

Getting Started in R for Biologists

Marguerite Butler

University of Hawaii

the R environment

An integrated suite of software facilities:

A fancy calculator

Data Management Handling and storage

Matrix Math: Manipulating matrices, vectors,
and arrays

Statistics: A large, integrated set of tools for
data analysis

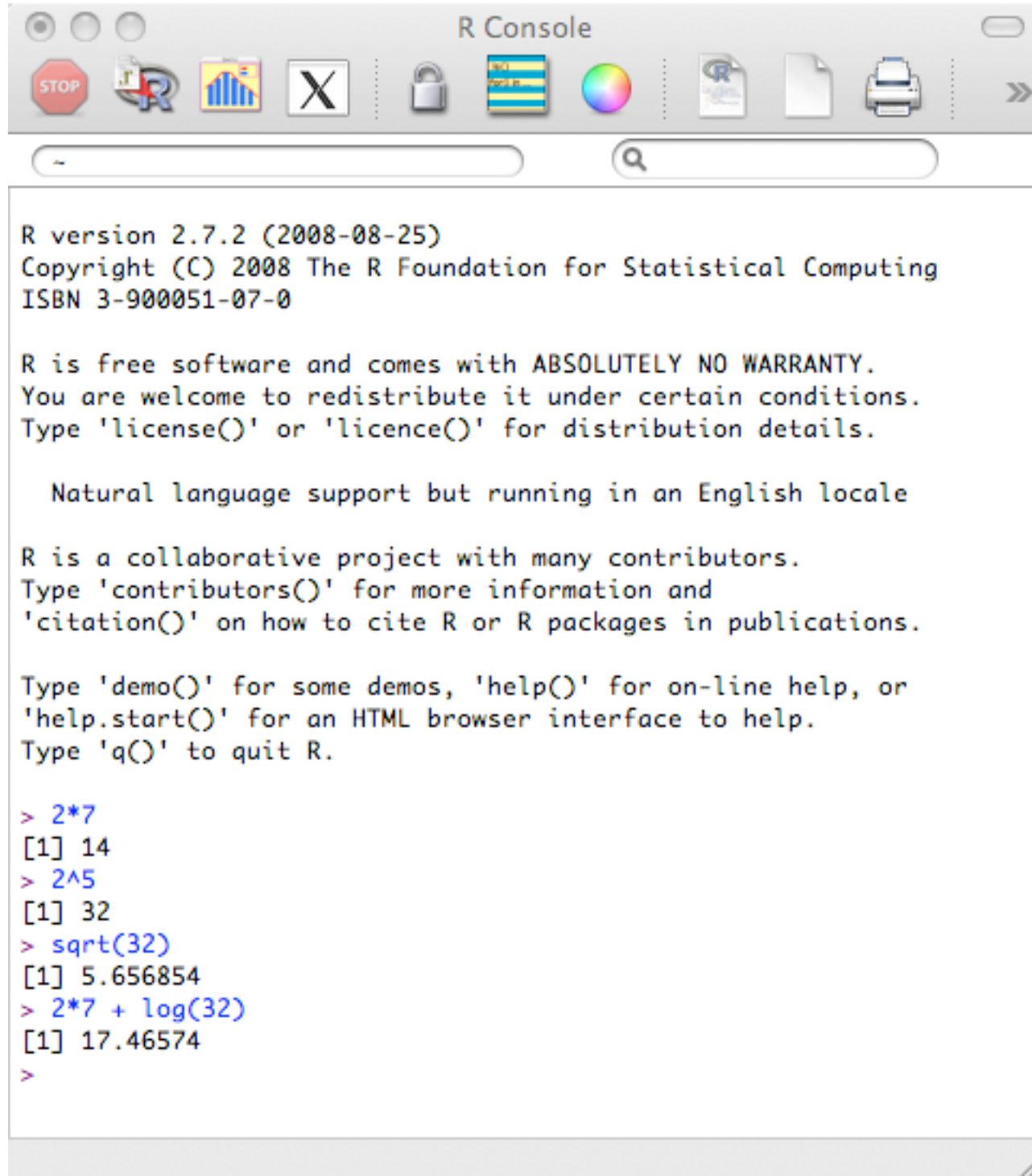
Graphics: Graphical facilities for data analysis
and display

Programming: Powerful programming
language ('S')

Open-Source Development Platform

the R environment

A fancy calculator



```
R Console
R version 2.7.2 (2008-08-25)
Copyright (C) 2008 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

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.

> 2*7
[1] 14
> 2^5
[1] 32
> sqrt(32)
[1] 5.656854
> 2*7 + log(32)
[1] 17.46574
>
```

the R environment

an integrated suite of software facilities:

Data Handling and Storage

```
> morph <- data.frame(species=LETTERS[1:5], size = rnorm(5, mean=15))
> morph
  species    size
1      A 13.38846
2      B 14.83139
3      C 16.68702
4      D 12.42916
5      E 17.32852

> eco <- data.frame(species=LETTERS[5:1], ecology = sample(c("a", "b", "c"), 5,
replace=TRUE))
> eco
  species ecology
1      E      a
2      D      b
3      C      a
4      B      c
5      A      b

> merge(morph, eco)
  species    size ecology
1      A 13.38846      b
2      B 14.83139      c
3      C 16.68702      a
4      D 12.42916      b
5      E 17.32852      a
```

R has facilities for basic database functions:
merging, matching, string matching, file access

the R environment

an integrated suite of software facilities:

Data Handling and Storage

Matrix Math

```
> x <- matrix( data= 1:6, nrow=2)
> x
      [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> t(x)
      [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
> y <- matrix( data = rnorm(9), nrow=3)
> y
      [,1]      [,2]      [,3]
[1,] 0.07969564 -0.04395246 -0.11727169
[2,] -0.01708504 -0.15159683  0.13944474
[3,] 0.56229980  0.25573414 -0.05902727
> x %*% y
      [,1]      [,2]      [,3]
[1,] 2.839940 0.7799277  0.005926158
[2,] 3.464850 0.8401126 -0.030928068
> solve(y)
      [,1]      [,2]      [,3]
[1,] 1.779046  2.170144  1.5922018
[2,] -5.154919 -4.078422  0.6066952
[3,] -5.386179  3.003348  0.8546460
```

matrix

matrix transpose

matrix multiplication

matrix inverse

the R environment

an integrated suite of software facilities:

