

Basic of R language

Dr. S. DEVAARUL

Learning aims

- Basic use of R and R help
- How to give R commands
- R data structures
- Reading and writing data
- Some more R commands (exercises)

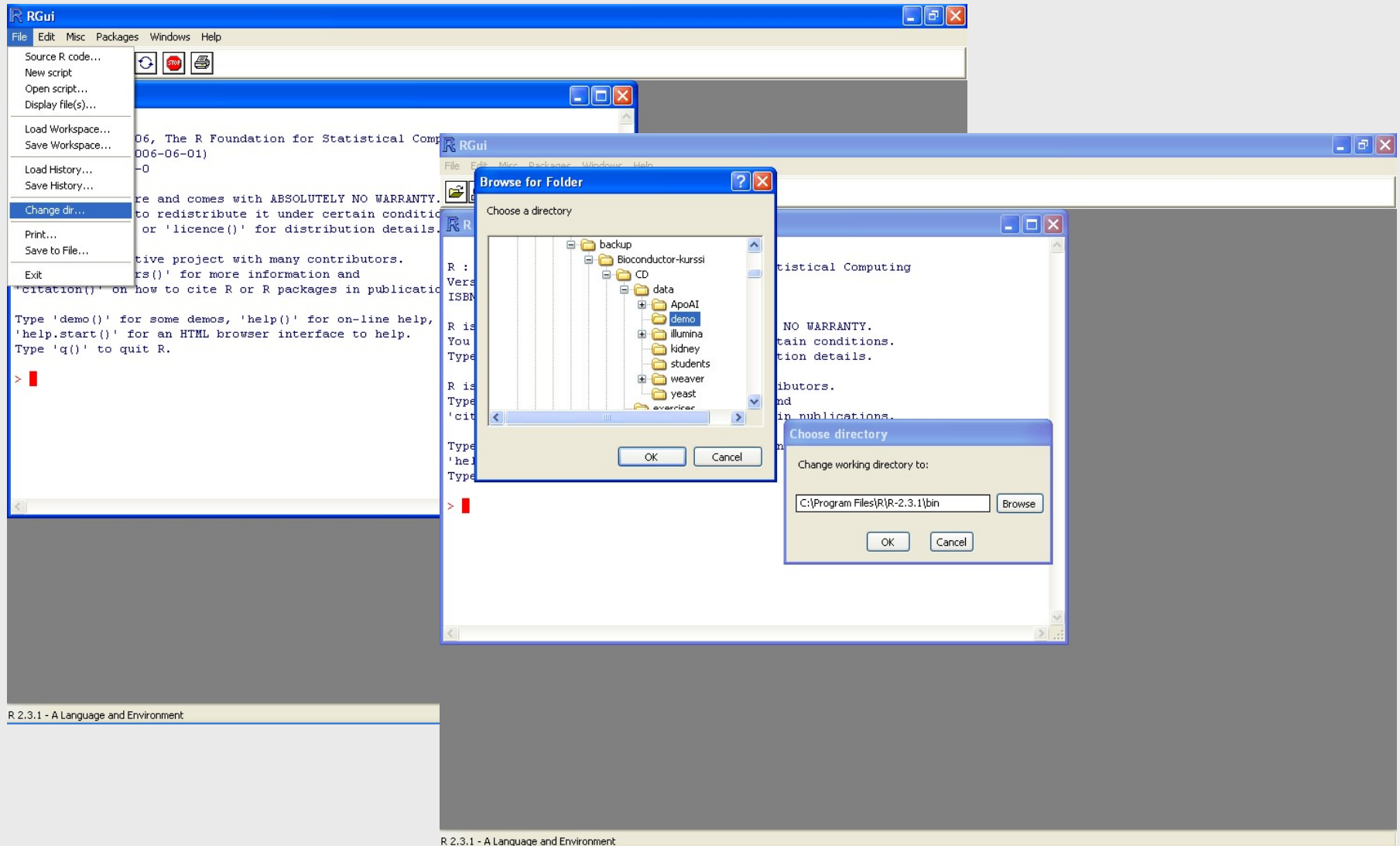
R project

- "R is a free software environment for statistical computing and graphics"
(<http://www.r-project.org>)
- "Bioconductor is a software project for the analysis of genomic data"
(<http://www.bioconductor.org>)
 - Currently works as an expansion to R

Packages

- R consists of a core and packages.
- Packages contain functions that are not available in the core.
- For example, Bioconductor code is distributed as several dozen of packages for R.
 - Software packages
 - Metadata (annotation) packages

Starting the work with R



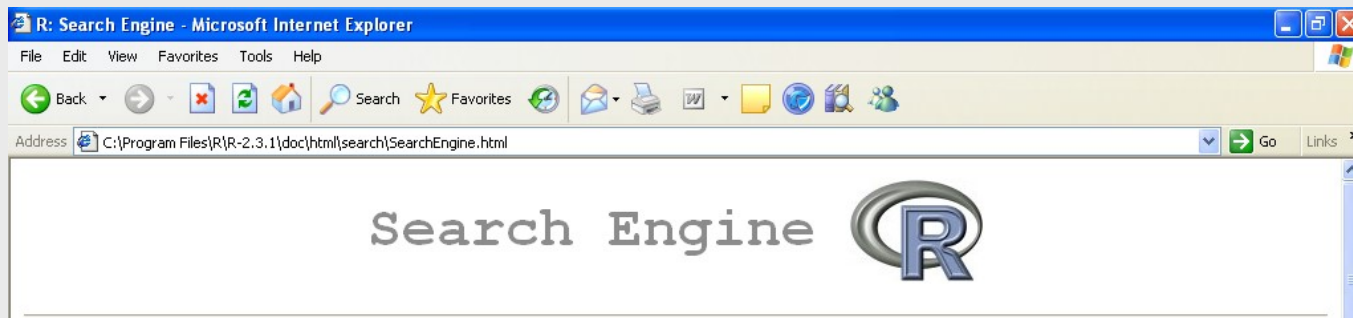
Start help

The image shows a screenshot of the RGui interface and a Microsoft Internet Explorer browser window. The RGui window is in the foreground, showing the R Console with the command `> help.start()` entered. The browser window is displaying the R help page, which is titled "Statistical Data Analysis" and features the R logo. The page contains several sections of links:

- Manuals**
 - [An Introduction to R](#)
 - [The R Language Definition](#)
 - [R Installation and Administration](#)
 - [Writing R Extensions](#)
 - [R Data Import/Export](#)
- Reference**
 - [Packages](#)
 - [Search Engine & Keywords](#)
- Miscellaneous Material**
 - [About R](#)
 - [License](#)
 - [Authors](#)
 - [Frequently Asked Questions](#)
 - [FAQ for Windows port](#)
 - [Resources](#)
 - [Thanks](#)

The browser window title is "The R Language - Microsoft Internet Explorer" and the address bar shows the path `C:\Program Files\R\R-2.3.1\doc\html\index.html`. The RGui window title is "RGui" and the menu bar includes "File", "Edit", "Misc", "Packages", "Windows", and "Help". The R Console window title is "R Console" and the status bar at the bottom of the RGui window shows "R 2.3.1 - A Language and Environment".

