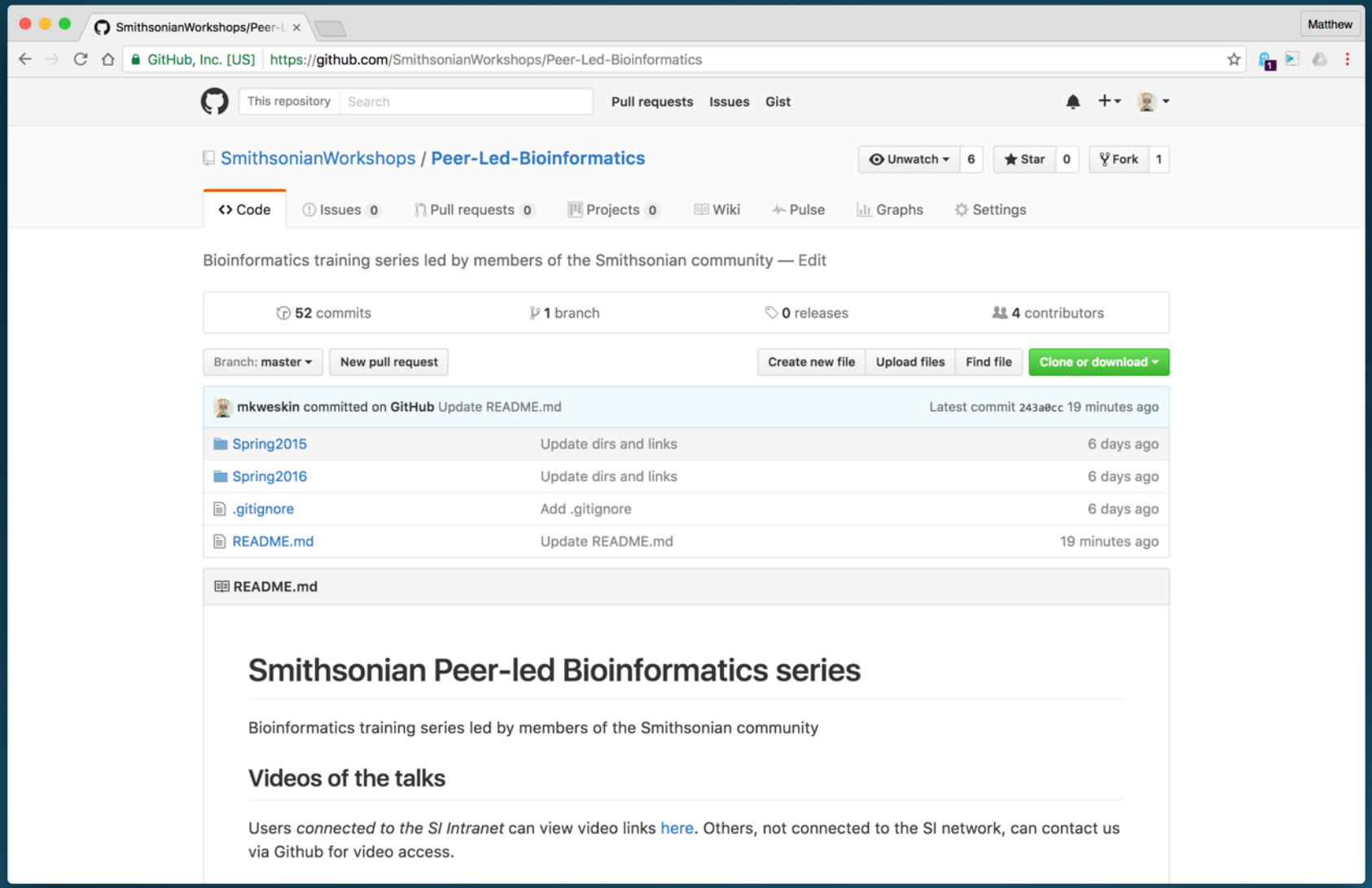
Introduction to R

Series Introduction

Github site:

https://github.com/SmithsonianWorkshops/Peer-Led-Bioinformatics



Date	Speaker	Topics	Packages
Thurs, Oct 13 (! **1pm**, WG33)	Matthew Kweskin & Kenneth (Tripp) Macdonald	• R resources, R studio, Data types and concepts, packages and task views, running R on Hydra; • Introductory R, subsetting data, exporting importing data, data classes, some statistics	
Thurs, Oct 20 (2pm, WG33)	Kenneth (Tripp) Macdonald	Introductory R, subsetting data, exporting importing data, data classes, some statistics (continuation of week 1)	
! **Tues**, Oct 25 (2pm, WG33)	Dietrich Gotzek	Assess MCMC convergence	CODA, BOA, and RWTY
Thurs, Nov 3 (2pm, WG33)	HC Lim	Manipulating phylogenetic trees, labeling, coloring	mainly APE
Thurs, Nov 10 (2pm, WG33)	Michael Lloyd	Loops, files manipulations and repetitive tasks, etc	
! **Tues**, Nov 15 (2pm, WG33)	Caroline Judy & Andrew Gottscho	population structure analysis	Adegenet
! **Tues**, Nov 22 (2pm, WG33)	Mike Trizna	Data wrangling	dplyr, magrittr, tidyr
Thurs, Dec 1 (2pm, WG33)	Carolyn Tepolt	Plotting and data wrangling	ggplot2, data.table
Thurs, Dec 8 (2pm, WG33)	Nicole Angeli	Displaying and analyzing geographic data, pattern analysis, or regression/population modeling.	phytools, spatstat, rgdal, shapefiles, rgeos, raster, and lattice
Thurs, Dec 15 (2pm, WG33)	Steven Calahan	Manipulating geospatial data	rgdal, ggplot2

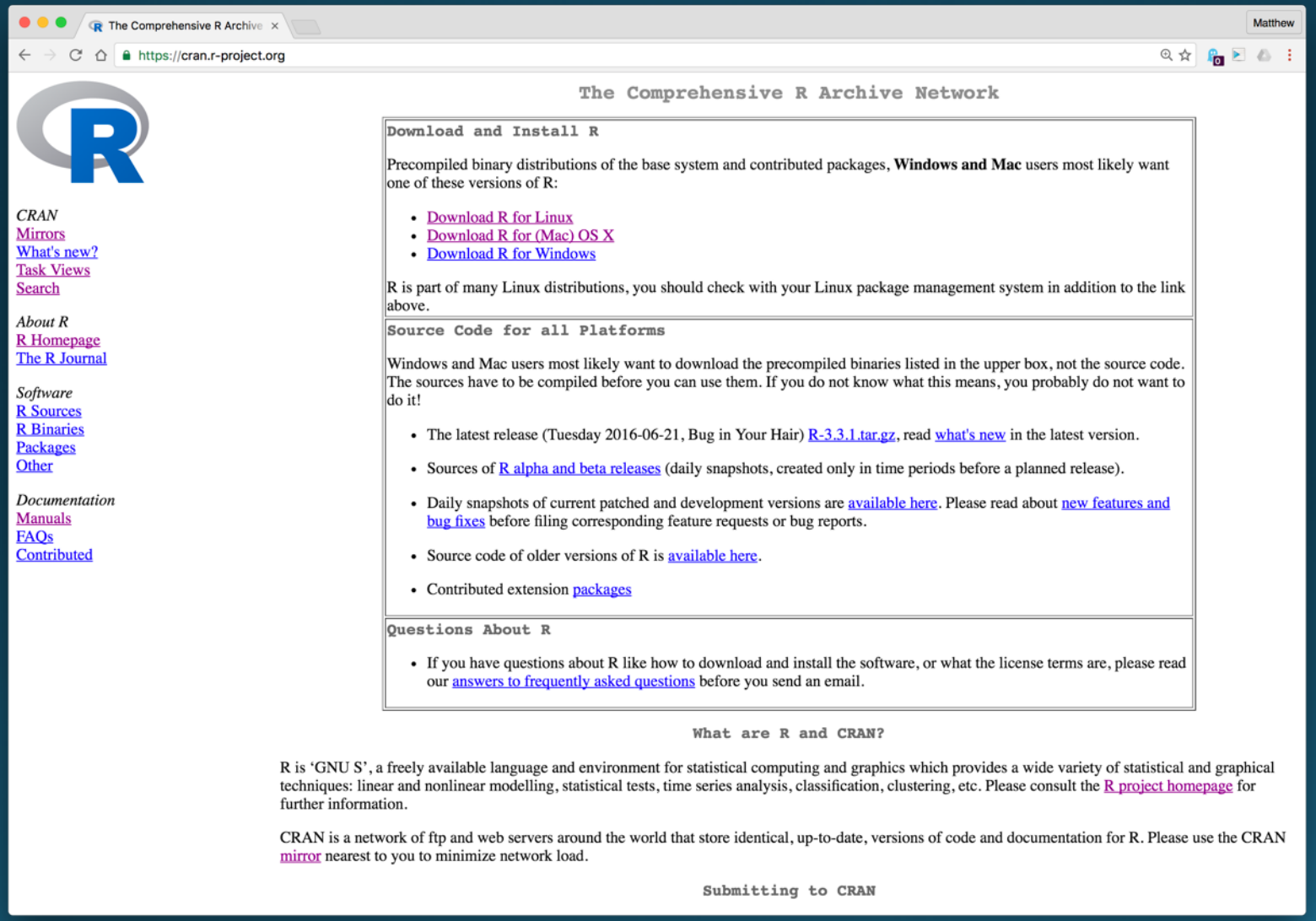
Today

- Matt
  - Basic R concepts and terminology, running on Desktops and Hydra
- Tripp (this week and next)
  - Working with data and stats: example with fluorescence data