Data Handling and Storage

Matrix Math

Statistics

Linear Models

ANOVA

Non-parametric Statistics

Multivariate Statistics

Time Series

Numerical Methods

Optimization

etc. etc.

the R environment

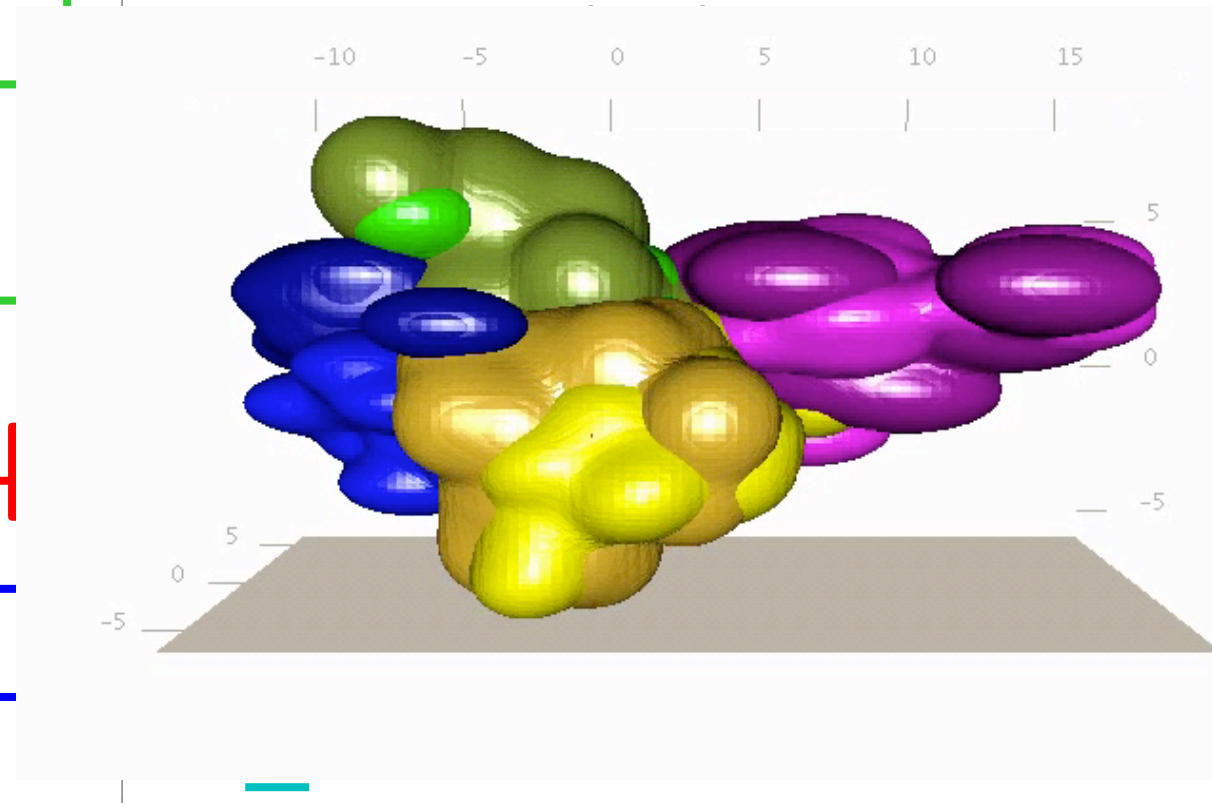
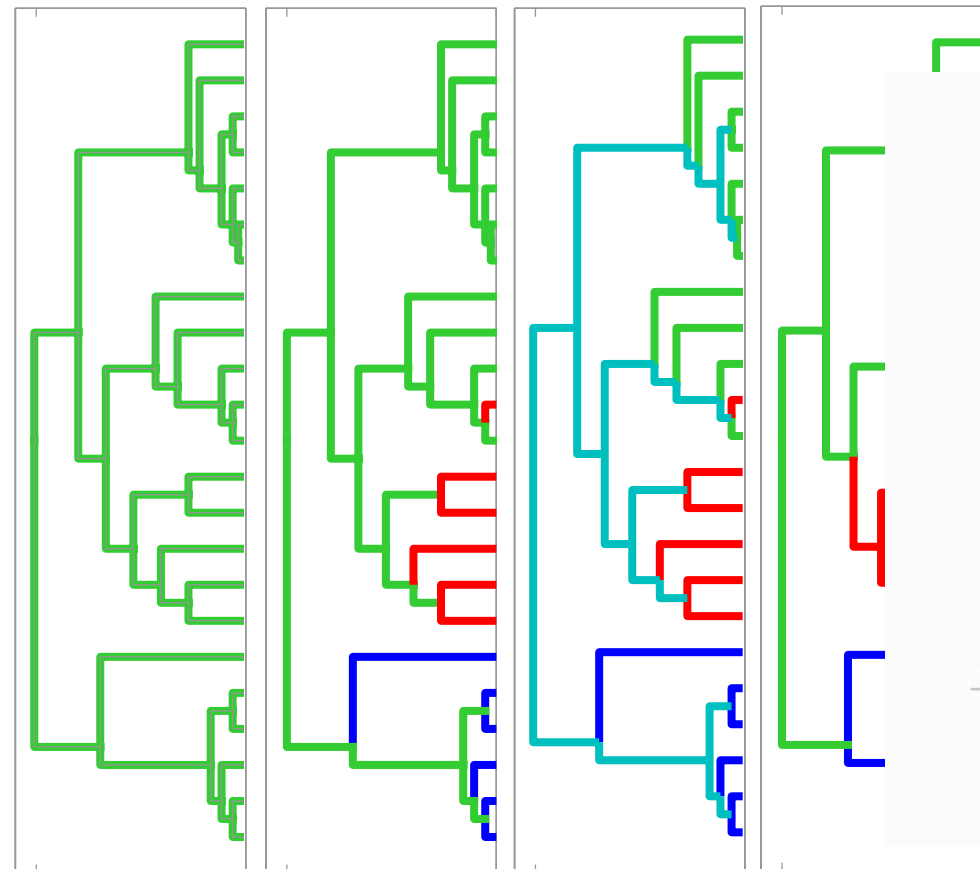
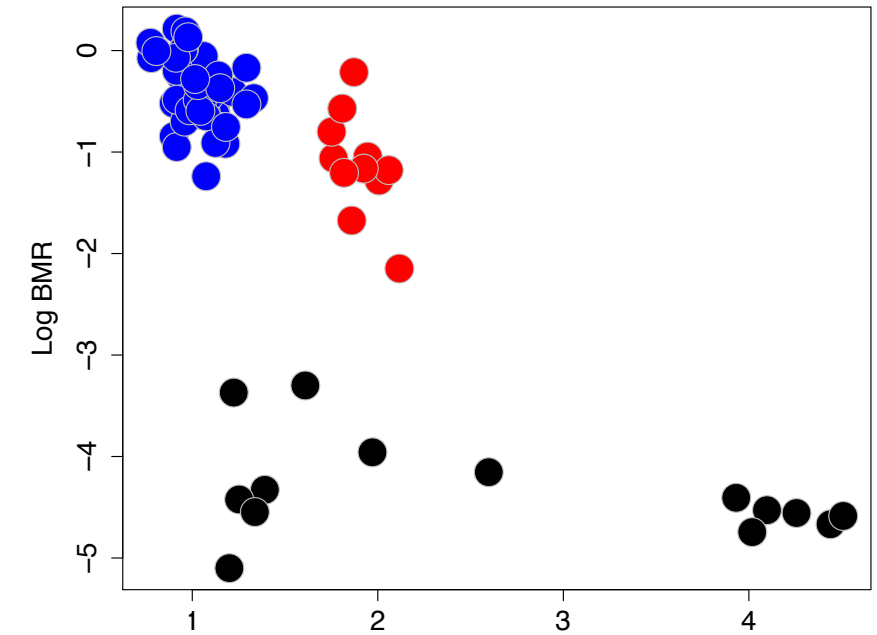
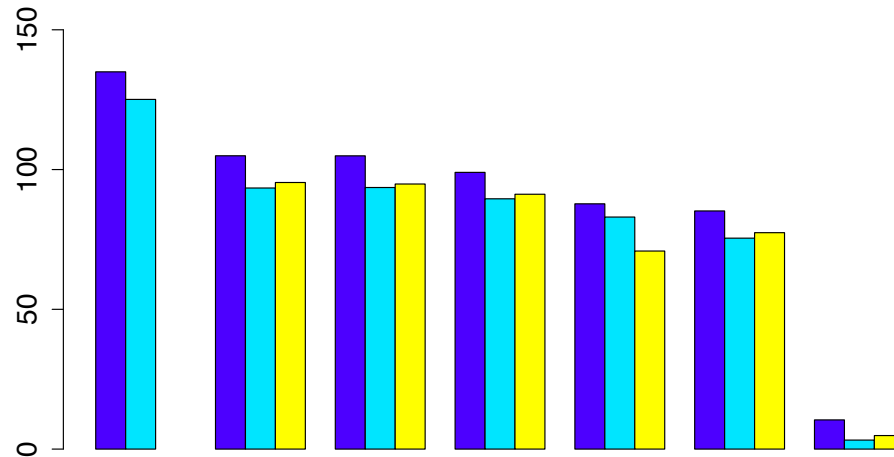
an integrated suite of software facilities:

