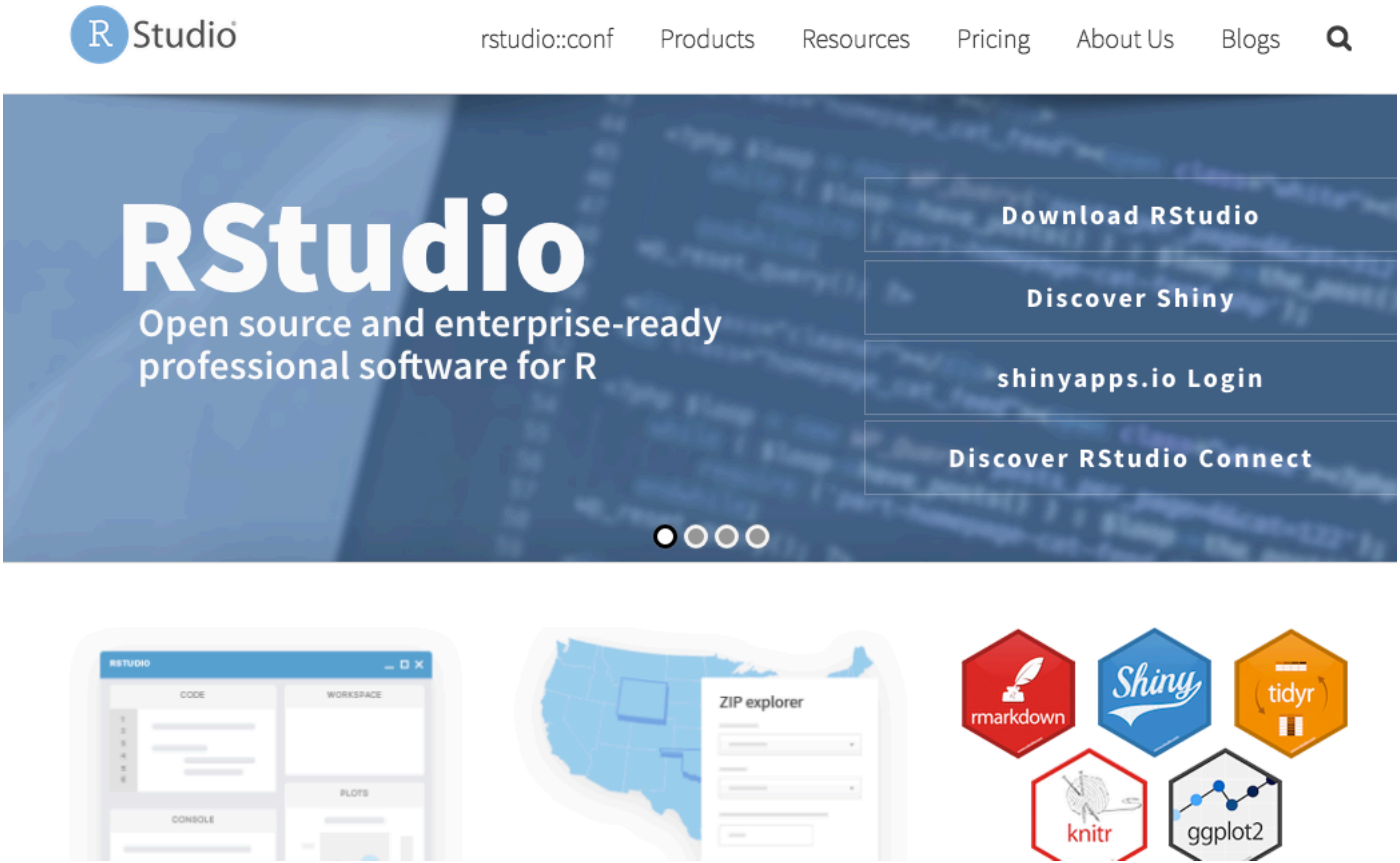
# Manuals and Tutorials

- Under “Manuals” on R website – several in depth tutorials; some basic, some advanced
- Basic introductions to several specific topics in R
  - http://www.cyclismo.org/tutorial/R/
- Various forums available which discuss ranges of errors that users encounter – When in doubt, Just Google and get the syntax!
- Many R books available:
  - General purpose R: e.g., R Cookbook (2011)
  - R in a Nutshell (2010)
  - Specific topics: e.g., Introductory statistics in R
  - Applied Statistical Genetics with R
  - The art of R programming (software design)
  - R Graphics Cookbook
  - Data Mining with R: Learning with Case Studies