Getting and Running R

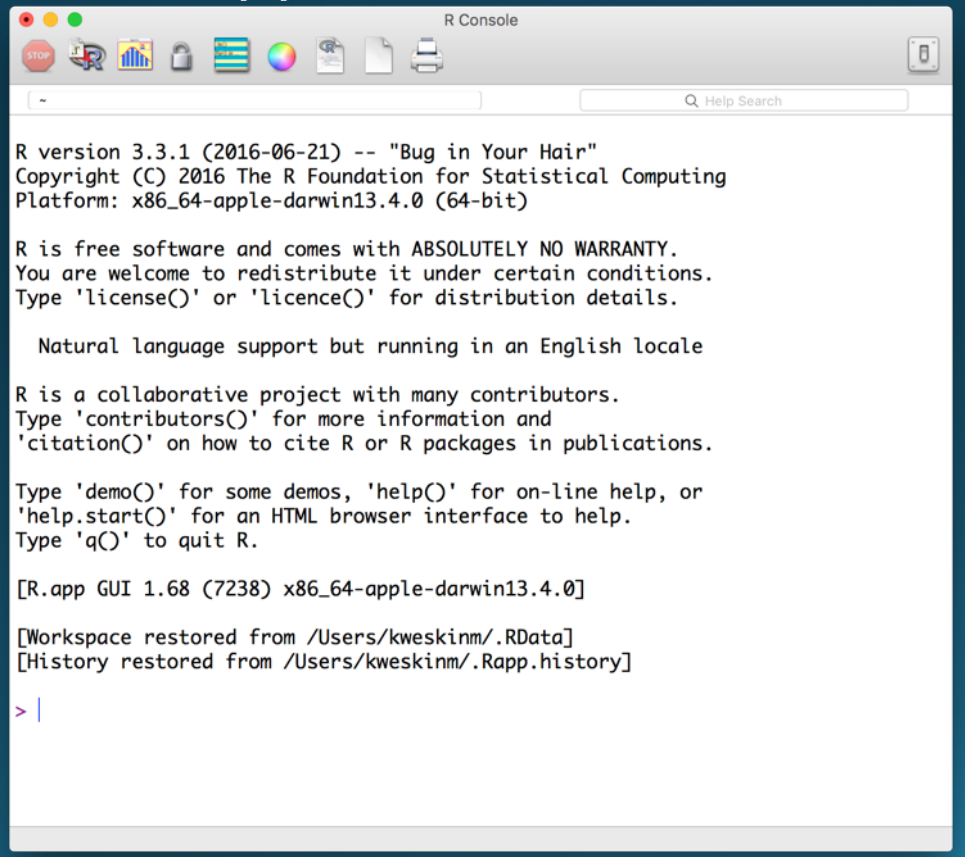
Running R: Desktops

https://cran.r-project.org



Running R: Desktops

- Command line: **`R`**
- R GUI application in /Applications (on Macs)