Data Handling

Matrix Math

Statistics

Graphics



the R environment

an integrated suite of software facilities:

Data Handling and Storage

(semi) Object-Oriented Design

Matrix Math

Conditional Expressions

(Loops)

Statistics

(Recursion)

Vectorized Calculations

Graphics

Functions

Packages

Programming

Extensibility

the R environment

an integrated suite of software facilities:

Data Handling and Storage

Matrix Math

Statistics

Graphics

Programming

Open-Source Community

the R environment

an integrated suite of software facilities:

R Homepage

Data Handling and Storage

Matrix Math

Statistics

Graphics

Programming

Open-Source Community

The R Project for Statistical Computing

http://www.r-project.org/

Most Visited ▾ LaTeX/BibTeX ▾ New Home ▾ Rhackathon ▾ Dog Health ▾ Latest Headlines ▾ Grants ▾ UH Websites ▾ Computer Prices ▾

The R Project for Statistical Computing

PCA 5 vars
princomp(x = data, cor = cor)

Clustering 4 groups

Factor 1 [41%]

Factor 3 [19%]

Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred [CRAN mirror](#).
- 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.

News:

- [R version 2.8.1](#) has been released on 2008-12-22.
- [R News 8/2](#) has been published on 2008-11-03.
- [DSC 2009](#), The 6th workshop on Directions in Statistical Computing, will be held at the Center for Health and Society, University of Copenhagen, Denmark, July 13-14, 2009.
- [useR! 2009](#), the R user conference, will be held at Agrocampus Rennes, France, July 8-10, 2009.
- [useR! 2008](#), has been held at Dortmund University, Germany, August 12-14, 2008.

This server is hosted by the [Department of Statistics and Mathematics](#) of the [WU Wien](#).