Help - Search engine



Search

You can search for keywords, function and data names and text in help

Usage: Enter a string in the text field below and hit RETURN.

linear regression Help page titles Keywords Object names

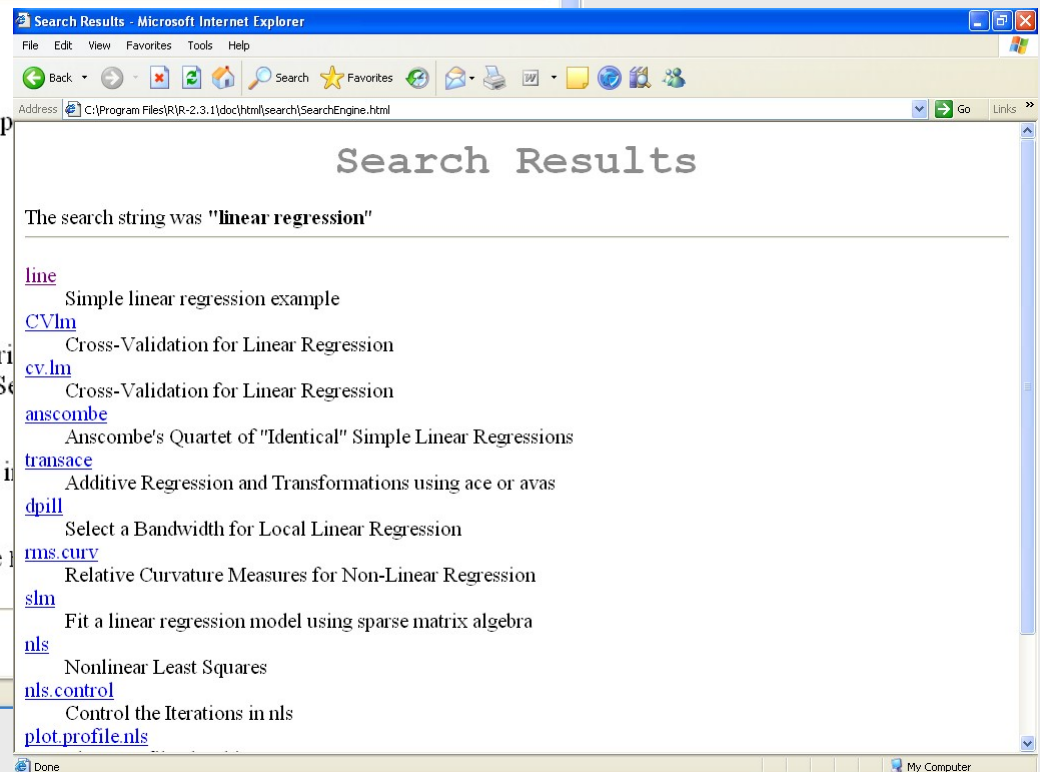
For search to work, you need Java installed and both Java and JavaScript. On the Mozilla/Netscape family of browsers you should see "Applet Security" bar. For help consult the [R Installation and Administration](#) manual.

On Mozilla-based browsers the links on the results page will become images; you can open a link in a new tab or window.

Even if this search does not work on your system, you can always use

Keywords

Applet SearchEngine started



Help - packages

R: Package Index - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Home Search Favorites
Address C:\Program Files\R\R-2.3.1\doc\html/packages.html
Package Index

[abind](#)
[acepack](#)
[affy](#)
[affydata](#)
[affyio](#)
[affyPLM](#)
[annaffy](#)
[annotate](#)
[aroma](#)
[aroma.light](#)

Combine multi-dimensional arrays
ace() and avas() for selecting regre
Methods for Affymetrix Oligonuc
Affymetrix Data for Demonstratio
Tools for parsing Affymetrix data
Methods for fitting probe-level mo
Annotation tools for Affymetrix bi
Annotation for microarrays
An R Object-oriented Microarray
Light-weight methods for normaliz
basic R data types
The R Base Package
Biobase: Base functions for Bioc
Interface to BioMart databases (e.g
String objects representing biologi
Bootstrap R (S-Plus) Functions (C
Companion to Applied Regression

[base](#)
[Biobase](#)
[biomaRt](#)
[Biostrings](#)
[boot](#)
[car](#)

R: Grapics related functions for Bioconductor - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Home Search Favorites
Address C:\Program Files\R\R-2.3.1\library\genepLOTter\html\00Index.html
Grapics related functions for Bioconductor

Documentation for package 'genepLOTter' version 1.10.0
User Guides and Package Vignettes
Read [overview](#) or browse [directory](#).
Help Pages
[alongChrom](#) A function for plotting expression data from an exprset for a given chromosome.
[amplicon.plot](#) Create an amplicon plot
[buildACMainLabel](#) A function for plotting expression data from an exprset for a given chromosome.
[cColor](#) A function for marking specific probes on a cPlot.
[closeHtmlPage](#) Open and close an HTML file for writing.
[connection-class](#) Virtual S4 classes for method dispatching
[cPlot](#) A plotting function for chromosomes.
[cScale](#) A function for mapping chromosome length to a number of points.
[cullACXPoints](#) A function for plotting expression data from an exprset for a given chromosome.