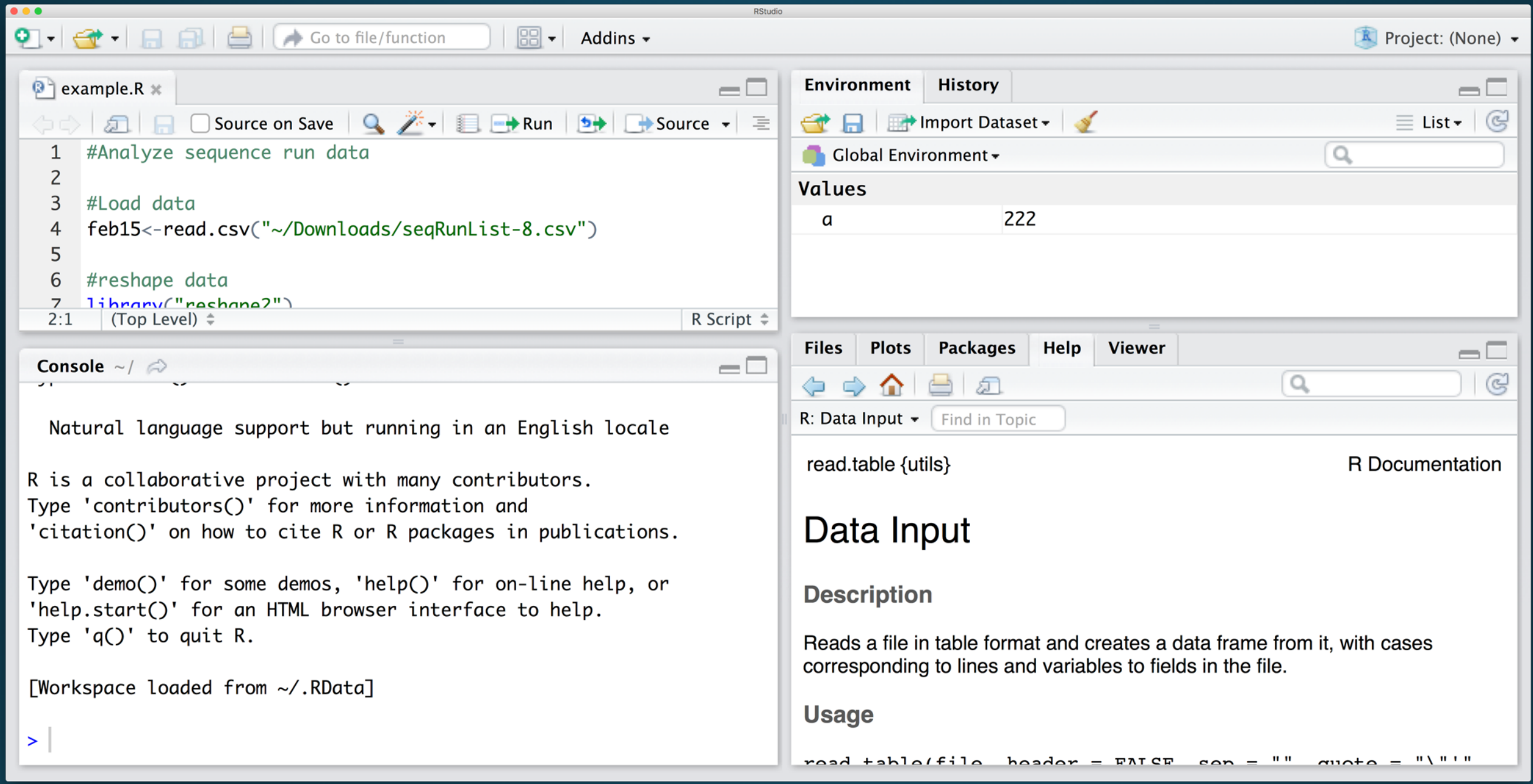
">" Is the interactive prompt in the console

Desktops: R studio

- RStudio.com
- Free, open source development environment

<demo>

- 3-4 panes: Console, Source (initially hidden), Env/History, packages/files/plots/etc
- Source: Running from the Source pane (cmd+return), tab-complete, context coloring
- Pop-up function tips


RStudio
Go to file/function
Addins
Project: (None)
example.R
Source on Save
Run
Source
1 #Analyze sequence run data
2
3 #Load data
4 feb15<-read.csv("~/Downloads/seqRunList-8.csv")
5
6 #reshape data
7 library("reshape2")
2:1
(Top Level)
R Script
Console ~/
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.
[Workspace loaded from ~/.RData]
>
Environment
History
Import Dataset
List
Global Environment
Values
a
222
Files
Plots
Packages
Help
Viewer
R: Data Input
Find in Topic
read.table {utils}
R Documentation
Data Input
Description
Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.
Usage

