

Introduction to R

Series Introduction

Github site:

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

The screenshot shows the GitHub repository page for the 'Peer-Led-Bioinformatics' repository. The repository has 52 commits, 1 branch, 0 releases, and 4 contributors. The latest commit was made 19 minutes ago by mkweskin. The repository description is "Bioinformatics training series led by members of the Smithsonian community". It includes sections for 'Videos of the talks' and a note about video access via the SI Intranet or GitHub.

SmithsonianWorkshops / Peer-Led-Bioinformatics

52 commits · 1 branch · 0 releases · 4 contributors

Branch: master · New pull request · Create new file · Upload files · Find file · Clone or download

mkweskin committed on GitHub Update README.md · Latest commit 243a0cc 19 minutes ago

Spring2015 · Update dirs and links · 6 days ago

Spring2016 · Update dirs and links · 6 days ago

.gitignore · Add .gitignore · 6 days ago

README.md · Update README.md · 19 minutes ago

README.md

Smithsonian Peer-led Bioinformatics series

Bioinformatics training series led by members of the Smithsonian community

Videos of the talks

Users connected to the SI Intranet can view video links [here](#). Others, not connected to the SI network, can contact us via Github for video access.

Date	Speaker	Topics	Packages
Thurs, Oct 13 (! 1pm, WG33)	Matthew Kweskin & Kenneth (Tripp) Macdonald	<ul style="list-style-type: none"> • 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 , tidyverse
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>

The screenshot shows a web browser window displaying the CRAN homepage at <https://cran.r-project.org>. The page title is "The Comprehensive R Archive Network". On the left, there's a sidebar with links for "CRAN Mirrors", "What's new?", "Task Views", "Search", "About R", "R Homepage", and "The R Journal". Below that are links for "Software", "R Sources", "R Binaries", "Packages", and "Other". Further down are links for "Documentation", "Manuals", "FAQs", and "Contributed". The main content area has three sections: "Download and Install R", "Source Code for all Platforms", and "Questions About R". Each section contains a bulleted list of links or instructions. At the bottom, there's a summary of what R is and how to use CRAN, along with a link to the R project homepage.

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 (Tuesday 2016-06-21, Bug in Your Hair) [R-3.3.1.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.

What are R and CRAN?

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the [R project homepage](#) for further information.

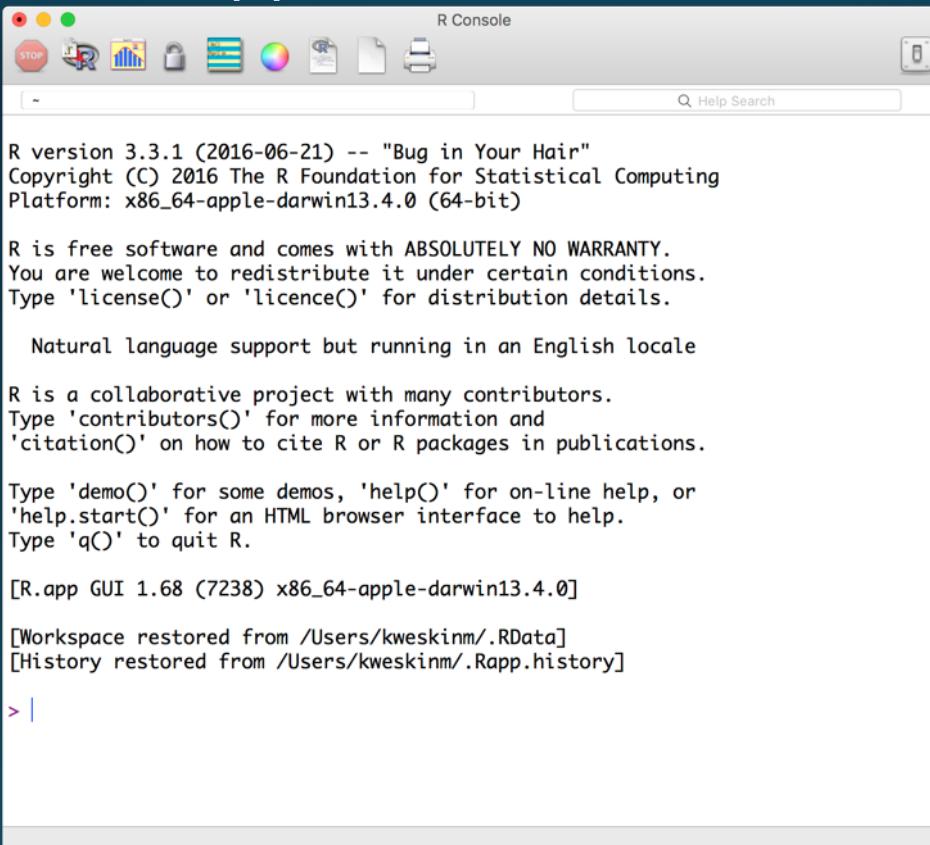
CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN [mirror](#) nearest to you to minimize network load.