Anatomy of a help file 1/2

The screenshot shows a Microsoft Internet Explorer window displaying the R documentation for the `mas5` function. The address bar shows the file path: `C:\Program Files\R\R-2.3.1\library\affy\html\mas5.html`. The page content is as follows:

mas5 {affy} ← **Function {package}**

MAS 5.0 expression measure

Description

This function converts an instance of `AffyBatch-class` into an instance of `exprSet-class` using our implementation of Affymetrix's MAS 5.0 expression measure. ← **General description**

Usage

```
mas5(object, normalize = TRUE, sc = 500, analysis = "absolute", ...)
```

 ← **Command and it's argument**

Arguments

- `object` an instance of `AffyBatch-class`
- `normalize` logical. If `TRUE` scale normalization is used after we obtain an instance of `exprSet-class`
- `sc` Value at which all arrays will be scaled to.
- `analysis` should we do absolute or comparison analysis, although "comparison" is still not implemented.
- `...` other arguments to be passed to `expresso`.

} ← **Detailed description of arguments**

Details

This function is a wrapper for `expresso` and `affy.scalevalue.exprSet`.

Value

Anatomy of a help file 2/2

R: MAS 5.0 expression measure - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <C:\Program Files\R\R-2.3.1\library\affy\html\mas5.html> Go Links

sc Value at which all arrays will be scaled to.
analysis should we do absolute or comparison analysis, although "comparison" is still not implemented.
... other arguments to be passed to [expresso](#).

Details

This function is a wrapper for [expresso](#) and [affy.scalevalue.exprSet](#).

Value

[exprSet-class](#)
The methods used by this function were implemented based upon available documentation. In particular a useful reference is Statistical Algorithms Description Document by Affymetrix. Our implementation is based on what is written in the documentation and as you might appreciate there are places where the documentation is less than clear. This function does not give exactly the same results. All source code of our implementation is available. You are free to read it and suggest fixes.
For more information visit this URL: <http://stat-www.berkeley.edu/users/bolstad/>

See Also

[expresso](#), [affy.scalevalue.exprSet](#)

Examples

```
data(affybatch.example)
eset <- mas5(affybatch.example)
```

[Package *affy* version 1.10.0 [Index](#)]

Description of how function actually works

What function returns

Related functions

Examples, can be run from R by:
example(mas5)

Functions or commands in R 1/3

- To use a function in a package, the package needs to be loaded in memory.
- Command for this is `library()`, for example:

```
library(affy)
```

- There are three parts in a command:
 - the command
 - brackets
 - Arguments inside brackets (these are not always present)

Functions or commands in R 2/3

- R is case sensitive, so take care when typing in the commands!
 - `library(affy)` works, but `Library(affy)` does not.
- Multiple commands can be written on the same line. Here we first remove missing values from the variable `year`, and then calculate its arithmetic average.
 - Writing:
 - `na.omit(year)`
 - `mean(year)`
 - Would be the same as
 - `mean(na.omit(year))`

Functions or commands in R 3/3

- Command can have many arguments. These are always given inside the brackets.
- Numeric (1, 2, 3...) or logic (T/F) values and names of existing objects are given for the arguments without quotes, but string values, such as file names, are always put inside quotes. For example:
 - `mas5(dat3, normalize=T, analysis="absolute")`

Data structures 1/6

- Vector
 - A list of numbers, such as (1,2,3,4,5)
 - R: `a<-c(1,2,3,4,5)`
 - Command `c` creates a vector that is assigned to object `a`
- Factor
 - A list of levels, either numeric or string
 - R: `b<-as.factor(a)`
 - Vector `a` is converted into a factor

Data structures 2/6

- Data frame
 - A table where columns can contain numeric and string values
 - R: `d<-data.frame(a, b)`
- Matrix
 - All columns must contain either numeric or string values, but these can not be combined
 - R: `e<-as.matrix(d)`
 - Data frame `d` is converted into a matrix `e`
 - R: `f<-as.data.frame(e)`
 - Matrix `e` is converted into a dataframe `f`

Data structures 3/6

- List
 - Contains a list of objects of possibly different types.
 - R: `g<-as.list(d)`
 - Converts a data frame `d` into a list `g`
- Class structures
 - Many of the Bioconductor functions create a formal class structure, such as an `AffyBatch` object.
 - They contain data in slots
 - Slots can be accessed using the `@`-operator:
 - `dat2@cdfName`

Data structures 4/6

- Some command need to get, for example, a matrix, and do not accept a data frame. Data frame would give an error message.
- To check the object type:
 - R: `class(d)`
- To check what fields there are in the object:
 - R: `d`
 - R: `str(d)`
- To check the size of the table/matrix:
 - R: `dim(d)`
- To check the length of a factor or vector:
 - R: `length(a)`

Data structures 5/6

- Some data frame related commands:
 - R: `names (d)`
 - Reports column names
 - R: `row.names (d)`
 - Reports row names
- These can also be used for giving the names for the data frame. For example:
 - R: `row.names (d) <- c ("a", "b", "c", "d", "e")`
 - Letters from a to e are used as the row names for data frame d
 - Note the quotes around the string values!
 - R: `row.names (d)`

Data structures 5/6

- Naming objects:
 - Never use command names as object names!
 - If your unsure whether something is a command name, type to the comman line first. If it gives an error message, you're safe to use it.
 - Object names can't start with a number
 - Never use special characters, such as å, ä, or ö in object names.
 - Underscore (`_`) is not usable, use dot (`.`) instead:
 - Not acceptable: `good_data`
 - Better way: `good.data`
 - Object names are case sensitive, just like commands

Reading data 1/2

- Command for reading in text files is:

```
read.table("suomi.txt", header=T, sep="\t")
```

- This examples has one command with three arguments: file name (in quotes), header that tells whether columns have titles, and sep that tells that the file is tab-delimited.

Reading data 2/2

- It is customary to save the data in an object in R. This is done with the assignment operator (`<-`):

```
dat<-read.table("suomi.txt", header=T, sep="\t")
```

- Here, the data read from file `suomi.txt` is saved in an object `dat` in R memory.
- The name of the object is on the left and what is assigned to the object is on the right.
- Command `read.table()` creates a data frame.

Using data frames

- Individual columns in the data frame can be accessed using one of the following ways:
 - Use its name:
 - `dat$year`
 - `dat` is the data frame, and `year` is the header of one of its columns. Dollar sign (\$) is an operator that accesses that column.
 - Split the data frame into variables, and use the names directly:
 - `attach(dat)`
 - `year`
 - Use subscripts

Subscripts 1/2

- Subscripts are given inside square brackets after the object's name:
 - `dat[,1]`
 - Gets the first column from the object `dat`
 - `dat[1,]`
 - Gets the first row from the object `dat`
 - `dat[1,1]`
 - Gets the first row and its first column from the object `dat`
- Note that `dat` is now an object, not a command!