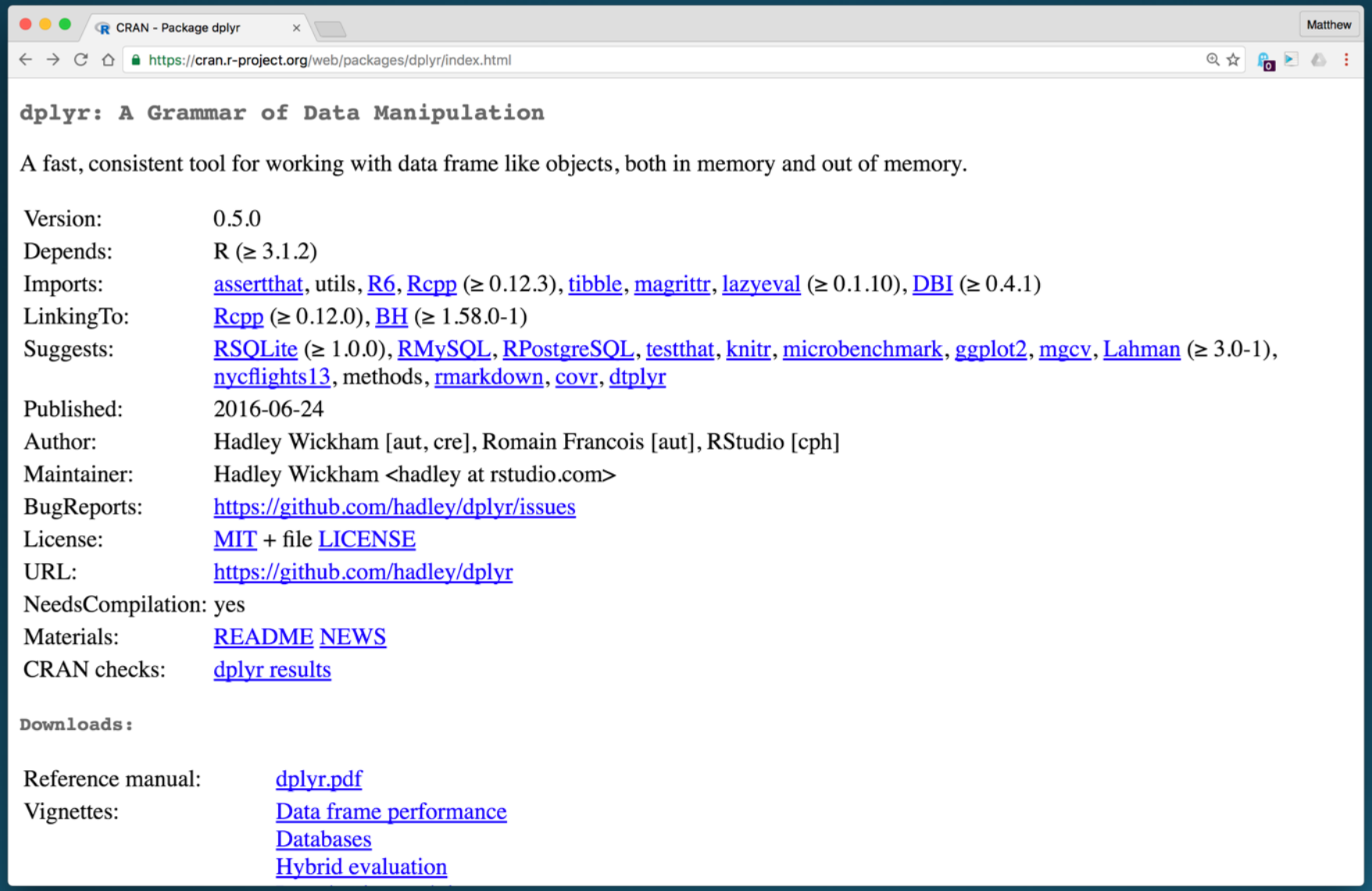
Running R: Servers (Hydra)

```
module load tools/R
```

- Currently installed: 3.2.1
- Run via job scripts, not interactively
- You can install packages on your own (we'll cover later)

Official Documentation: Online

https://cran.r-project.org/web/packages/



dplyr: A Grammar of Data Manipulation

A fast, consistent tool for working with data frame like objects, both in memory and out of memory.

Version:	0.5.0
Depends:	R (≥ 3.1.2)
Imports:	assertthat, utils, R6, Rcpp (≥ 0.12.3), tibble, magrittr, lazyeval (≥ 0.1.10), DBI (≥ 0.4.1)
LinkingTo:	Rcpp (≥ 0.12.0), BH (≥ 1.58.0-1)
Suggests:	RSQLite (≥ 1.0.0), RMySQL, RPostgreSQL, testthat, knitr, microbenchmark, ggplot2, mgcv, Lahman (≥ 3.0-1), nycflights13, methods, rmarkdown, covr, dtplyr
Published:	2016-06-24
Author:	Hadley Wickham [aut, cre], Romain Francois [aut], RStudio [cph]
Maintainer:	Hadley Wickham <hadley at rstudio.com>
BugReports:	https://github.com/hadley/dplyr/issues
License:	MIT + file LICENSE
URL:	https://github.com/hadley/dplyr
NeedsCompilation:	yes
Materials:	README NEWS
CRAN checks:	dplyr results

Downloads:

Reference manual:	dplyr.pdf
Vignettes:	Data frame performance Databases Hybrid evaluation

Official Documentation: in R

```
> ?function
```

- Documentation page for function (exact search)

```
> ??keyword
```

- Search documentation for keyword

```
> help(package="ape")
```

- All the functions for a package

- GUI: Search boxes in Help windows; f1 in RStudio

Data concepts and terminology

Variables

```
> a <- 10
```

"a gets the value of 10"

```
> a
[1] 10

> a+3
[1] 13
```

- "<-" is assignment
- "[1]" is an index.