Submitting to CRAN

Running R: Desktops



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



The screenshot shows the R Console window on a Mac OS X desktop. The window title is "R Console". The content area displays the R startup message:

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
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.

[R.app GUI 1.68 (7238) x86_64-apple-darwin13.4.0]

[Workspace restored from /Users/kweskinm/.RData]
[History restored from /Users/kweskinm/.Rapp.history]

> |
```

">" 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

Project: (None)

example.R *

Source on Save | Run | Source | List | G

```
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
```

Environment History

Import Dataset | Global Environment | Values

a 222

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]

> |

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

```
read.table(file, header = FALSE, sep = "", quote = "\"\"",
```

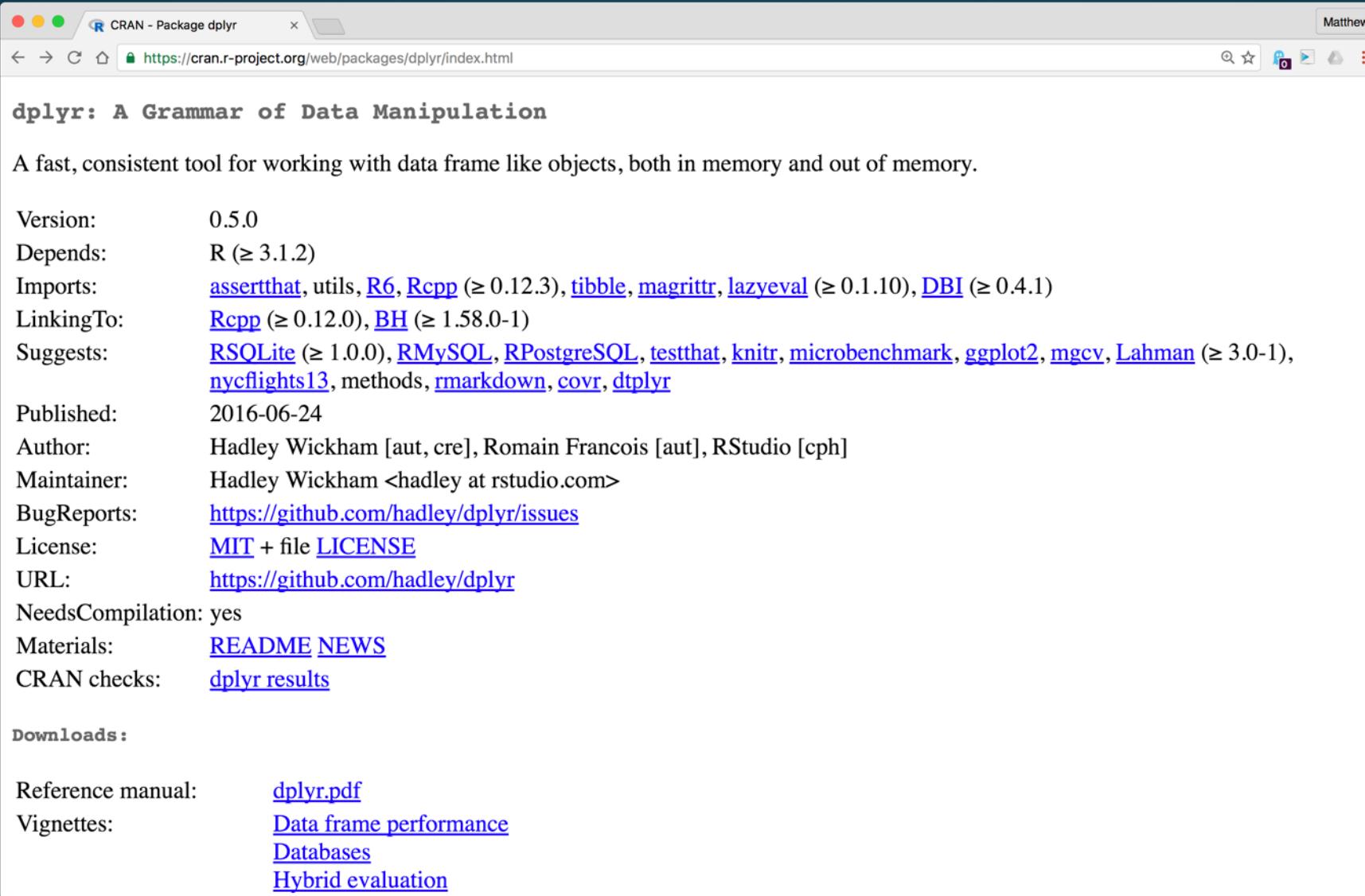
Running R: Servers (Hydra)