Done

the R environment

CRAN (Comprehensive R Archive Network)

an integrated suite

Data Handling and Storage

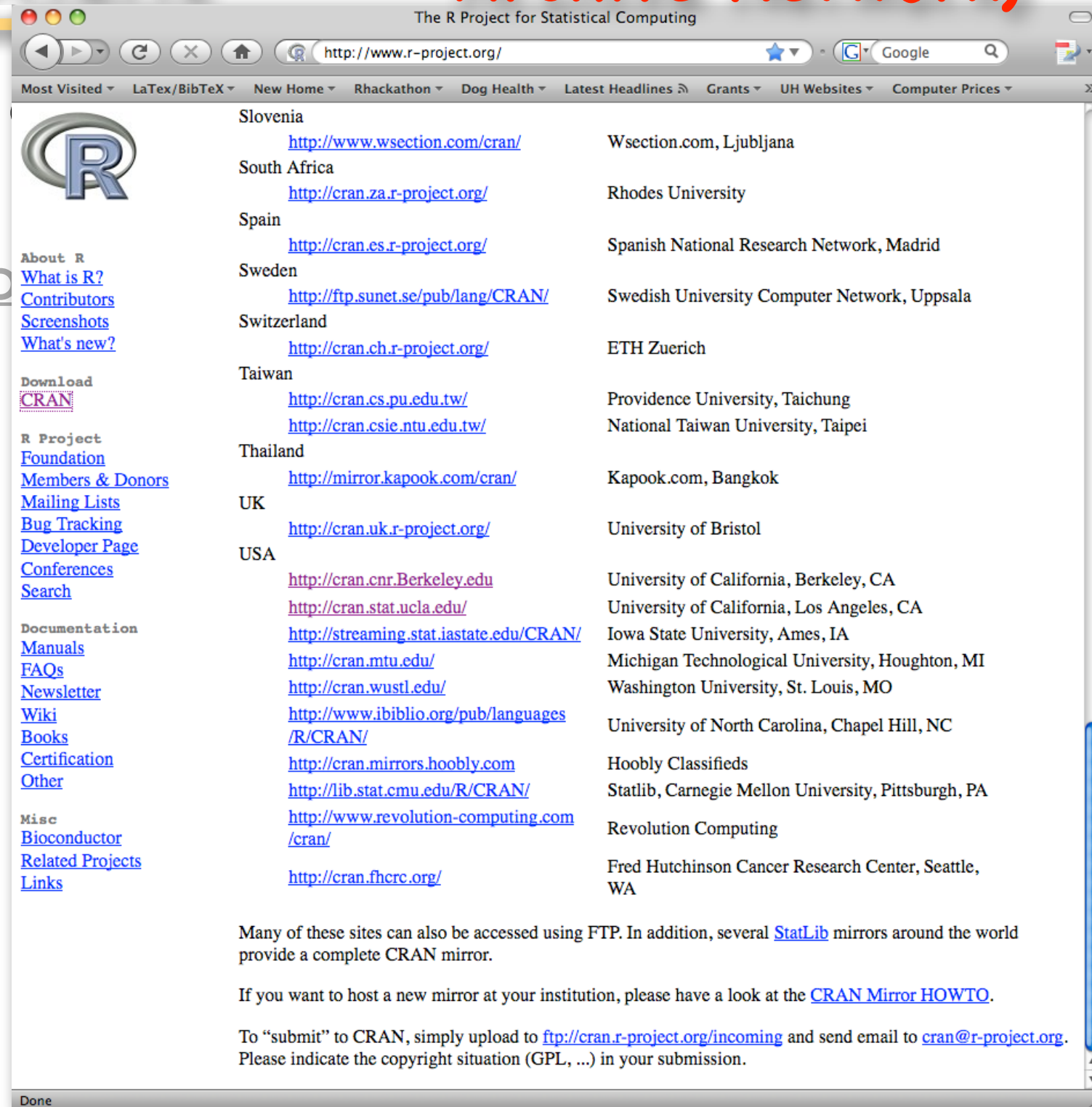
Matrix Math

Statistics

Graphics

Programming


Open-Source Community



The R Project for Statistical Computing

http://www.r-project.org/

Most Visited ▾ LaTeX/BibTeX ▾ New Home ▾ Rhackathon ▾ Dog Health ▾ Latest Headlines ▾ Grants ▾ UH Websites ▾ Computer Prices ▾



About R
[What is R?](#)
[Contributors](#)
[Screenshots](#)
[What's new?](#)

Download
[CRAN](#)

R Project
[Foundation](#)
[Members & Donors](#)
[Mailing Lists](#)
[Bug Tracking](#)
[Developer Page](#)
[Conferences](#)
[Search](#)

Documentation
[Manuals](#)
[FAQs](#)
[Newsletter](#)
[Wiki](#)
[Books](#)
[Certification](#)
[Other](#)

Misc
[Bioconductor](#)
[Related Projects](#)
[Links](#)

Slovenia	http://www.wsection.com/cran/	Wsection.com, Ljubljana
South Africa	http://cran.za.r-project.org/	Rhodes University
Spain	http://cran.es.r-project.org/	Spanish National Research Network, Madrid
Sweden	http://ftp.sunet.se/pub/lang/CRAN/	Swedish University Computer Network, Uppsala
Switzerland	http://cran.ch.r-project.org/	ETH Zuerich
Taiwan	http://cran.cs.pu.edu.tw/ http://cran.csie.ntu.edu.tw/	Providence University, Taichung National Taiwan University, Taipei
Thailand	http://mirror.kapook.com/cran/	Kapook.com, Bangkok
UK	http://cran.uk.r-project.org/	University of Bristol
USA	http://cran.cnr.Berkeley.edu http://cran.stat.ucla.edu/ http://streaming.stat.iastate.edu/CRAN/ http://cran.mtu.edu/ http://cran.wustl.edu/ http://www.ibiblio.org/pub/languages/R/CRAN/ http://cran.mirrors.hoobly.com http://lib.stat.cmu.edu/R/CRAN/ http://www.revolution-computing.com/cran/ http://cran.fhrc.org/	University of California, Berkeley, CA University of California, Los Angeles, CA Iowa State University, Ames, IA Michigan Technological University, Houghton, MI Washington University, St. Louis, MO University of North Carolina, Chapel Hill, NC Hoobly Classifieds Statlib, Carnegie Mellon University, Pittsburgh, PA Revolution Computing Fred Hutchinson Cancer Research Center, Seattle, WA

Many of these sites can also be accessed using FTP. In addition, several [StatLib](#) mirrors around the world provide a complete CRAN mirror.