Variable names:

- Case sensitive
- Can't start with a number

Data types

Numeric

```
> a <- 10
```

Character

```
> b <- "hello world"
```

Logic

```
> c <- TRUE
T TRUE
F FALSE
```

Operations perform differently on different data types

Vector

One dimensional sequence of values

Used extensively


```
> myvector <- c(23,13,10)
> myvector
[1] 23 13 10
```

`c()` first function shown. "c" stands for "combine"

```
> myvector2 <- c("one", 2, 3)
> myvector2
[1] "one" "2"   "3"
```

Every element is made the ***same*** data type

Matrix

Two dimensional sequence of values

1	2	3
4	5	6

Every element is made the ***same*** data type

```
> mat <- matrix(c(1,2,3,4,5,6),nrow=2)
> mat
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

List

One dimensional, Each element can be a ***different*** data type


```
> lst<-list("one",2,3)
> lst
[[1]]
[1] "one"

[[2]]
[1] 2

[[3]]
[1] 3
```

`[[1]]` is element index in list

Data Frame

Two dimensional, Each element can be a ***different*** data type

Used extensively

SampleID	Age	Length	Width
"2016oct13-01"	5	13.1	4.3
"2016oct13-02"	6	11.2	3.9
"2016oct13-03"	5	8.2	3.1

Columns can be used as a vector

Typically read from external source

Summary

Dimensions	Same	Multiple
	Data types	
1	Vector	List
2	Matrix	Data Frame

Functions and Packages

Functions

- Functions are in add-on packages or part of R base

```
function(argument1, argument2=TRUE,
argument3="character",…)
```

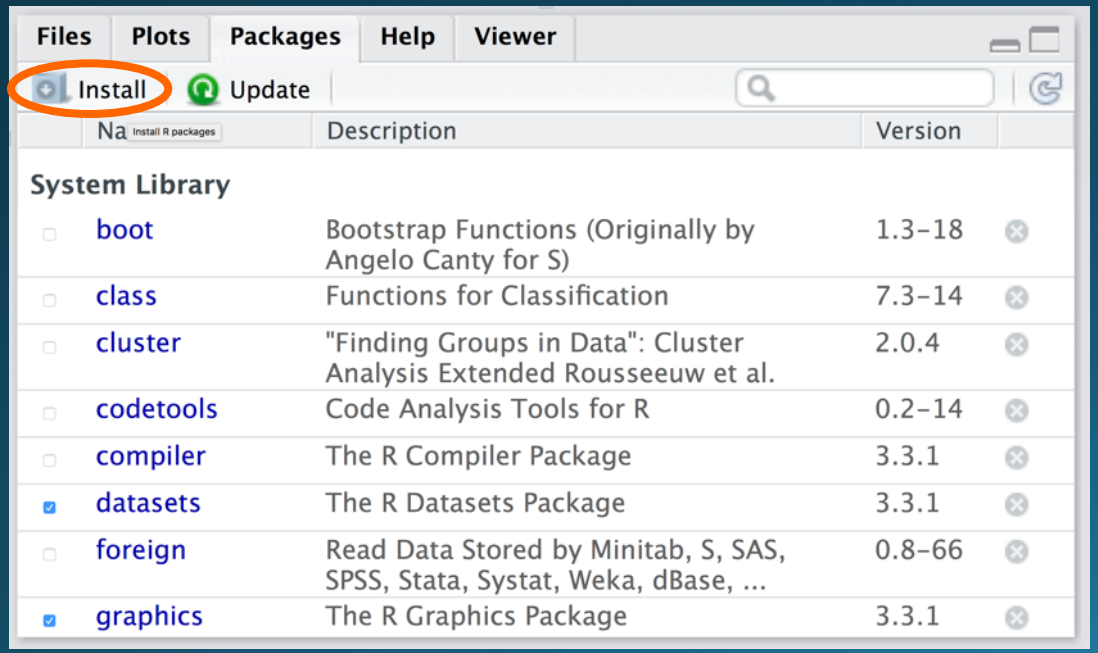
```
> sqrt(9)
  [1] 3

> x<-c(13,2)
> mean(x)
  [1] 7.5

> round(6.89, digits=1)
  [1] 6.9
```

Install packages

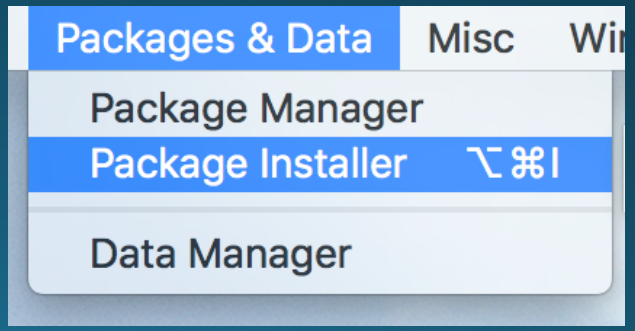
- Command line:

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

- GUI:



Mirror: copy of central CRAN repository. Choose a local one.



Install packages: Hydra