```
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/>



The screenshot shows a web browser window with the title "CRAN - Package dplyr". The URL in the address bar is <https://cran.r-project.org/web/packages/dplyr/index.html>. The page content is as follows:

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: [RSSQLite](#) (≥ 1.0.0), [RMySQL](#), [RPostgreSQL](#), [testthat](#), [knitr](#), [microbenchmark](#), [ggplot2](#), [mgcv](#), [Lahman](#) (≥ 3.0-1), [nycflights13](#), methods, [rmarkdown](#), [covr](#), [dplyr](#)
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

23	13	10
----	----	----

```
> 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

Every element is made the
same data type

1	2	3
4	5	6

```
> 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

"one"	2	3
-------	---	---

```
> 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

<u>SampleID</u>	<u>Age</u>	<u>Length</u>	<u>Width</u>
"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

		Data types	
		Same	Multiple
Dimensions	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
```

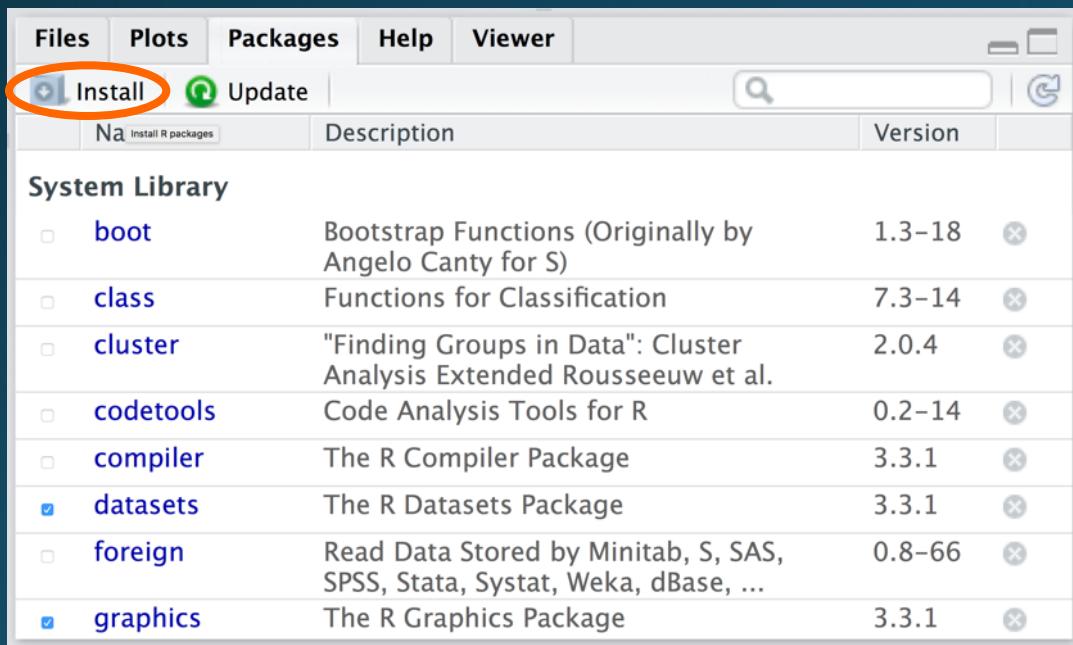
```
> round(6.89, digits=1)  
[1] 6.9
```