If you want to host a new mirror at your institution, please have a look at the [CRAN Mirror HOWTO](#).

To "submit" to CRAN, simply upload to <ftp://cran.r-project.org/incoming> and send email to cran@r-project.org. Please indicate the copyright situation (GPL, ...) in your submission.

Done

the R environment Support for all Major Platforms

an integrated suite

Data Handling and

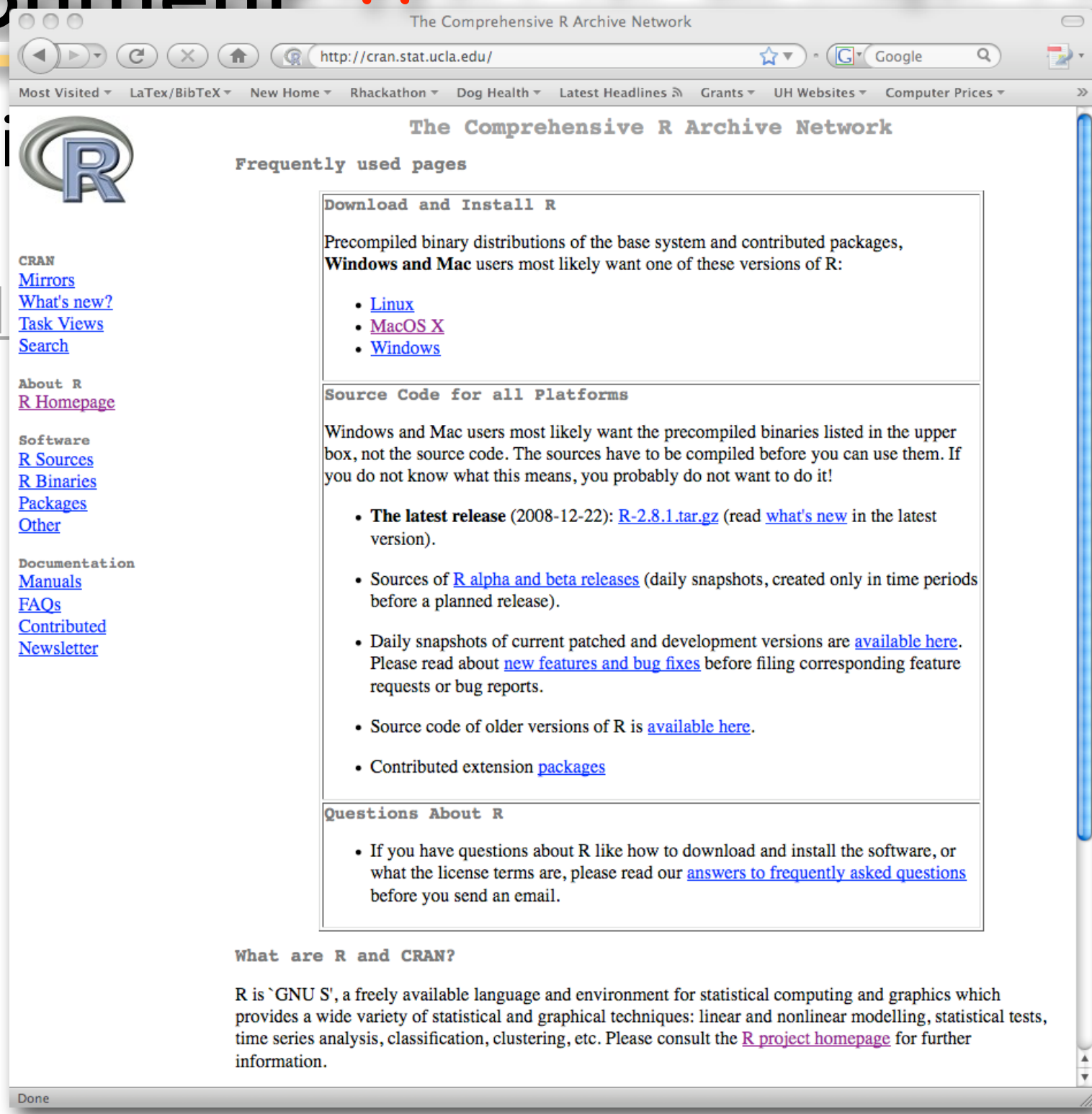
Matrix Math

Statistics

Graphics

Programming

Open-Source Community



the R environment

an integrated suite

Data Handling and Storage

Matrix Math

Statistics

Graphics

Programming

Open-Source Community

Contributed Packages

The screenshot shows the CRAN website at <http://cran.stat.ucla.edu/>. The page title is "Contributed Packages". The main content includes:

- Installation of Packages**: Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages from this directory. The manual [R Installation and Administration](#) (also contained in the R base sources) explains the process in detail.
- CRAN Task Views**: allow you to browse packages by topic and provide tools to automatically install all packages for special areas of interest. Currently, 22 views are available.
- Daily Package Check Results**: All packages are tested regularly on machines running [Debian GNU/Linux](#). Packages are also checked under MacOS X and Windows, but only at the day the package appears on CRAN. The results are summarized in the [check summary](#) (some [timings](#) are also available). Additional details for Windows checking and building can be found in the [Windows check summary](#).
- Writing Your Own Packages**: The manual [Writing R Extensions](#) (also contained in the R base sources) explains how to write new packages and how to contribute them to CRAN.
- Available Bundles and Packages**: Currently, the CRAN package repository features 1630 objects including 1622 packages and 8 bundles containing 34 packages, for a total of 1656 available packages.