# Working in R

- Can work interactively (line by line)
- In Batch mode (run a whole file with code at once)
- Linux Command Line: Rscript filename.r
- In linux, to run interactively type R in terminal window In Windows, Open the R program with interface

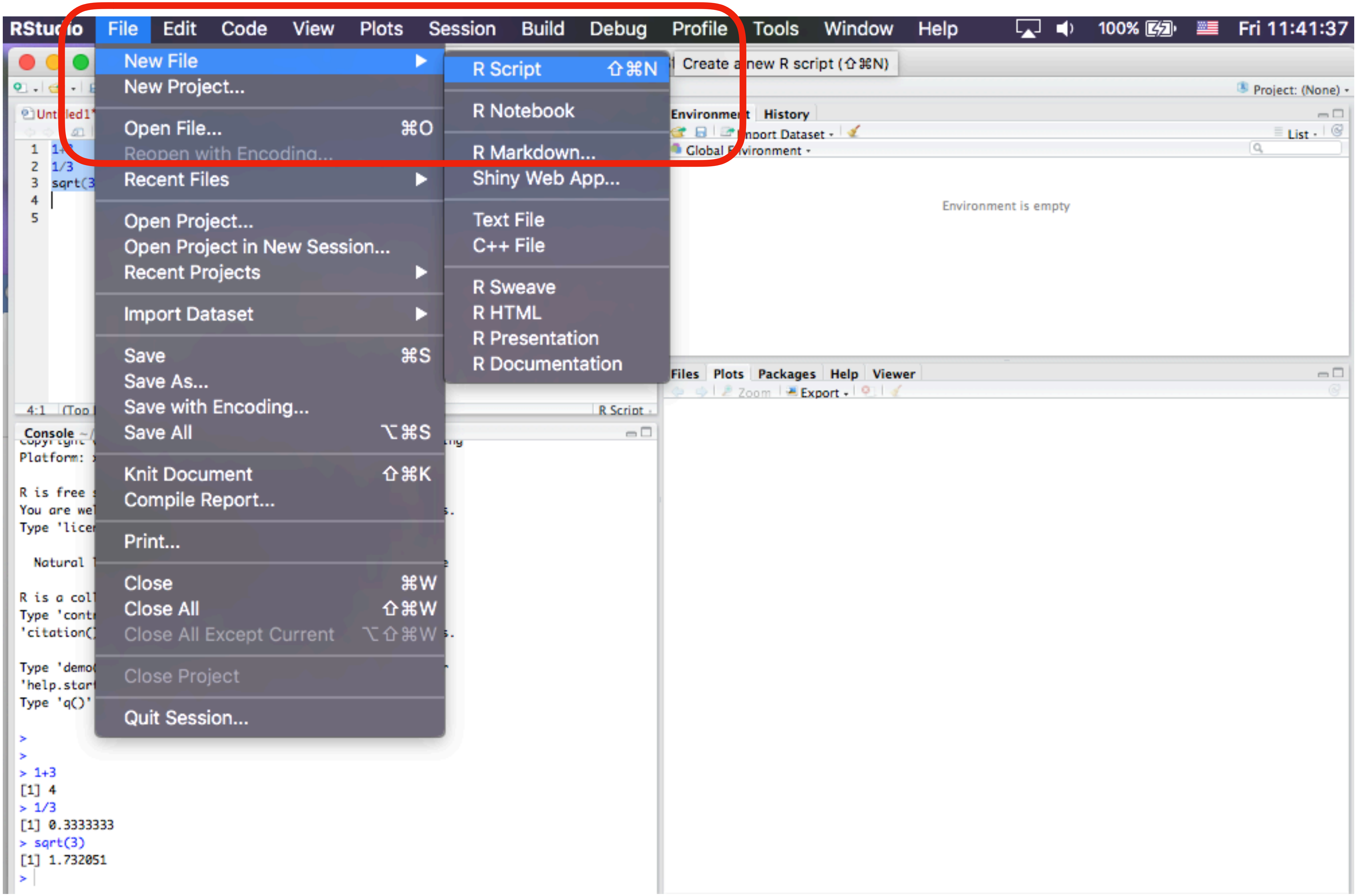
# Graphical Interface to Command Line R



RStudio

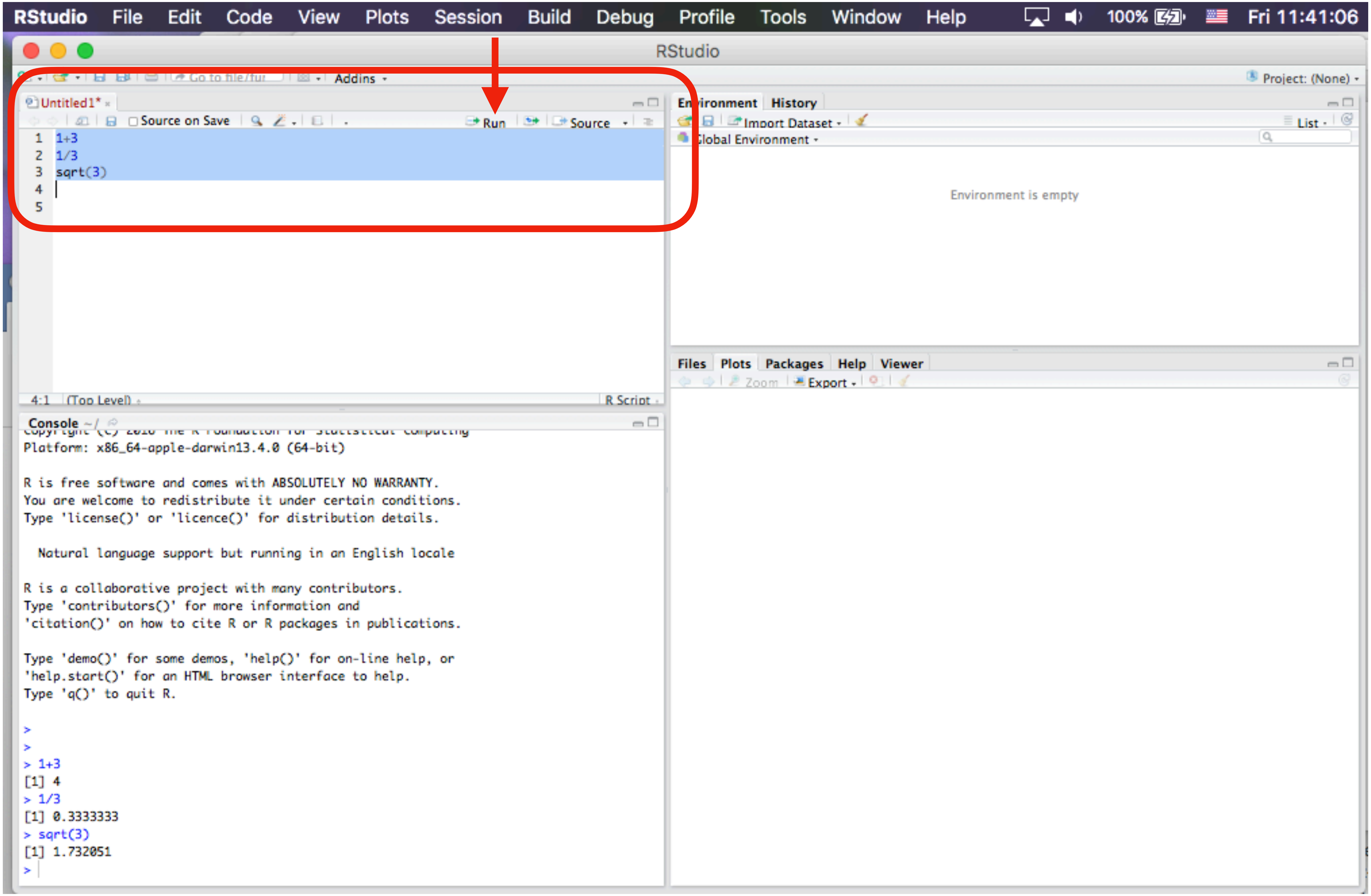
Open source and enterprise-ready professional software for R