Install packages

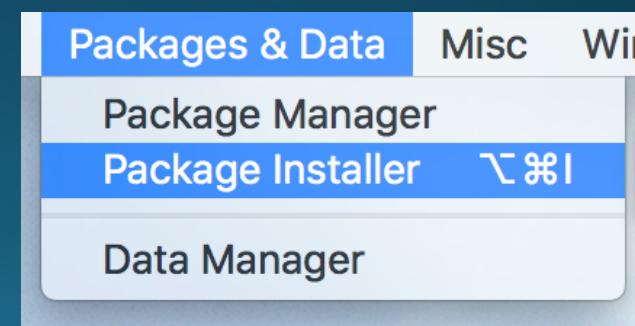
- Command line:

```
> 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

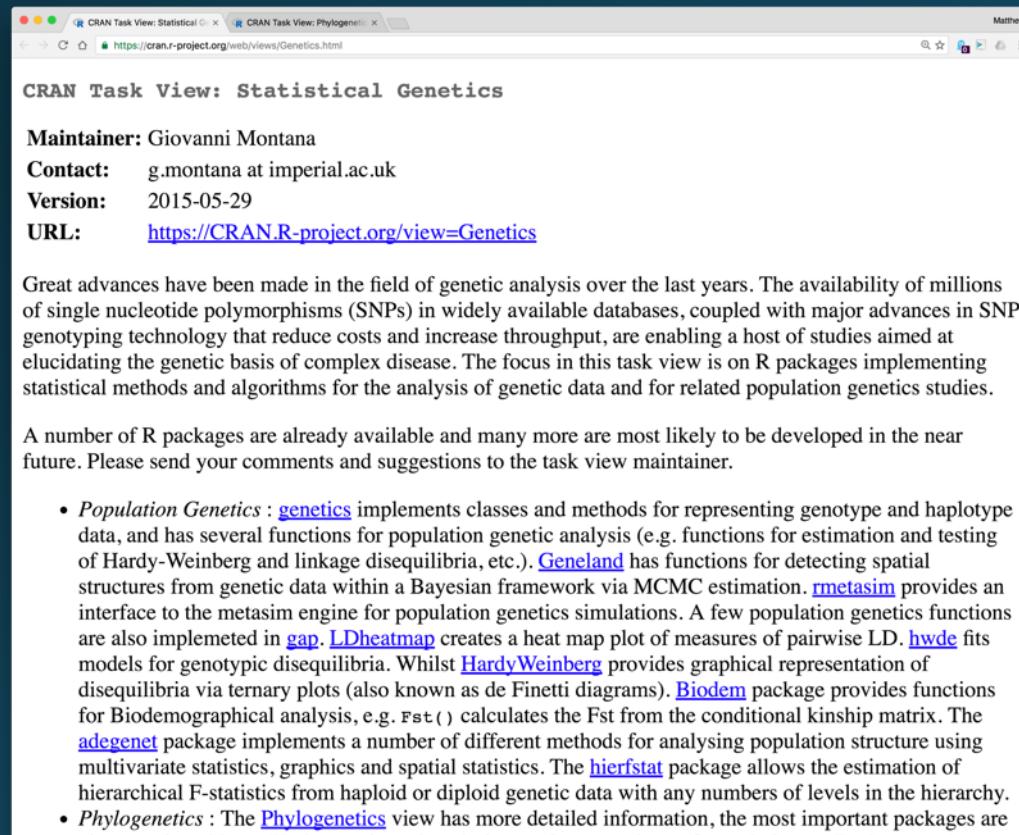
...Would you like to use a personal library instead? (y/n) y

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

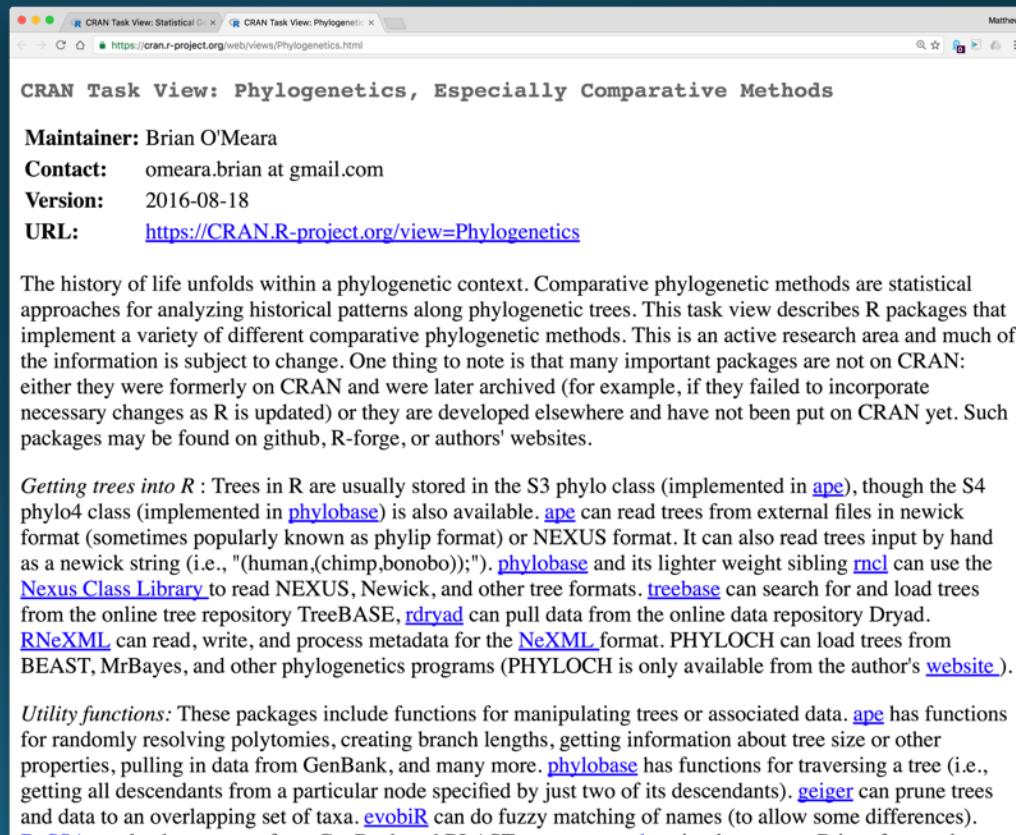
<https://cran.r-project.org/web/views/>

Statistical Genetics



The screenshot shows the CRAN Task View: Statistical Genetics page. It includes contact information for the maintainer (Giovanni Montana), version (2015-05-29), and URL (<https://CRAN.R-project.org/view=Genetics>). A paragraph discusses the rapid advances in genetic analysis and the availability of R packages for population genetics. Below this is a section for population genetics packages, listing `genetics`, `Geneland`, `rmetasim`, `gap`, `LDheatmap`, `hwde`, `HardyWeinberg`, `Biodem`, `adegenet`, `hierfstat`, and `Phylogenetics`. A note at the bottom indicates that the `Phylogenetics` view has more detailed information.

Phylogenetics

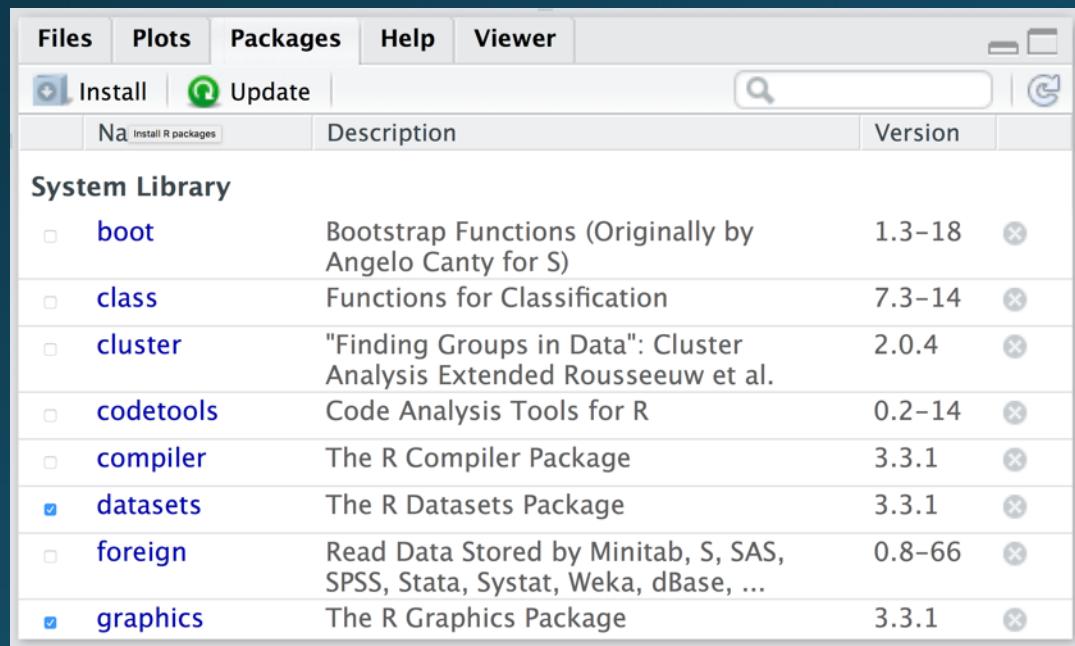


The screenshot shows the CRAN Task View: Phylogenetics, Especially Comparative Methods page. It includes contact information for the maintainer (Brian O'Meara), version (2016-08-18), and URL (<https://CRAN.R-project.org/view=Phylogenetics>). A paragraph describes the history of life within a phylogenetic context and the various comparative phylogenetic methods implemented in R packages. Below this is a detailed section on trees in R, mentioning `ape`, `phylobase`, `ncl`, `treebase`, `dryad`, `RNeXML`, `BEAST`, `MrBayes`, and `PHYLOCH`. Another section covers utility functions for manipulating trees and associated data, including `ape`, `phylobase`, `geiger`, `evobiR`, and `BiSSA`.

"ctv" package will install groups of packages

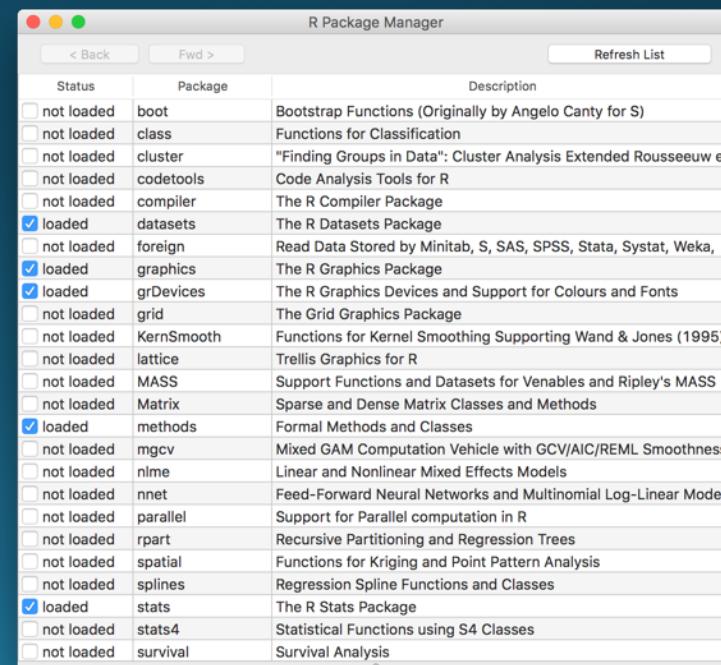
Using a package

- Command line:
> **library("ggplot2")**
- GUI:



The screenshot shows the RStudio interface with the "Packages" tab selected. The title bar includes "Files", "Plots", "Packages", "Help", and "Viewer". Below the tabs, there are buttons for "Install" and "Update". A search bar and a refresh icon are also present. The main area is titled "System Library" and displays a table of packages. The columns are "Name", "Description", and "Version". The "Status" column indicates whether each package is loaded or not. The "Status" column contains icons: a grey square for "not loaded" and a blue square with a checkmark for "loaded".

Name	Description	Version	Status
System Library			
boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-18	
class	Functions for Classification	7.3-14	
cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.0.4	
codetools	Code Analysis Tools for R	0.2-14	
compiler	The R Compiler Package	3.3.1	
datasets	The R Datasets Package	3.3.1	
foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, ...	0.8-66	
graphics	The R Graphics Package	3.3.1	



The screenshot shows the "R Package Manager" window. The title bar includes "Back", "Fwd >", and "Refresh List". The main area displays a table of packages. The columns are "Status", "Package", and "Description". The "Status" column contains icons: a grey square for "not loaded" and a blue square with a checkmark for "loaded".

Status	Package	Description
	boot	Bootstrap Functions (Originally by Angelo Canty for S)
	class	Functions for Classification
	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.
	codetools	Code Analysis Tools for R
	compiler	The R Compiler Package
	datasets	The R Datasets Package
	foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, ...
	graphics	The R Graphics Package
	grDevices	The R Graphics Devices and Support for Colours and Fonts
	grid	The Grid Graphics Package
	KernSmooth	Functions for Kernel Smoothing Supporting Wand & Jones (1995)
	lattice	Trellis Graphics for R
	MASS	Support Functions and Datasets for Venables and Ripley's MASS
	Matrix	Sparse and Dense Matrix Classes and Methods
	methods	Formal Methods and Classes
	mgcv	Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness
	nlme	Linear and Nonlinear Mixed Effects Models
	nnet	Feed-Forward Neural Networks and Multinomial Log-Linear Models
	parallel	Support for Parallel computation in R
	rpart	Recursive Partitioning and Regression Trees
	spatial	Functions for Kriging and Point Pattern Analysis
	splines	Regression Spline Functions and Classes
	stats	The R Stats Package
	stats4	Statistical Functions using S4 Classes
	survival	Survival Analysis