Below this, there is a list of packages with their descriptions:

Package Name	Description
ADaCGH	Analysis of data from aCGH experiments
AER	Applied Econometrics with R
AIS	Tools to look at the data ("Ad Inidicia Spectata")
ALS	multivariate curve resolution alternating least squares (MCR-ALS)
AMORE	A MORE flexible neural network package
ARES	Allelic richness estimation, with extrapolation beyond the sample size
AcceptanceSampling	Creation and evaluation of Acceptance Sampling Plans
AdMit	Adaptive Mixture of Student-t distributions
AdaptFit	Adaptive Semiparametric Regression
AlgDesign	AlgDesign
Amelia	Amelia II: A Program for Missing Data
AnalyzefMRI	Functions for analysis of fMRI datasets stored in the ANALYZE or NIFTI format

the R environment

CRAN Package "home pages"

an integrated suite

Data Handling and S

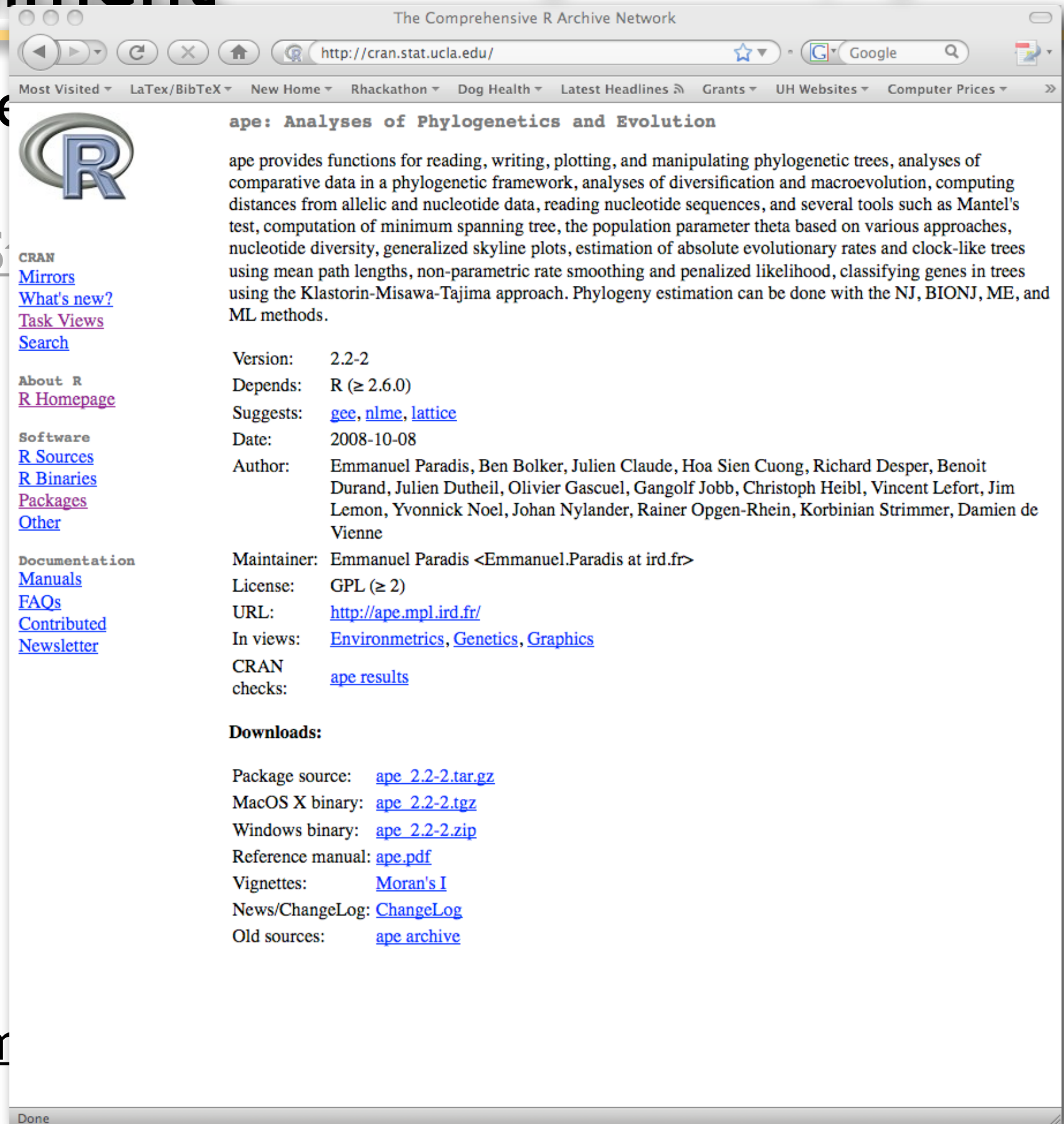
Matrix Math

Statistics

Graphics

Programming

Open-Source Comm



The screenshot shows a web browser window titled "The Comprehensive R Archive Network" with the URL "http://cran.stat.ucla.edu/". The browser's address bar and search engine (Google) are visible. The page content is for the CRAN package "ape: Analyses of Phylogenetics and Evolution".

ape: Analyses of Phylogenetics and Evolution

ape provides functions for reading, writing, plotting, and manipulating phylogenetic trees, analyses of comparative data in a phylogenetic framework, analyses of diversification and macroevolution, computing distances from allelic and nucleotide data, reading nucleotide sequences, and several tools such as Mantel's test, computation of minimum spanning tree, the population parameter theta based on various approaches, nucleotide diversity, generalized skyline plots, estimation of absolute evolutionary rates and clock-like trees using mean path lengths, non-parametric rate smoothing and penalized likelihood, classifying genes in trees using the Klastorin-Misawa-Tajima approach. Phylogeny estimation can be done with the NJ, BIONJ, ME, and ML methods.