Subscripts 2/2

- Subscripts can be used for, e.g., extracting a subset of the data:
 - `dat[which(dat$year>1900),]`
 - Now, this takes a bit of pondering to work out...
 - First we have the object `dat`, and we are accessing a part of it, because its name is followed by the square brackets
 - Then we have one command (`which`) that makes an evaluation whether the column `year` in the object `dat` has a value higher than 1900.
 - Last the subscript ends with a comma, that tells us that we are accessing rows.
 - So this command takes all the rows that have a year higher 1900 from the object `dat` that is a data frame.

Writing tables

- To write a table:
 - `write.table(dat, "dat.txt", sep="\t")`
 - Here an object `dat` is written to a file called `dat.txt`. This file should be tab-delimited (argument `sep`).
- To capture what is written on the screen:
 - `sink("output.txt")`
 - `dat`
 - `sink()`
 - Here, output written on the screen should be written to a file `output.txt` instead. Contents of the object `dat` are written to the named file. Last, the file is closed.
 - Note that if you accidentally omit the last command, you'll not be able to see any output on the screen, because output is still redirected to a file!

Quitting R

- Use command `q()` or menu choice File->Exit.
- R asks whether to save workspace image. If you do, all the object currently in R memory are written to a file `.Rdata`, and all command will be written a file `.Rhistory`.
- These can be loaded later, and you can continue your work from where you left it.
- Loading can be done after starting R using the manu choices File->Load Workspace and File->Load History.

In summary 1/2

- Commands can be recognized from the brackets "()" that follow them. If you calculate how many bracket pairs there are, you'll be able to identify the number of commands.
 - `pData(dat) <- pd`
- Assignment to an object is denoted by "<-" or "->" or "=". If you see a notation "=", you'll be looking at a comparison operator.
 - Many other notations can be found from the documentation for the Base package or R.
- Table-like objects are often followed by square brackets "[]". Square brackets never associate with commands, only objects.
 - `dat[,1]`
- Special characters \$ and @ are used denoting individual columns in a data frame or an individual slot in a class type of an object, respectively.
 - `dat$year`
 - `dat2@cdfName`

In summary 2/2

- If you encounter a new command during the exercises, and you'd like to know what it does, please consult the documentation. All R commands are listed nowhere, and the only way to get to know new commands is to read the documentation files, so we'd like you to practise this yourself.
- You'll probably see command and notations that were not introduced in this talk. This is intentional, because we thought that these things are best handled on a situational basis. In such cases, please ask for more clarifications if needed.
- If you run into problems, please ask for help from the teachers. That's why we are here!

Installing R

Downloading R

The R Project for Statistical Computing - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print Mail Stop

Address <http://www.r-project.org/> Go Links

The R Project for Statistical Computing

Download CRAN

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

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

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

Misc
[Bioconductor](#)

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.

<http://cran.r-project.org/mirrors.html> Internet

Downloading R

The screenshot shows the CRAN Mirrors page in Microsoft Internet Explorer. The browser title is "The R Project for Statistical Computing - Microsoft Internet Explorer". The address bar shows "http://www.r-project.org/". The page content includes the R logo and the heading "CRAN Mirrors". Below the heading, a message states: "The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you:". A list of mirrors is provided, organized by country. A red box highlights the Austria mirror entry.

Country	URL	Location
Australia	http://cran.au.r-project.org/	PlanetMirror, Brisbane
Australia	http://cran.melb.unimelb.edu.au/	University of Melbourne
Austria	http://cran.at.r-project.org/	Technische Universitaet Wien
Brazil	http://cran.br.r-project.org/	Universidade Federal do Parana
Brazil	http://www.insecta.ufv.br/CRAN/	Federal University of Vicosa
Brazil	http://cran.fiocruz.br/	Oswaldo Cruz Foundation, Rio de Janeiro
Brazil	http://lmq.esalq.usp.br/CRAN/	University of Sao Paulo, Piracicaba
Brazil	http://www.vps.fmvz.usp.br/CRAN/	University of Sao Paulo, Sao Paulo
Canada	http://cran.stat.sfu.ca/	Simon Fraser University, Burnaby
Canada	http://probability.ca/cran/	University of Toronto
China	http://www.lnbe.seu.edu.cn/CRAN/	Southeast University, Nanjing
Denmark	http://cran.dk.r-project.org/	dotsrc.org, Aalborg
France	http://cran.fr.r-project.org/	CICT, Toulouse
France	http://cran.univ-lyon1.fr/	Dept. of Biometry & Evol. Biology, University of Lyon
France	http://mirror.internet.tp/cran/	Boese Internet, Paris
Germany		

Navigation links on the left side of the page include: About R, What is R?, Contributors, Screenshots, What's new?, Download CRAN, R Project Foundation, Members & Donors, Mailing Lists, Bug Tracking, Developer Page, Search, Documentation, Manuals, FAQs, Newsletter, Books, Other, Misc, Bioconductor.

Downloading R

The Comprehensive R Archive Network - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print Mail Stop

Address <http://cran.at.r-project.org/> Go Links

The Comprehensive R Archive Network

Frequently used pages

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:

- [Linux](#)
- [MacOS X](#)
- [Windows \(95 and later\)](#)

Source Code for all Platforms

Windows and Mac users most likely want 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** (2006-04-24): [R-2.3.0.tar.gz](#) (read [what's new](#) in the latest version).
- 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

CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

Documentation
[Manuals](#)
[FAQs](#)
[Contributed](#)
[Newsletter](#)

Done Internet

Downloading R

The Comprehensive R Archive Network - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print Mail Stop

Address <http://cran.at.r-project.org/> Go Links

R-2.3.0 for Windows

This directory contains a binary distribution of R-2.3.0 to run on Windows 95, 98, ME, NT4.0, 2000 and XP on Intel/clone chips.

Patches to this release are incorporated in the [r-patched snapshot build](#).

A build of the development version (which will eventually become the next major release of R) is available in the [r-devel snapshot build](#).