```
$ module load tools/R
$ R
> install.packages("ggplot2")
```

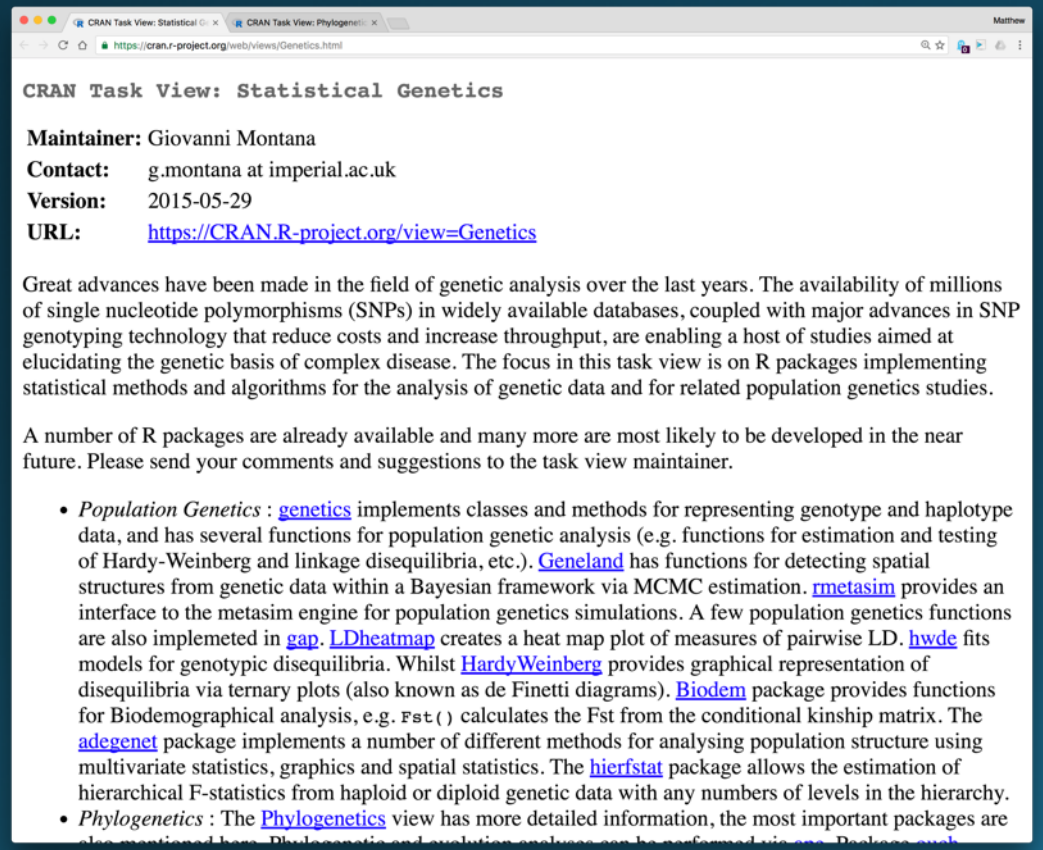
- Run install steps (only) interactively without a job script.
- Only need to install one time