Version: 2.2-2
Depends: R (≥ 2.6.0)
Suggests: [gee](#), [nlme](#), [lattice](#)
Date: 2008-10-08
Author: Emmanuel Paradis, Ben Bolker, Julien Claude, Hoa Sien Cuong, Richard Desper, Benoit Durand, Julien Dutheil, Olivier Gascuel, Gangolf Jobb, Christoph Heibl, Vincent Lefort, Jim Lemon, Yvonnick Noel, Johan Nylander, Rainer Opgen-Rhein, Korbinian Strimmer, Damien de Vienne
Maintainer: Emmanuel Paradis <Emmanuel.Paradis at ird.fr>
License: GPL (≥ 2)
URL: <http://ape.mpl.ird.fr/>
In views: [Environmetrics](#), [Genetics](#), [Graphics](#)
CRAN checks: [ape results](#)

Downloads:

Package source: [ape 2.2-2.tar.gz](#)
MacOS X binary: [ape 2.2-2.tgz](#)
Windows binary: [ape 2.2-2.zip](#)
Reference manual: [ape.pdf](#)
Vignettes: [Moran's I](#)
News/ChangeLog: [ChangeLog](#)
Old sources: [ape archive](#)

On the left side of the browser window, there is a sidebar with the R logo and several navigation links:

- CRAN
- [Mirrors](#)
- [What's new?](#)
- [Task Views](#)
- [Search](#)
- About R
- [R Homepage](#)
- Software
- [R Sources](#)
- [R Binaries](#)
- [Packages](#)
- [Other](#)
- Documentation
- [Manuals](#)
- [FAQs](#)
- [Contributed](#)
- [Newsletter](#)

the R environment

Task Views

an integrated suite of s

Data Handling and Storage

Matrix Math

Statistics

Graphics

Programming

Open-Source Community

The Comprehensive R Archive Network

http://cran.stat.ucla.edu/

CRAN Task Views

[Bayesian](#) Bayesian Inference

[ChemPhys](#) Chemometrics and Computational Physics

[Cluster](#) Cluster Analysis & Finite Mixture Models

[Distributions](#) Probability Distributions

[Econometrics](#) Computational Econometrics

[Environmetrics](#) Analysis of Ecological and Environmental Data

[ExperimentalDesign](#) Design of Experiments (DoE) & Analysis of Experimental Data

[Finance](#) Empirical Finance

[Genetics](#) Statistical Genetics

[Graphics](#) Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization

[gR](#) gRaphical Models in R

[MachineLearning](#) Machine Learning & Statistical Learning

[Multivariate](#) Multivariate Statistics

[NaturalLanguageProcessing](#) Natural Language Processing

[Optimization](#) Optimization and Mathematical Programming

[Pharmacokinetics](#) Analysis of Pharmacokinetic Data

[Psychometrics](#) Psychometric Models and Methods

[Robust](#) Robust Statistical Methods

[SocialSciences](#) Statistics for the Social Sciences

[Spatial](#) Analysis of Spatial Data

[Survival](#) Survival Analysis

[TimeSeries](#) Time Series Analysis

To automatically install these views, the ctv package needs to be installed, e.g., via

```
install.packages("ctv")
library("ctv")
and then the views can be installed via install.views or update.views (which first assesses which of the
packages are already installed and up-to-date), e.g.,
install.views("Econometrics")
or
update.views("Econometrics")
```


the R environment

an integrated suite of software

Data Handling and Storage

Matrix Math

Statistics

Graphics

Programming

Open-Source Community

Task Views

The screenshot shows a web browser window displaying the CRAN Task View for 'Statistical Genetics'. The browser's address bar shows 'http://cran.stat.ucla.edu/'. The page title is 'CRAN Task View: Statistical Genetics'. The maintainer is Giovanni Montana, with contact 'g.montana at imperial.ac.uk' and version '2008-12-08'. The main text describes the availability of millions of SNPs and the focus on R packages for genetic analysis. A list of R packages is provided, including `genetics`, `Geneland`, `rmetasim`, `hapsim`, `gap`, `popgen`, `hierfstat`, `LDheatmap`, `mapLD`, `hwde`, `HardyWeinberg`, `Biodem`, `kinship`, `ape`, `apTreeshape`, `ouch`, `PHYLOGR`, `stepwise`, `phangor`, `ibdreg`, and `multic`.

The Comprehensive R Archive Network

http://cran.stat.ucla.edu/

Most Visited ▾ LaTeX/BibTeX ▾ New Home ▾ Rhackathon ▾ Dog Health ▾ Latest Headlines ▾ Grants ▾ UH Websites ▾ Computer Prices ▾

CRAN Task View: Statistical Genetics

Maintainer: Giovanni Montana
Contact: g.montana at imperial.ac.uk
Version: 2008-12-08

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.