In this directory:

README.R-2.3.0	Installation and other instructions.
CHANGES	New features of this Windows version.
NEWS	New features of all versions.
R-2.3.0-win32.exe	Setup program (about 27 megabytes). Please download this from a mirror near you . This corresponds to the file named SetupR.exe or rwXXXX.exe in pre-2.2.0 releases.
old	The previous release.
md5sum.txt	md5sum output for the setup program. A Windows GUI version of md5sum is available at http://www.md5summer.org/ ; a Windows command line version is available at http://www.etree.org/md5com.html .

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information, including upgrade advice.

Note to webmasters: A stable link which will redirect to the current Windows binary release is [<CRAN MIRROR>/bin/windows/base/release.htm](http://cran.at.r-project.org/bin/windows/base/release.htm).

CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

Documentation
[Manuals](#)
[FAQs](#)
[Contributed](#)
[Newsletter](#)

<http://cran.at.r-project.org/bin/windows/base/R-2.3.0-win32.exe> Internet

Installing R for Windows

- Execute the R-2.3.0-win32.exe with administrator privileges
- Once the program is installed, run the R program by clicking on its icon
- R 2.2.1 with Bioconductor 1.7.0 is installed on corona.csc.fi, also
- R 2.3.1 is in works

Downloading Bioconductor

The screenshot shows the Bioconductor website in a Microsoft Internet Explorer browser window. The browser title is "Welcome to Bioconductor — bioconductor.org - Microsoft Internet Explorer". The address bar shows "http://www.bioconductor.org/". The website header features the Bioconductor logo and the text "BioConductor is an open source and open development software project for the analysis and comprehension of genomic data." Below the header is a navigation menu with links for "home", "what is it?", "download", "documentation", "people", and "publications". A search bar is located on the right side of the page. The main content area is divided into two columns. The left column, titled "QUICK LINKS", lists several links: "What is it?", "Install - How To", "FAQ", "For Developers", "BioC Workshops", "How to up-load packages", "Metadata", and "BioC News". The right column, titled "project news", lists two news items: "2006-04-06 Changes in BioC Devel, March 2006" and "2006-02-08 Changes in BioC Devel, February 2006", with a "More..." link below. Below the "project news" section is a "Latest News" section with three items: "Bioconductor 1.8 released 27 April, 2006. This release is designed for R 2.3.0. View the packages here", "Upcoming: Computational and Statistical short course on Microarrays. 18 June - 23 June, 2006, Brixen-Bressanone (BZ), Italy.", and "Upcoming: Workshops on S, R and Bioconductor. 7 July - 8 July 2006, Auckland, New Zealand." Below the "Latest News" section is an "Annoucement: BioC2006 Conference" section with the text "Bioconductor User and Developer Conference August 3-4, 2006 Seattle, WA, USA" and a "Detailed information" link. The browser's taskbar at the bottom shows the Start button, several open applications (Google, Microsoft, Welcome t..., statistics, RGui), and the system clock showing 09:05.

Welcome to Bioconductor — bioconductor.org - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print Mail Print Preview Stop

Address http://www.bioconductor.org/ Go Links

BIOCONDUCTOR
open source software for bioinformatics

BioConductor is an open source and open development software project for the analysis and comprehension of genomic data.

home what is it? download documentation people publications

search

project news

- ▶ [2006-04-06](#)
Changes in BioC Devel, March 2006
- ▶ [2006-02-08](#)
Changes in BioC Devel, February 2006

[More...](#)

QUICK LINKS

- ▶ [What is it?](#)
- ▶ [Install - How To](#)
- ▶ [FAQ](#)
- ▶ [For Developers](#)
- ▶ [BioC Workshops](#)
- ▶ [How to up-load packages](#)
- ▶ [Metadata](#)
- ▶ [BioC News](#)

Latest News

Bioconductor 1.8 released 27 April, 2006. This release is designed for R 2.3.0. View the packages [here](#)

Upcoming: [Computational and Statistical short course on Microarrays.](#)
18 June - 23 June, 2006, Brixen-Bressanone (BZ), Italy.

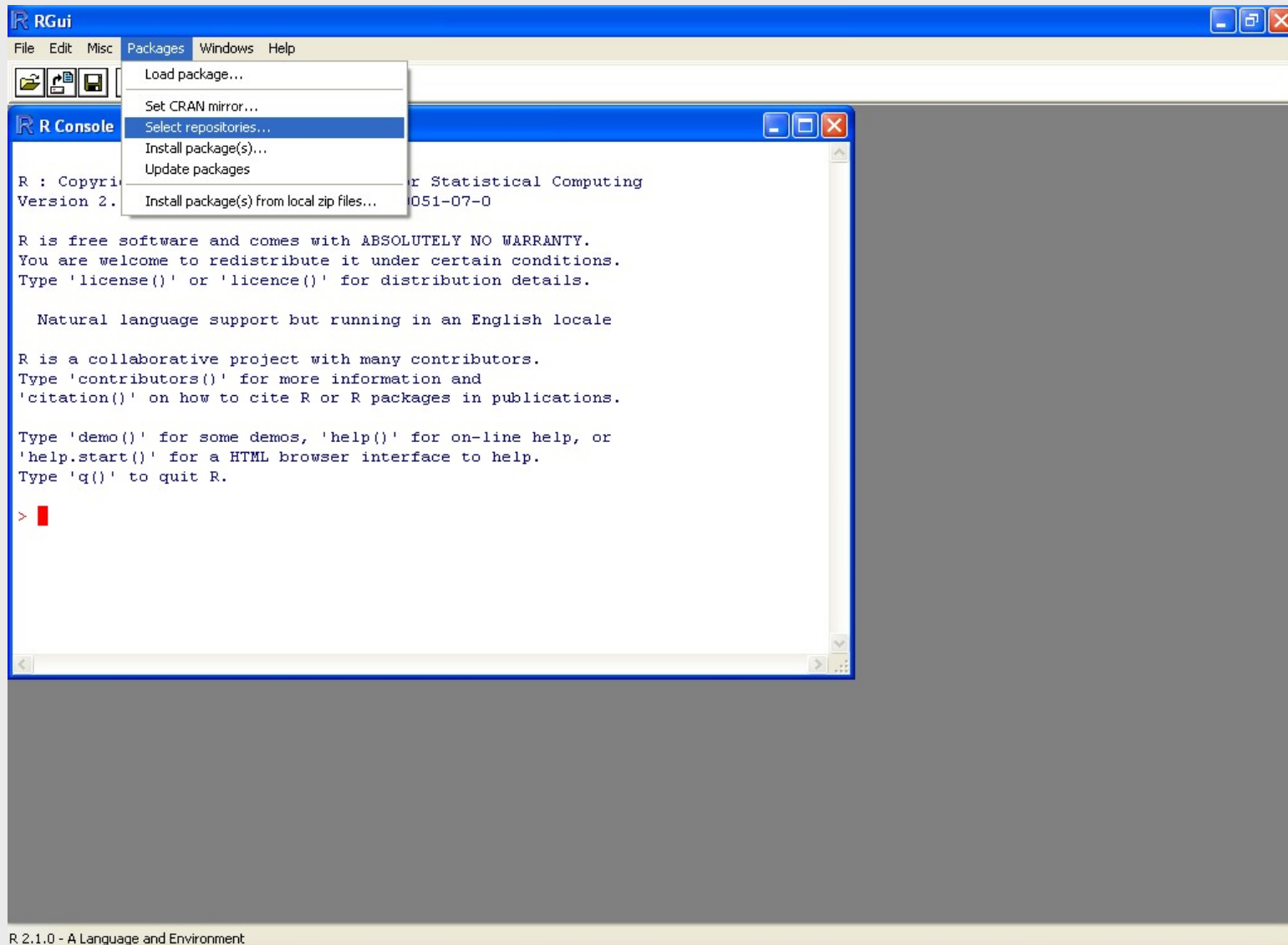
Upcoming: [Workshops on S, R and Bioconductor.](#)
7 July - 8 July 2006, Auckland, New Zealand.

Annoucement: BioC2006 Conference

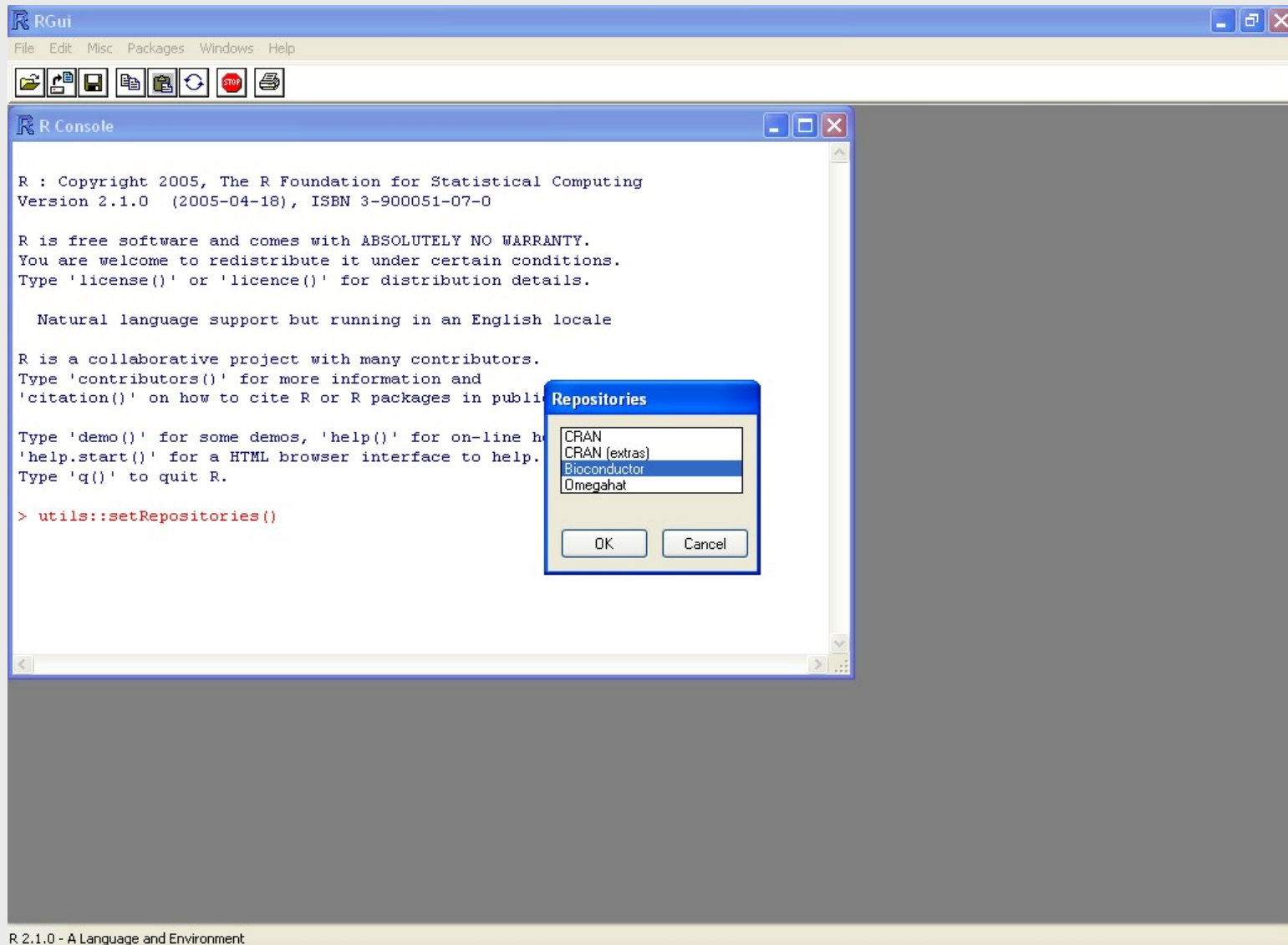
Bioconductor User and Developer Conference
August 3-4, 2006
Seattle, WA, USA
[Detailed information](#)

start rw2010 Google - ... Microsoft ... Welcome t... statistics RGui FI 09:05

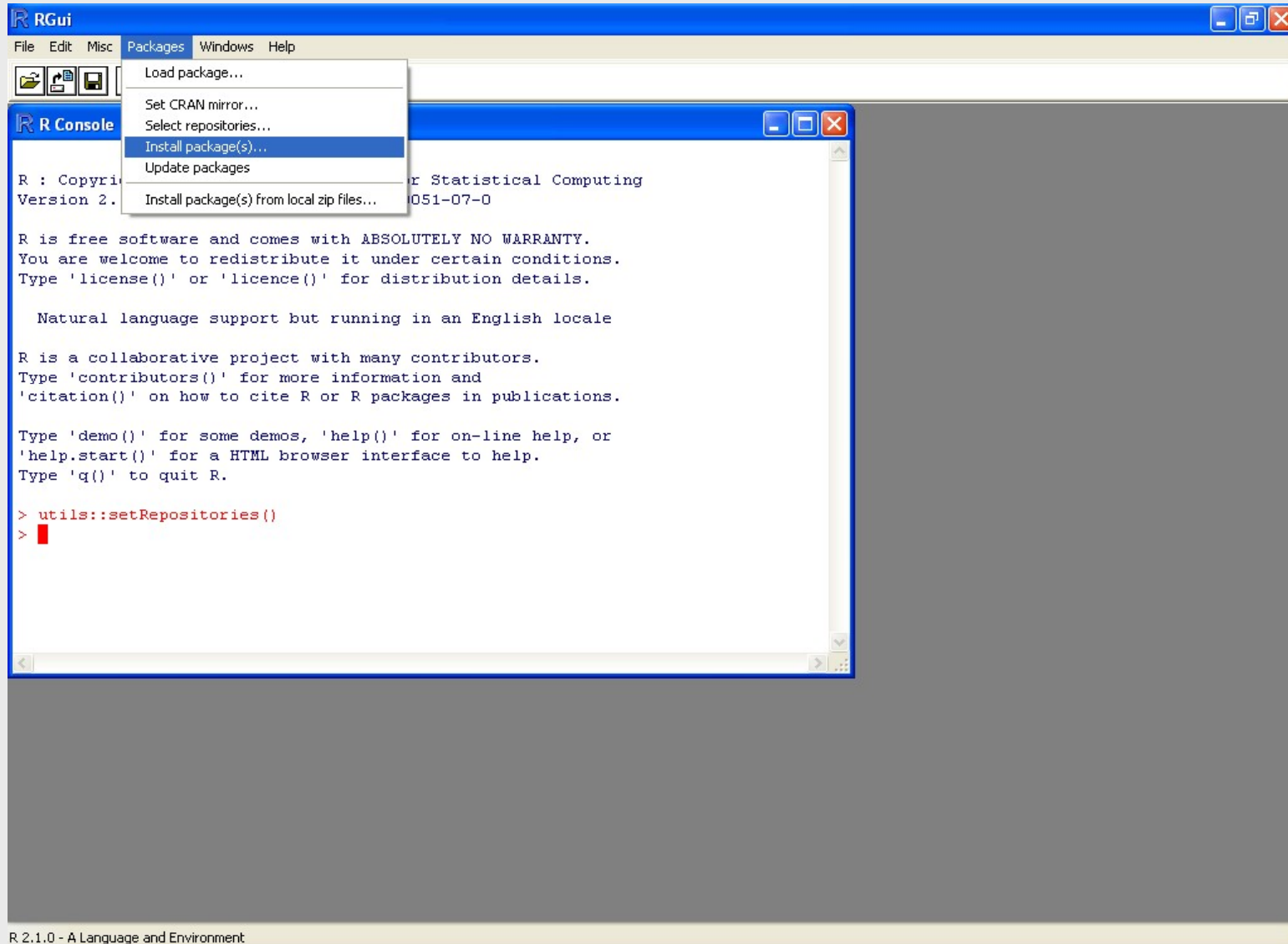
Installing Bioconductor



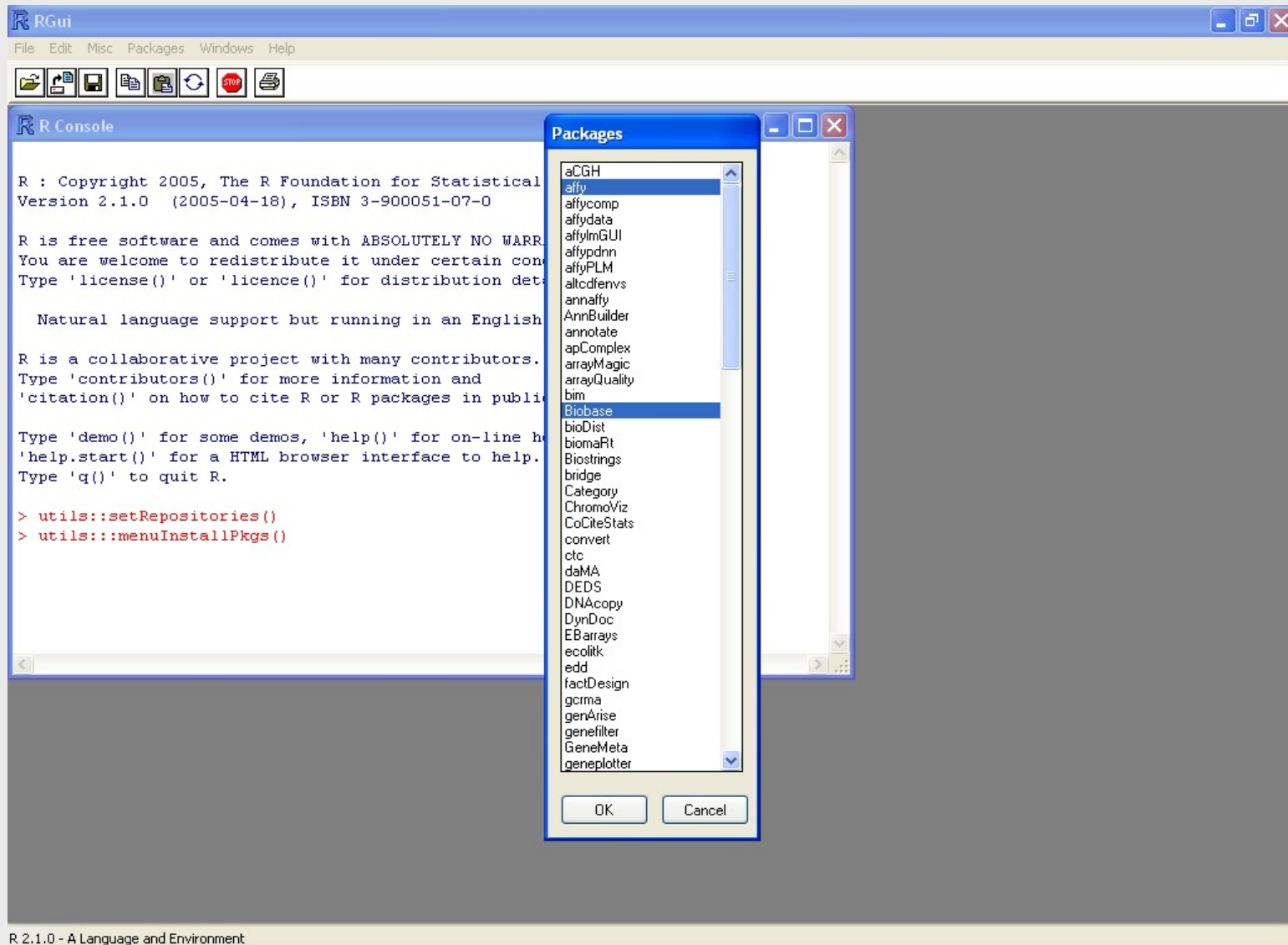
Installing Bioconductor



Installing Bioconductor



Installing Bioconductor



Installing Bioconductor (the best way)

- Alternatively, you can install Bioconductor using a script:

```
source("http://www.bioconductor.org/biocLite.R")  
biocLite()
```

```
biocLite(c("hgu133a", "hgu133acdf",  
          "hgu133aprobe", "ygs98", "ygs98cdf",  
          "ygs98probe"))
```


Linear Models & Descriptive Statistics

- **Has functions for all common statistics**
- **summary() gives lowest, mean, median, first, third quartiles, highest for numeric variables**
- **stem() gives stem-leaf plots**
- **table() gives tabulation of categorical variables**

Basics

- **Highly Functional**
 - Everything done through functions
 - Strict named arguments
 - Abbreviations in arguments OK (e.g. T for TRUE)
- **Object Oriented**
 - Everything is an object
 - “< -” is an assignment operator
 - “X <- 5”: X GETS the value 5

Data Structures

- **Supports virtually any type of data**
- **Numbers, characters, logicals (TRUE/FALSE)**
- **Arrays of virtually unlimited sizes**
- **Simplest: Vectors and Matrices**
- **Lists: Can Contain mixed type variables**
- **Data Frame: Rectangular Data Set**

Data Structure in R

	Linear	Rectangular
All Same Type	VECTORS	MATRIX*
Mixed	LIST	DATA FRAME

Reading Data: summary

- **Directly using a vector e.g.: `x <- c(1,2,3...)`**
- **Using `scan` and `read.table` function**
- **Using `matrix` function to read data matrices**
- **Using `data.frame` to read mixed data**
- **`library(foreign)` for data from other programs**

Accessing Variables

- **edit(<mydataobject>)**
- **Subscripts essential tools**
 - **x[1]** identifies first element in vector x
 - **y[1,]** identifies first row in matrix y
 - **y[,1]** identifies first column in matrix y
- **\$ sign for lists and data frames**
 - **myframe\$age** gets age variable of myframe
 - **attach(dataframe)** -> extract by variable name

Subset Data

- **Using subset function**
 - `subset()` will subset the dataframe
- **Subscripting from data frames**
 - `myframe[,1]` gives first column of myframe
- **Specifying a vector**
 - `myframe[1:5]` gives first 5 rows of data
- **Using logical expressions**
 - `myframe[myframe[,1], < 5,]` gets all rows of the first column that contain values less than 5

Graphics

- **Plot an object, like: `plot(num.vec)`**
 - here plots against index numbers
- **Plot sends to graphic devices**
 - can specify which graphic device you want
 - `postscript`, `gif`, `jpeg`, etc...
 - you can turn them on and off, like: `dev.off()`
- **Two types of plotting**
 - high level: graphs drawn with one call
 - Low Level: add additional information to existing graph

Programming in R

- **Functions & Operators typically work on entire vectors**
- **Expressions surrounded by `{}`**
- **Codes separated by newlines, “`;`” not necessary**
- **You can write your own functions and use them**

Statistical Functions in R

- **Descriptive Statistics**
- **Statistical Modeling**
 - **Regressions: Linear and Logistic**
 - **Probit, Tobit Models**
 - **Time Series**
- **Multivariate Functions**
- **Inbuilt Packages, contributed packages**

Descriptive Statistics

- **Has functions for all common statistics**
- **summary() gives lowest, mean, median, first, third quartiles, highest for numeric variables**
- **stem() gives stem-leaf plots**
- **table() gives tabulation of categorical variables**

Statistical Modeling

- **Over 400 functions**
 - `lm`, `glm`, `aov`, `ts`
- **Numerous libraries & packages**
 - `survival`, `coxph`, `tree` (recursive trees), `nls`, ...
- **Distinction between factors and regressors**
 - **factors: categorical, regressors: continuous**
 - **you must specify factors unless they are obvious to R**
 - **dummy variables for factors created automatically**
- **Use of `data.frame` makes life easy**

How to model

- **Specify your model like this:**
 - $y \sim x_i + c_i$, where
 - y = outcome variable, x_i = main explanatory variables, c_i = covariates, + = add terms
 - Operators have special meanings
 - + = add terms, : = interactions, / = nesting, so on...
- **Modeling -- object oriented**
 - each modeling procedure produces objects

Synopsis of Operators

Operato	Usually means	In Formula means
+ or -	add or subtract	add or remove terms
*	multiplication	main effect and
/	division	interactions main effect and nesting
:	sequence	interaction only
^	exponentiation	limiting interaction
%in%	no specific	depths nesting only

Modeling Example: Regression

carReg <- lm(speed~dist, data=cars)

carReg = becomes an object

**to get summary of this regression, we
type**

summary(carReg)

to get only coefficients, we type

coef(carReg), or carReg\$coef

don't want intercept? add 0, so

carReg <- lm(speed~0+dist, data=cars)

Multivariate Techniques

- **Several Libraries available**
 - mva, hmisc, glm,
 - **MASS: discriminant analysis and multidim scaling**
- **Econometrics packages**
 - dse (multivariate time series, state-space models), ineq: for measuring inequality, poverty estimation, its: for irregular time series, sem: structural equation modeling, and so on...

[<http://www.mayin.org/ajayshah/>]

Summarizing...

- **Effective data handling and storage**
- **large, coherent set of tools for data analysis**
- **Good graphical facilities and display**
 - on screen
 - on paper
- **well-developed, simple, effective programming**