**https://www.rstudio.com/**



```
1+3
1/3
sqrt(3
```

```
>
>
> 1+3
[1] 4
> 1/3
[1] 0.3333333
> sqrt(3)
[1] 1.732051
>
```

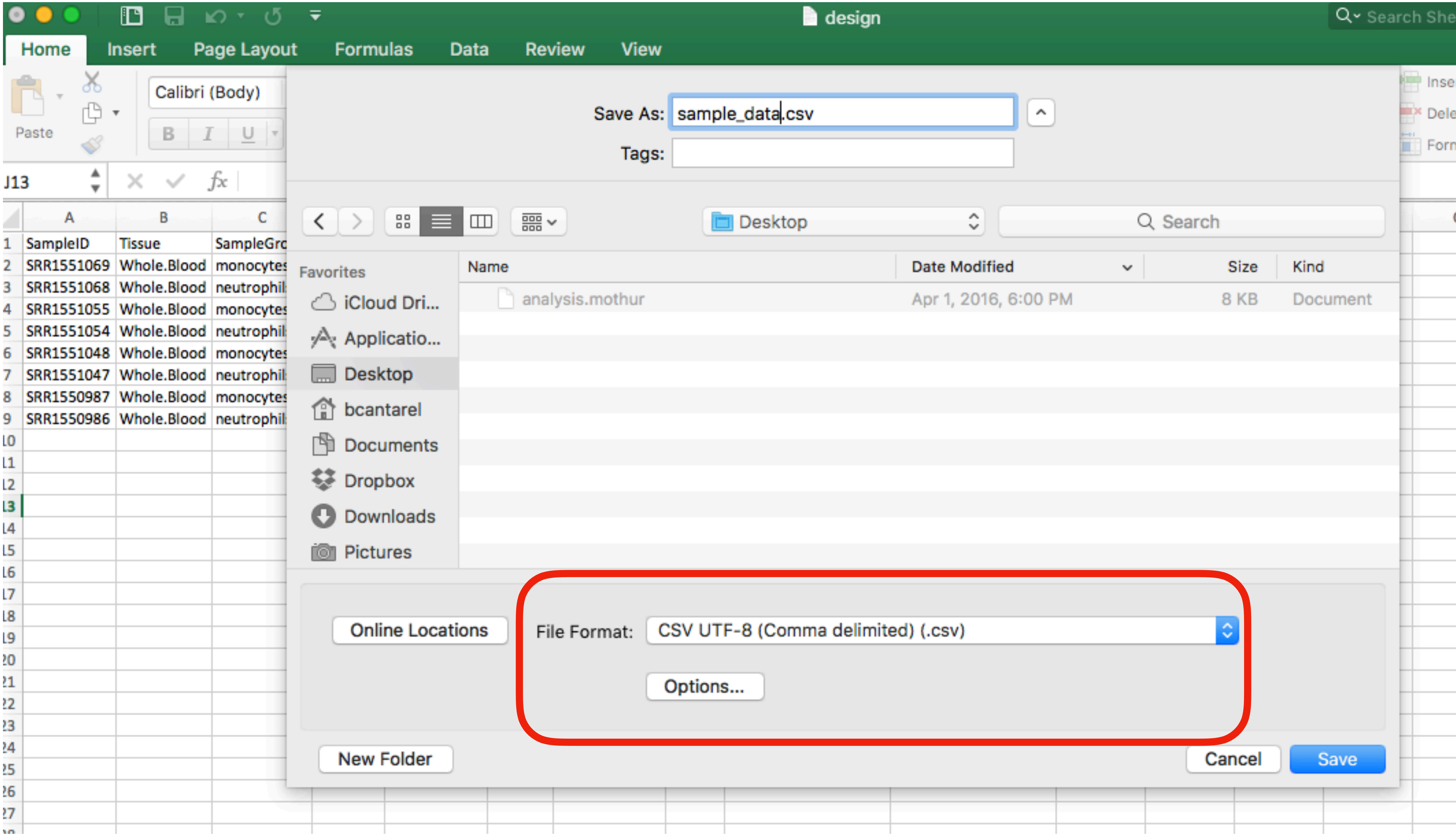


```
1+3
1/3
sqrt(3)
```

```
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (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.

>
>
> 1+3
[1] 4
> 1/3
[1] 0.3333333
> sqrt(3)
[1] 1.732051
>
```

# Go to r_intro.md

# Excel To Data Frame



|  | A | B | C |
|---|---|---|---|
| 1 | SampleID | Tissue | SampleGr |
| 2 | SRR1551069 | Whole.Blood | monocytes |
| 3 | SRR1551068 | Whole.Blood | neutrophil |
| 4 | SRR1551055 | Whole.Blood | monocytes |
| 5 | SRR1551054 | Whole.Blood | neutrophil |
| 6 | SRR1551048 | Whole.Blood | monocytes |
| 7 | SRR1551047 | Whole.Blood | neutrophil |
| 8 | SRR1550987 | Whole.Blood | monocytes |
| 9 | SRR1550986 | Whole.Blood | neutrophil |

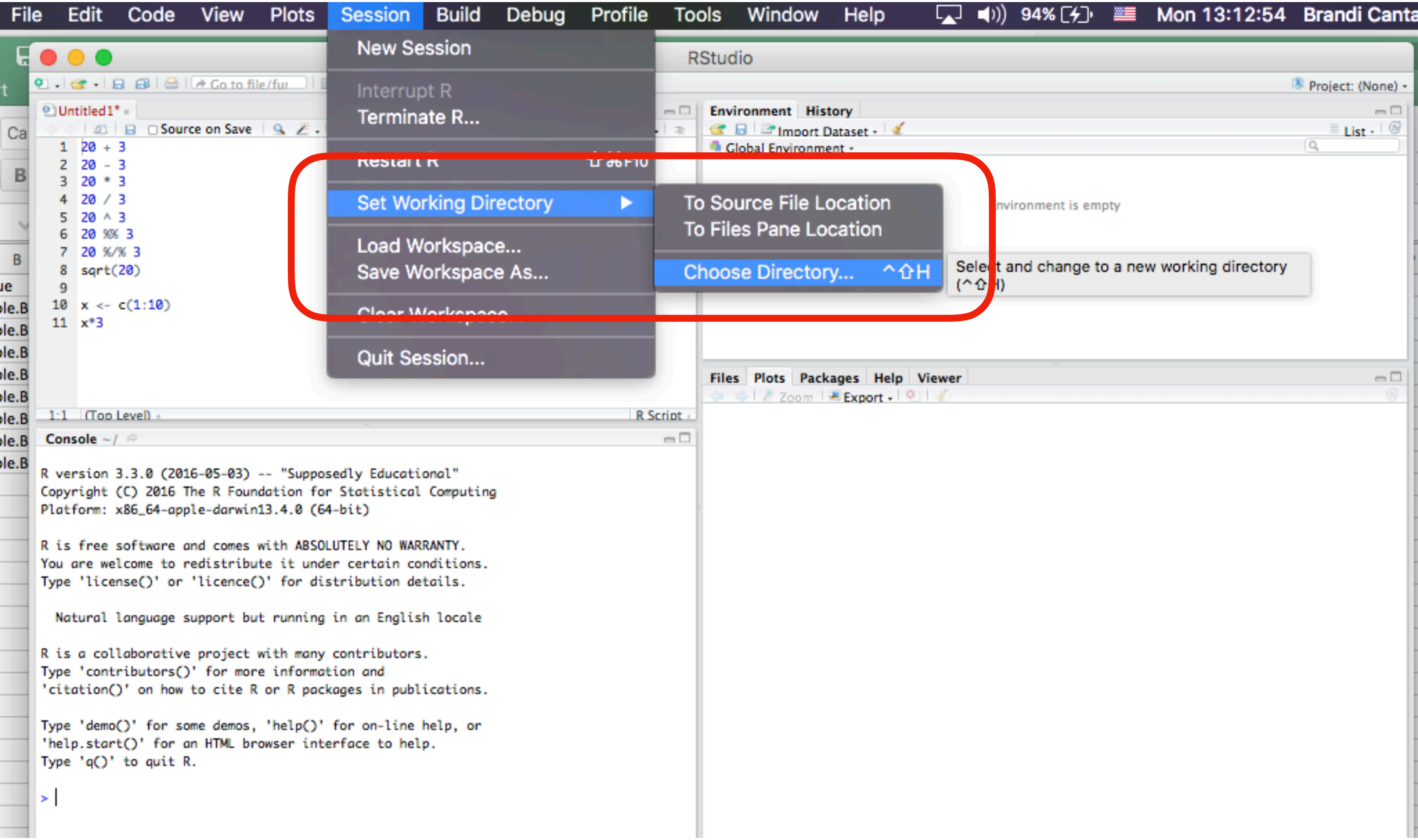
Save As: sample_data.csv

Tags:

| Name | Date Modified | Size | Kind |
|---|---|---|---|
| analysis.mothur | Apr 1, 2016, 6:00 PM | 8 KB | Document |

File Format: CSV UTF-8 (Comma delimited) (.csv)

# Excel To Data Frame

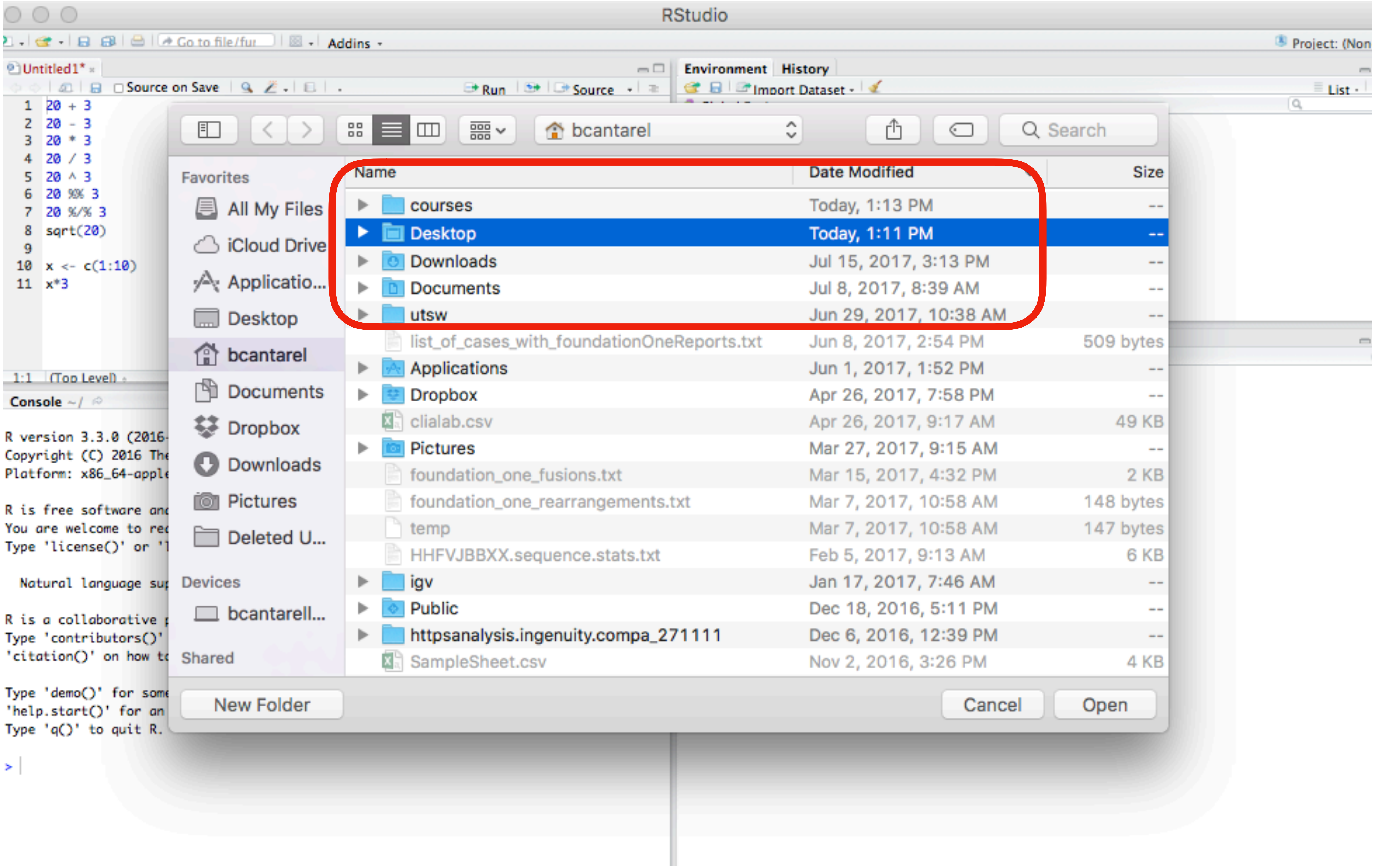


Select and change to a new working directory (^⇧H)

```
20 + 3
20 - 3
20 * 3
20 / 3
20 ^ 3
20 %% 3
20 %/% 3
sqrt(20)

x <- c(1:10)
x*3
```

```
R version 3.3.0 (2016-05-03) -- "Supposedly Educational"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (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.

> |
```

# Excel To Data Frame



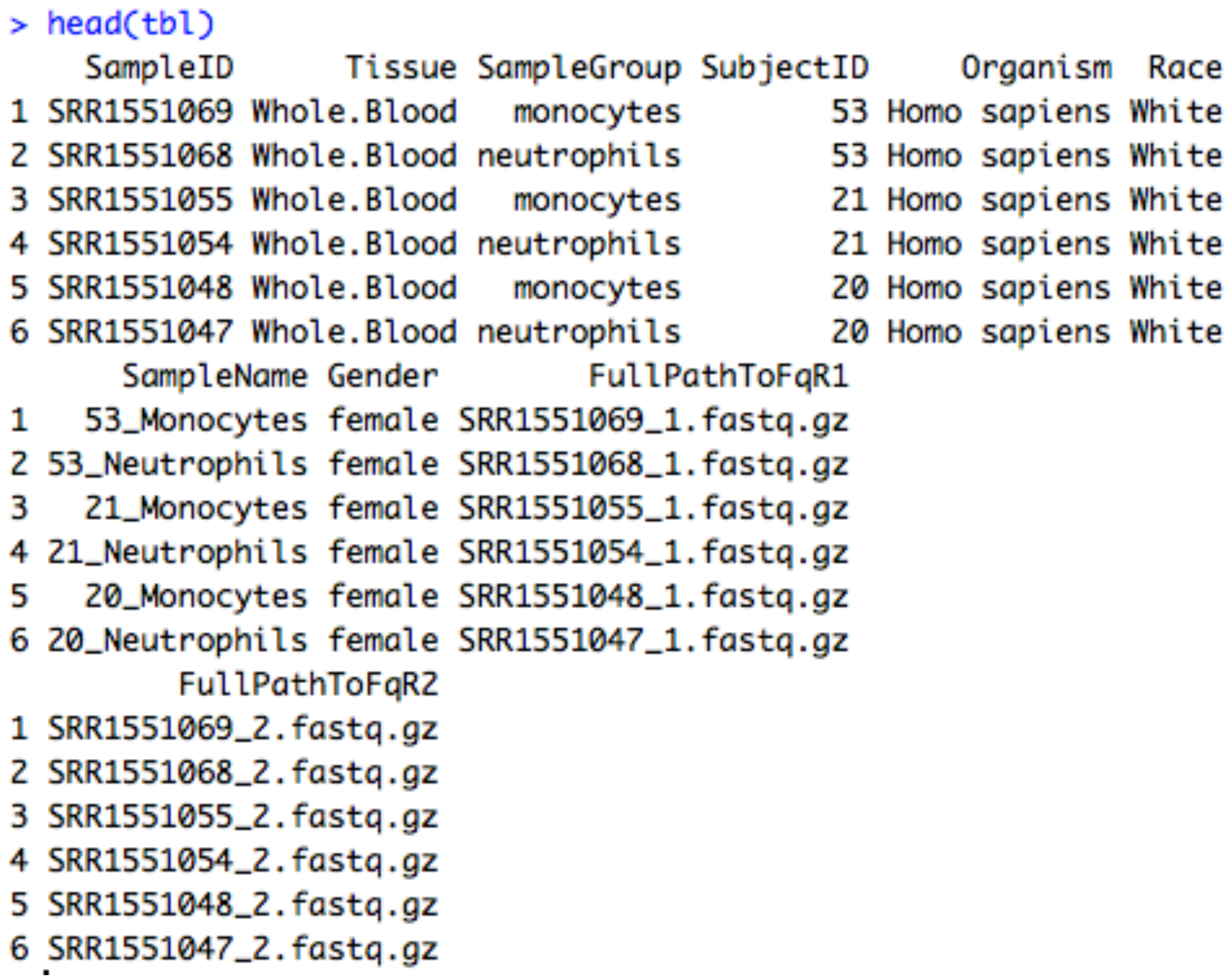
```
20 + 3
20 - 3
20 * 3
20 / 3
20 ^ 3
20 %% 3
20 %/% 3
sqrt(20)

x <- c(1:10)
x*3
```

```
R version 3.3.0 (2016-
Copyright (C) 2016 The
Platform: x86_64-apple

R is free software and
You are welcome to red
Type 'license()' or '1

  Natural language sup

R is a collaborative p
Type 'contributors()'
'citation()' on how to

Type 'demo()' for some
'help.start()' for an
Type 'q()' to quit R.

>
```

| Name | Date Modified | Size |
|---|---|---|
| courses | Today, 1:13 PM | -- |
| Desktop | Today, 1:11 PM | -- |
| Downloads | Jul 15, 2017, 3:13 PM | -- |
| Documents | Jul 8, 2017, 8:39 AM | -- |
| utsw | Jun 29, 2017, 10:38 AM | -- |
| list_of_cases_with_foundationOneReports.txt | Jun 8, 2017, 2:54 PM | 509 bytes |
| Applications | Jun 1, 2017, 1:52 PM | -- |
| Dropbox | Apr 26, 2017, 7:58 PM | -- |
| clialab.csv | Apr 26, 2017, 9:17 AM | 49 KB |
| Pictures | Mar 27, 2017, 9:15 AM | -- |
| foundation_one_fusions.txt | Mar 15, 2017, 4:32 PM | 2 KB |
| foundation_one_rearrangements.txt | Mar 7, 2017, 10:58 AM | 148 bytes |
| temp | Mar 7, 2017, 10:58 AM | 147 bytes |
| HHFVJBBXX.sequence.stats.txt | Feb 5, 2017, 9:13 AM | 6 KB |
| igv | Jan 17, 2017, 7:46 AM | -- |
| Public | Dec 18, 2016, 5:11 PM | -- |
| httpsanalysis.ingenuity.compa_271111 | Dec 6, 2016, 12:39 PM | -- |
| SampleSheet.csv | Nov 2, 2016, 3:26 PM | 4 KB |

# Excel To Data Frame

- `setwd("~/Desktop")`
- `tbl <- read.csv(file=“sample_data.csv",header=TRUE)`



```
> head(tbl)
    SampleID      Tissue SampleGroup SubjectID     Organism  Race
1 SRR1551069 Whole.Blood   monocytes        53 Homo sapiens White
2 SRR1551068 Whole.Blood neutrophils        53 Homo sapiens White
3 SRR1551055 Whole.Blood   monocytes        21 Homo sapiens White
4 SRR1551054 Whole.Blood neutrophils        21 Homo sapiens White
5 SRR1551048 Whole.Blood   monocytes        20 Homo sapiens White
6 SRR1551047 Whole.Blood neutrophils        20 Homo sapiens White
      SampleName Gender        FullPathToFqR1
1   53_Monocytes female SRR1551069_1.fastq.gz
2 53_Neutrophils female SRR1551068_1.fastq.gz
3   21_Monocytes female SRR1551055_1.fastq.gz
4 21_Neutrophils female SRR1551054_1.fastq.gz
5   20_Monocytes female SRR1551048_1.fastq.gz
6 20_Neutrophils female SRR1551047_1.fastq.gz
         FullPathToFqR2
1 SRR1551069_2.fastq.gz
2 SRR1551068_2.fastq.gz
3 SRR1551055_2.fastq.gz
4 SRR1551054_2.fastq.gz
5 SRR1551048_2.fastq.gz
6 SRR1551047_2.fastq.gz
```