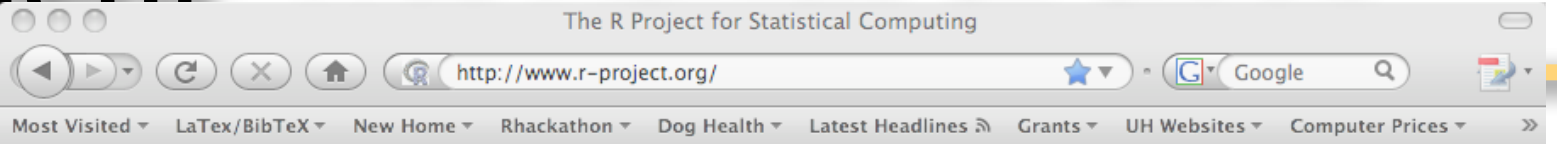
- *Population Genetics* : `genetics` implements classes and methods for representing genotype and haplotype data, and has several functions for population genetic analysis (e.g. functions for estimation and testing of Hardy-Weinberg and linkage disequilibria, etc.). `Geneland` has functions for detecting spatial structures from genetic data within a Bayesian framework via MCMC estimation. `rmetasim` provides an interface to the metasim engine for population genetics simulations. `hapsim` simulates haplotype data with pre-specified allele frequencies and LD patterns. A few population genetics functions are also implemented in `gap` and `popgen`. `popgen` has functions for clustering SNP genotype data and SNP simulation from a Multinomial-Dirichlet model. `hierfstat` allows the estimation of hierarchical F-statistics from haploid or diploid genetic data. `LDheatmap` creates a heat map plot of measures of pairwise LD. `mapLD` measures linkage disequilibrium and constructs haplotype blocks. `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. Package `kinship` offers some functions for analysis on pedigrees. The `adegenet` implements a number of different methods for analysing population structure using multivariate statistics, graphics and spatial statistics.
- *Phylogenetics* : Phylogenetic and evolution analyses can be performed via `ape` and `apTreeshape`. Package `ouch` provides Ornstein-Uhlenbeck models for phylogenetic comparative hypotheses. `PHYLOGR` is a suite of functions for the analysis of phylogenetically simulated data sets and phylogenetically-based GLS model fitting. `stepwise` implements a method for stepwise detection of recombination breakpoints in sequence alignments. `phangor` estimates phylogenetic trees and networks using maximum likelihood, maximum parsimony, distance methods and Hadamard conjugation.
- *Linkage* : There are few native packages for performing parametric or non-parametric linkage analysis from within R itself, the calculations must be performed using external packages. However, there are a number of ancillary R packages that facilitate interface with these stand-alone programs and using the results for further analysis and presentation. `ibdreg` uses Identity By Descent (IBD) Non-Parametric Linkage (NPL) statistics for related pairs calculated externally to test for genetic linkage with covariates by regression modelling. `multic` also utilises IBD sharing statistics calculated externally for

Done

the R environment

Free Manuals on CRAN!

an integrated suite of



Data Handling and Storage

Matrix Math

Statistics

Graphics

Programming

Open-Source Community

- About R
 - [What is R?](#)
 - [Contributors](#)
 - [Screenshots](#)
 - [What's new?](#)
- Download
 - [CRAN](#)
- R Project
 - [Foundation](#)
 - [Members & Donors](#)
 - [Mailing Lists](#)
 - [Bug Tracking](#)
 - [Developer Page](#)
 - [Conferences](#)
 - [Search](#)
- Documentation
 - [Manuals](#)
 - [FAQs](#)
 - [Newsletter](#)
 - [Wiki](#)
 - [Books](#)
 - [Certification](#)
 - [Other](#)
- Misc
 - [Bioconductor](#)
 - [Related Projects](#)
 - [Links](#)



The R Manuals

edited by the R Development Core Team.

Current Version: 2.8.1 (December 2008)

The following manuals for R were created on Debian Linux and may differ from the manuals for Mac or Windows on platform-specific pages, but most parts will be identical for all platforms. The correct version of the manuals for each platform are part of the respective R installations. Here they can be downloaded as PDF files or directly browsed as HTML:

- **An Introduction to R** is based on the former "Notes on R", gives an introduction to the language and how to use R for doing statistical analysis and graphics. [[browse HTML](#) | [download PDF](#)]
- A draft of **The R language definition** documents the language *per se*. That is, the objects that it works on, and the details of the expression evaluation process, which are useful to know when programming R functions. [[browse HTML](#) | [download PDF](#)]
- **Writing R Extensions** covers how to create your own packages, write R help files, and the foreign language (C, C++, Fortran, ...) interfaces. [[browse HTML](#) | [download PDF](#)]
- **R Data Import/Export** describes the import and export facilities available either in R itself or via packages which are available from CRAN. [[browse HTML](#) | [download PDF](#)]
- **R Installation and Administration** [[browse HTML](#) | [download PDF](#)]
- **R Internals**: a guide to the internal structures of R and coding standards for the core team working on R itself. [[browse HTML](#) | [download PDF](#)]
- **The R Reference Index**: contains all help files of the R standard and recommended packages in printable form. [[download PDF, 14MB](#)]

Translations of manuals into other languages than English are available from the [contributed documentation](#) section (only a few translations are available).

The latex or texinfo sources of the latest version of these documents are contained in every R source distribution (in the subdirectory `doc/manual` of the extracted archive). Older versions of the manual can be found in the respective [archives of the R sources](#). The HTML versions of the manuals are also part of most R installations (accessible using function `help.start()`).

Done

Why Code?

Does your desktop look like this?

Are your results reproducible?

<https://blog.learningtree.com/reproducible-results-data-science/>



Can you customize your analyses and models to your data and needs?

R works in RAM

Computations are fast

Computers today have >1GB RAM

On 32bit systems the max size of all data objects in an R workspace is about 2GB

On 64bit systems the max size is limited by total RAM

*** All meaningful analyses are saved as scripts which are run in RAM —> A big departure from “point and click” (one-off) analyses ***

Let's Play with  !

Assignment: Basic rules of syntax

R can **save** information in variables or objects

Assignment works by two types of operators:

Equal sign: right side stored in left side

> `x = 6` (put 6 into x)

Arrow: assignment direction follows arrow

> `x <- 6` (put 6 into x)

> `6 -> x` (put 6 into x)

> `6 = x` (error! cannot put x into 6)

Example: `rnorm(x)`

Let's create a script that illustrates the central limit theorem

Common Sources of Error

1) Typos! Computers are very anal that way.

- > `length = 6` # is not the same as
- > `lengths = 6`

2) R is case sensitive

- > `length != Length`

3) Using () when should use [] and vice versa

- > `mean(x)` # use () for functions
- > `mean[x]` # error
- > `x[5]` # select an element of a vector, matrix, data.frame, etc.
- > `x(5)` # error

4) No comma or comma in the wrong place

- > `x[5,3]` # fifth row, third column of x
- > `x[5 3]` # error
- > `x[5,3,]` # error

Common Sources of Error

5) Forgetting quotes for character strings (R will assume it's another named object or variable)

```
> treatment = c("a", "b", "c")
```

```
> treatment == a      # error - R thinks a is another object
```