```
…Would you like to use a personal library
instead?  (y/n) y
Would you like to create a personal
library~/R/x86_64-unknown-linux-gnu-library/3.2
to install packages into?  (y/n) y
```

Packages: Task Views

https://cran.r-project.org/web/views/

Statistical Genetics



CRAN Task View: Statistical Genetics

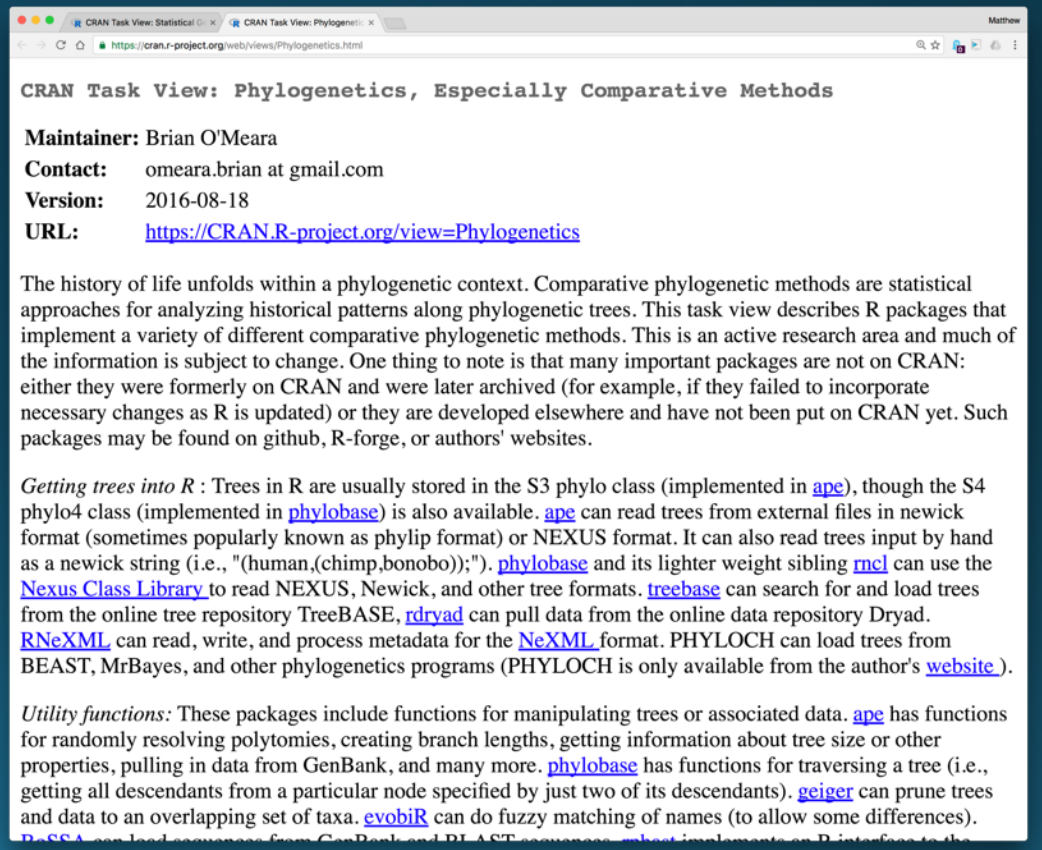
Maintainer: Giovanni Montana
Contact: g.montana at imperial.ac.uk
Version: 2015-05-29
URL: https://CRAN.R-project.org/view=Genetics

Great advances have been made in the field of genetic analysis over the last years. The availability of millions of single nucleotide polymorphisms (SNPs) in widely available databases, coupled with major advances in SNP genotyping technology that reduce costs and increase throughput, are enabling a host of studies aimed at elucidating the genetic basis of complex disease. The focus in this task view is on R packages implementing statistical methods and algorithms for the analysis of genetic data and for related population genetics studies.

A number of R packages are already available and many more are most likely to be developed in the near future. Please send your comments and suggestions to the task view maintainer.

- *Population Genetics* : genetics implements classes and methods for representing genotype and haplotype data, and has several functions for population genetic analysis (e.g. functions for estimation and testing of Hardy-Weinberg and linkage disequilibria, etc.). Geneland has functions for detecting spatial structures from genetic data within a Bayesian framework via MCMC estimation. rmetasim provides an interface to the metasim engine for population genetics simulations. A few population genetics functions are also implemeted in gap. LDheatmap creates a heat map plot of measures of pairwise LD. hwde fits models for genotypic disequilibria. Whilst HardyWeinberg provides graphical representation of disequilibria via ternary plots (also known as de Finetti diagrams). Biodem package provides functions for Biodemographical analysis, e.g. `Fst()` calculates the Fst from the conditional kinship matrix. The adegenet package implements a number of different methods for analysing population structure using multivariate statistics, graphics and spatial statistics. The hierfstat package allows the estimation of hierarchical F-statistics from haploid or diploid genetic data with any numbers of levels in the hierarchy.
- *Phylogenetics* : The Phylogenetics view has more detailed information, the most important packages are

Phylogenetics



CRAN Task View: Phylogenetics, Especially Comparative Methods

Maintainer: Brian O'Meara
Contact: omeara.brian at gmail.com
Version: 2016-08-18
URL: https://CRAN.R-project.org/view=Phylogenetics

The history of life unfolds within a phylogenetic context. Comparative phylogenetic methods are statistical approaches for analyzing historical patterns along phylogenetic trees. This task view describes R packages that implement a variety of different comparative phylogenetic methods. This is an active research area and much of the information is subject to change. One thing to note is that many important packages are not on CRAN: either they were formerly on CRAN and were later archived (for example, if they failed to incorporate necessary changes as R is updated) or they are developed elsewhere and have not been put on CRAN yet. Such packages may be found on github, R-forge, or authors' websites.

Getting trees into R : Trees in R are usually stored in the S3 phylo class (implemented in ape), though the S4 phylo4 class (implemented in phylobase) is also available. ape can read trees from external files in newick format (sometimes popularly known as phylip format) or NEXUS format. It can also read trees input by hand as a newick string (i.e., "(human,(chimp,bonobo));"). phylobase and its lighter weight sibling rncl can use the Nexus Class Library to read NEXUS, Newick, and other tree formats. treebase can search for and load trees from the online tree repository TreeBASE, rdryad can pull data from the online data repository Dryad. RNeXML can read, write, and process metadata for the NeXML format. PHYLOCH can load trees from BEAST, MrBayes, and other phylogenetics programs (PHYLOCH is only available from the author's website).

Utility functions: These packages include functions for manipulating trees or associated data. ape has functions for randomly resolving polytomies, creating branch lengths, getting information about tree size or other properties, pulling in data from GenBank, and many more. phylobase has functions for traversing a tree (i.e., getting all descendants from a particular node specified by just two of its descendants). geiger can prune trees and data to an overlapping set of taxa. evobiR can do fuzzy matching of names (to allow some differences).

"ctv" package will install groups of packages

Using a package

- Command line:

```
> library("ggplot2")
```

- GUI: