

Basic Introduction to R and Bioconductor

Aedin Culhane May 23, 2011 (aedin@jimmy.harvard.edu)

R (www.r-project.org) is an open source interactive computer system for visualization and analysis of statistical data. Bioconductor (www.bioconductor.org) is a project within R. The Bioconductor project is a source of many resources relevant to data analysis for high-throughput biology.

Installing R and Bioconductor

1. Download R from <http://www.r-project.org>. The primary portal for R is <http://cran.r-project.org>. Many local mirrors exist and these may provide quicker downloads.

R can be download precompiled for Linux, Mac OS and Windows. If you are using Linux, R can be installed using yum install.



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

About R
[R Homepage](#)
[The R Journal](#)

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

Documentation
[Manuels](#)
[FAQs](#)
[Contributed](#)

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

- [Linux](#)
- [MacOS X](#)
- [Windows](#)

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** (2011-04-13): [R-2.13.0.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.

To install R simply, click on download and select the version appropriate for your operating system.

For the most users, this precompiled version of R is sufficient. However if you wish to write an extension package for R/Bioconductor, you need to install R from source. Please refer to the R FAQ for directions on this. Windows users will need to install unix tools/perl/latex, please refer to the excellent guide at <http://www.murdoch-sutherland.com/Rtools/>.

R Extension Packages

On <http://cran.r-project.org/> or the mirror you are using. To find out about contributed package please click on the link **Contributed extension packages** at the bottom of the main download page. This will link to a page with lots of information on contributed packages and to “CRAN Task Views”.


Contributed Packages

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, 28 views are available.

CRAN task views, provide over 25 categories of contributed packages which makes installation of these packages simpler.



CRAN Task Views

Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
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
HighPerformanceComputing	High-Performance and Parallel Computing with R
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Phylogenetics	Phylogenetics, Especially Comparative Methods
Psychometrics	Psychometric Models and Methods
ReproducibleResearch	Reproducible Research
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")
```

For example to install genetics packages type

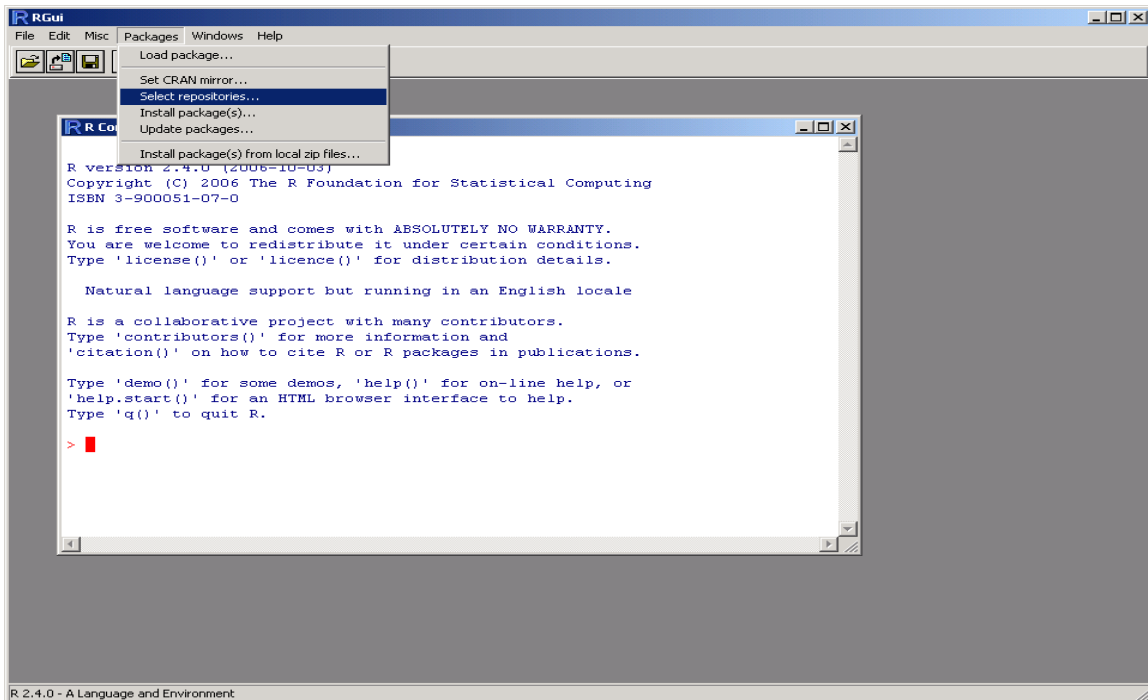
```
install.packages("ctv")
library(ctv)
install.views("Genetics")
```

Note these categories contain many packages, for example this Genetics contains 29 packages.

Installing Individual Contributed or Extension Packages

Although a team of statisticians and programmers maintain R, many independent groups submit contributed packages. This is a very extensive resource of hundreds of programs and will not be install by default.

R extensions can be installed easily. R extensions are stored in repositories. Select the package repositories using **Packages -> Select repositories**



The R repositories are:

CRAN: basic R distribution

CRAN (extras)- Contributed R packages. There are several hundred.

Bioconductor: The [Bioconductor Project](#) produces an open source software framework that will assist biologists and statisticians working in bioinformatics, with primary

emphasis on inference using DNA microarrays. A CRAN style R package repository is available via <http://www.bioconductor.org/>.

Omegahat: The Omegahat Project for Statistical Computing provides a variety of open-source software for statistical applications, with special emphasis on web-based software, Java, the Java virtual machine, and distributed computing. But recently many other plugins including are available, into RGoogleStorage, RGoogleDocs, ROpenOffice, RAmazonDBRest, RAmazonS3, R2GoogleMaps etc. CRAN style R package repository is available via <http://www.omegahat.org/>

Select one repository, it may provide a select of CRAN mirrors, select one close by. You can now download extension packages.

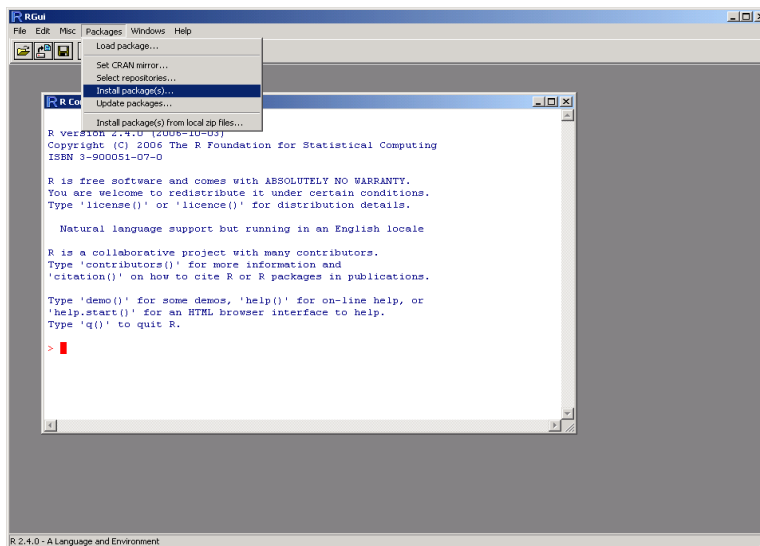
There are several ways to install extension packages

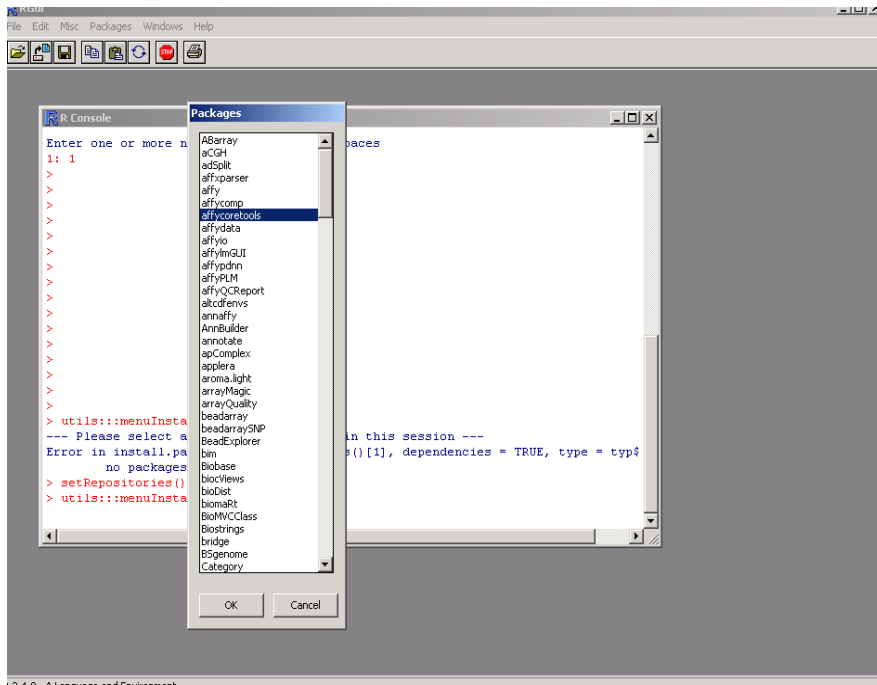
1. From the GUI, select **Packages -> Install Packages**. Note only the packages from the selected repositories will be shown.

2. From command line type

```
install.packages("packagename")
```

3. Download a compressed copy of the package from the R website and install from local zip or .tar.gz file





Alternatively one can install a package and specific the repository using the following command within R:

```
install.packages("packagename", repos = "http://www.omegahat.org/R")
```

Installing Bioconductor

The recommended method to install Bioconductor is using the following script, which can be sourced directly:

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

This will automatically download core bioconductor packages. Due to the number and size of Bioconductor, many useful packages will not be installed, so these can be installed as described above.

Using a script to simplify Installation of R and Bioconductor

Once you are familiar with R and Bioconductor, you may wish to script the above progress so that updates to R are simpler ;-)

I use the following script, which I saved to a plain text file called bioC_install.R

```
### Packages for R.
install.packages("ade4")
install.packages("scatterplot3d")
install.packages("Rcurl")
install.packages("DBI")
install.packages("RMySQL")
install.packages("impute")
install.packages("fastICA")
install.packages("e1071")

### BioC Packages for R.
source("http://www.bioconductor.org/biocLite.R")
biocLite()
biocLite("made4")
biocLite("hgu133plus2")
biocLite("hgu133plus2cdf")
biocLite("hgu133plus2probe")
biocLite("Heatplus")
biocLite("biomaRt")
```

To install these packages, I simply type command

```
source("./bioC_install.R")
```

and then go for coffee ;-)

R and Bioconductor are pretty “intelligent”, if you wish to install a package that has dependencies (needs other packages) it will automatically install these.

Introduction to Bioconductor Packages

Bioconductor (<http://www.bioconductor.org>) has a really nice “task views” interface to ipacackages in Bioconductor. To find out more about these click on **Packages link** or click **Install** and this will give you a link to **BioCViews**

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, more than [460 packages](#), and an active user community.

Use Bioconductor for...

- Microarrays**
Import Affymetrix, Illumina, Nimblegen, Agilent, and other platforms. Perform quality assessment, normalization, differential expression, clustering, classification, gene set enrichment, genetical genomics and other workflows for expression, exon, copy number, SNP, methylation and other assays. Access GEO, ArrayExpress, Biomart, UCSC, and other community resources.
- High Throughput Assays**
Import, transform, edit, analyze and visualize flow cytometric, mass spec, HTqPCR, cell-based, and other assays.
- Sequence Data**
Import fasta, fastq, ELAND, MAQ, BWA, Bowtie, BAM, gff, bed, wig, and other sequence formats. Trim, transform, align, and manipulate sequences. Perform quality assessment, CHIP-seq, differential expression, RNA-seq, and other workflows. Access the Sequence Read Archive.
- Annotation**
Use microarray probe, gene, pathway, gene ontology, homology and other annotations. Access GO, KEGG, NCBI, Biomart, UCSC, vendor, and other sources.

[Mailing Lists](#) [Subscribe »](#) [Events](#) [News](#)

BioCViews are grouped into 3 categories,

Annotation Data
Experiment Data.
Software

Bioconductor version 2.8 (Release)

- ▶ AnnotationData (594)
- ▶ ExperimentData (79)
- ▶ Software (467)

Click on each of these clicks to explore the packages in Bioconductor

Bioconductor Software Packages

Software packages are sub divided into the categories:

Bioconductor version 2.8 (Release)	
▶	AnnotationData (594)
▶	ExperimentData (79)
▼	Software (467)
▶	Annotation (61)
▶	AssayDomains (182)
▶	AssayTechnologies (289)
▼	Bioinformatics (270)
	Classification (34)
	Clustering (38)
	MultipleComparisons (31)
	Preprocessing (83)
	QualityControl (46)
	SequenceMatching (7)
	TimeCourse (11)
▶	BiologicalDomains (46)
▼	Infrastructure (194)
	DataImport (40)
	DataRepresentation (10)
	GraphsAndNetworks (40)
	GUI (8)
	Visualization (78)

Each contains a long list of contributed packages. For example clicking on Bioinformatics -> PreProcessing returns a long list of packages with methods for handling many different data types

Packages

Software > Bioinformatics > Preprocessing

▪ [ABarray](#) ▪ [affy](#) ▪ [affycomp](#) ▪ [AffyExpress](#) ▪ [affyILM](#) ▪ [affyImGUI](#) ▪ [affyPara](#) ▪ [affypdnn](#) ▪ [affyPLM](#) ▪ [AffyTiling](#) ▪ [Aqi4x44PreProcess](#) ▪ [AqiMicroRna](#) ▪ [altcdfenvs](#) ▪ [aroma.light](#) ▪ [ArrayTools](#) ▪ [beadarray](#) ▪ [beadarraySNP](#) ▪ [BUS](#) ▪ [CALIB](#) ▪ [cellHTS2](#) ▪ [CGHcall](#) ▪ [CGHnormaliter](#) ▪ [codelink](#) ▪ [crlmm](#) ▪ [DEGseq](#) ▪ [dyebias](#) ▪ [ExiMiR](#) ▪ [farms](#) ▪ [frma](#) ▪ [frmaTools](#) ▪ [qcrma](#) ▪ [genArise](#) ▪ [Harshlight](#) ▪ [HELP](#) ▪ [HTqPCR](#) ▪ [imageHTS](#) ▪ [limma](#) ▪ [limmaGUI](#) ▪ [LMGene](#) ▪ [lumi](#) ▪ [LVSmRNA](#) ▪ [maCorrPlot](#) ▪ [maigesPack](#) ▪ [makecdfenv](#) ▪ [makePlatformDesign](#) ▪ [MANOR](#) ▪ [marray](#) ▪ [MBCB](#) ▪ [MEDIPS](#) ▪ [methylumi](#) ▪ [Mfuzz](#) ▪ [MiChip](#) ▪ [multiscan](#) ▪ [nnNorm](#) ▪ [NTW](#) ▪ [oligo](#) ▪ [OLIN](#) ▪ [OLINqui](#) ▪ [oneChannelGUI](#) ▪ [puma](#) ▪ [qpcrNorm](#) ▪ [qrac](#) ▪ [rama](#) ▪ [RBioinf](#) ▪ [RefPlus](#) ▪ [Ringo](#) ▪ [rMAT](#) ▪ [RNAinteract](#) ▪ [RNAither](#) ▪ [Rolexa](#) ▪ [RPA](#) ▪ [SAGx](#) ▪ [simpleaffy](#) ▪ [snapCGH](#) ▪ [snm](#) ▪ [spotSegmentation](#) ▪ [Starr](#) ▪ [stepNorm](#) ▪ [TargetSearch](#) ▪ [tilingArray](#) ▪ [TurboNorm](#) ▪ [vsn](#) ▪ [xps](#)

and each package page give details, help and a link to a vignette (tutorial).

affy

Methods for Affymetrix Oligonucleotide Arrays

Bioconductor version: Release (2.8)

The package contains functions for exploratory oligonucleotide array analysis. The dependence on tkWidgets only concerns few convenience functions. 'affy' is fully functional without it.

Author: Rafael A. Irizarry, Laurent Gautier, Benjamin Milo Bolstad, and Crispin Miller with contributions from Magnus Astrand, Leslie M. Cope, Robert Gentleman, Jeff Gentry, Conrad Halling, Wolfgang Huber, James MacDonald, Benjamin I. P. Rubinstein, Christopher Workman, John Zhang

Maintainer: Rafael A. Irizarry

To install this package, start R and enter:

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

Documentation

- [PDF](#) [R Script](#) 1. Primer
 - [PDF](#) [R Script](#) 2. Built-in Processing Methods
 - [PDF](#) [R Script](#) 3. Custom Processing Methods
 - [PDF](#) [R Script](#) 4. Import Methods
- [Reference Manual](#)

Details

biocViews [Microarray](#), [OneChannel](#), [Preprocessing](#)

Depends R, [Biobase](#)

Imports [affyio](#), [Biobase](#), graphics, grDevices, methods, [preprocessCore](#), stats, utils

Suggests [tkWidgets](#), [affydata](#)

System

Requirements

License [LGPL \(>= 2.0\)](#)

URL

Depends On Me [AffyExpress](#), [AqiMicroRna](#), [ArrayTools](#), [ExiMiR](#), [LMGene](#), [LVSmiRNA](#), [MLP](#), [RPA](#), [RefPlus](#), [Starr](#), [affyContam](#), [affyPLM](#), [affyPara](#), [affyQCReport](#), [affycoretools](#), [affyImGUI](#), [affypdnn](#), [altcdfenvs](#), [arrayMvout](#), [bqx](#), [dualKS](#), [farms](#), [frmaTools](#), [qcrma](#), [loqitT](#), [maDB](#), [panp](#), [plw](#), [puma](#), [qpcrNorm](#), [rHVDM](#), [simpleaffy](#), [sscore](#), [webbioc](#)

Imports Me [AffyTiling](#), [ArrayExpress](#), [ArrayTools](#), [GEOsubmission](#), [HTqPCR](#), [Harshlight](#), [TurboNorm](#), [affyILM](#), [affyQCReport](#), [arrayQualityMetrics](#), [farms](#), [frma](#), [qcrma](#), [lumi](#), [makecdfenv](#), [plier](#), [plw](#), [puma](#), [pvac](#), [simpleaffy](#), [tilingArray](#), [vsn](#)

Suggests Me [AnnotationDbi](#), [BiocCaseStudies](#), [Biostrings](#), [BufferedMatrixMethods](#), [ExpressionView](#), [GeneRegionScan](#), [TurboNorm](#), [beadarraySNP](#), [beadarray](#), [factDesign](#), [hexbin](#), [limma](#), [made4](#), [oneChannelGUI](#), [siqgenes](#)

Version 1.30.0

Bioconductor Annotation Data Packages

Almost 600 bioconductor packages are annotation packages. These are an incredible value resource in bioconductor. These packages provide annotation on the genes on microarrays.

This resource can be searched by organism, chip manufacturer, chip name etc

For example click on chip manufacturer -> AffymetrixChip or Organism -> homo sapiens to get a list of relevant annotation packages.

Each Affymetrix array is represented by 3 annotation packages, .db, cdf and probe. For example the human GeneChip Human Genome U133 Plus 2.0 Array has the following

hgu133plus2.db	annotation data (chip hgu133plus2) assembled using data from public repositories
hgu133plus2cdf	chip description file (cdf)
hgu133plus2probe	probe sequence data (probe).

The latter two are compiled from Affymetrix or manufacturer information

To install an annotation package, follow the same guidelines as package installation

```
source("http://www.bioconductor.org/biocLite.R")
biocLite("hgu133plus2.db")
```

To find out about a package:

```
> library(hgu133plus2.db)
> hgu133plus2()

> hgu133plus2()
Quality control information for hgu133plus2:

This package has the following mappings:

hgu133plus2ACCNUM has 54675 mapped keys (of 54675 keys)
hgu133plus2ALIAS2PROBE has 73685 mapped keys (of 110538 keys)
hgu133plus2CHR has 40772 mapped keys (of 54675 keys)
hgu133plus2CHRLNGTHS has 93 mapped keys (of 93 keys)
hgu133plus2CHRLOC has 39212 mapped keys (of 54675 keys)
hgu133plus2CHRLOCEND has 39212 mapped keys (of 54675 keys)
hgu133plus2ENSEMBL has 37294 mapped keys (of 54675 keys)
hgu133plus2ENSEMBL2PROBE has 18042 mapped keys (of 19887 keys)
hgu133plus2ENTREZID has 40801 mapped keys (of 54675 keys)
hgu133plus2ENZYME has 4616 mapped keys (of 54675 keys)
hgu133plus2ENZYME2PROBE has 928 mapped keys (of 936 keys)
hgu133plus2GENENAME has 40801 mapped keys (of 54675 keys)
```

```
hgu133plus2GO has 35250 mapped keys (of 54675 keys)
hgu133plus2GO2ALLPROBES has 13288 mapped keys (of 13360 keys)
hgu133plus2GO2PROBE has 10091 mapped keys (of 10161 keys)
hgu133plus2MAP has 40571 mapped keys (of 54675 keys)
hgu133plus2OMIM has 27795 mapped keys (of 54675 keys)
hgu133plus2PATH has 11297 mapped keys (of 54675 keys)
hgu133plus2PATH2PROBE has 214 mapped keys (of 214 keys)
hgu133plus2PFAM has 39581 mapped keys (of 54675 keys)
hgu133plus2PMID has 40160 mapped keys (of 54675 keys)
hgu133plus2PMID2PROBE has 276342 mapped keys (of 283543 keys)
hgu133plus2PROSITE has 39581 mapped keys (of 54675 keys)
hgu133plus2REFSEQ has 40248 mapped keys (of 54675 keys)
hgu133plus2SYMBOL has 40801 mapped keys (of 54675 keys)
hgu133plus2UNIGENE has 40632 mapped keys (of 54675 keys)
hgu133plus2UNIPROT has 37123 mapped keys (of 54675 keys)
```

Additional Information about this package:

```
DB schema: HUMANCHIP_DB
DB schema version: 2.1
Organism: Homo sapiens
Date for NCBI data: 2010-Sep7
Date for GO data: 20100904
Date for KEGG data: 2010-Sep7
Date for Golden Path data: 2010-Mar22
Date for IPI data: 2010-Aug19
Date for Ensembl data: 2010-Aug5
```

Bioconductor Experiment Data Packages

Experiment data packages contain published data pre-prepared for Bioconductor. For example these packages include data from leukemia gene expression studies (Golub et al., 1999), colon cancer gene expression profiling (Alon et al. 1999), etc. Many of the tutorials (vignettes) in Bioconductor use these data in exercises.

Bioconductor and R Help

The first place to find help on R and Bioconductor is their website. These provide an excellent source of information.

In particular it is worth examining the manuals, FAQ and searchable archives of the mailing lists. **Do subscribe to the Bioconductor and R mailing lists.** When subscribing click the option “daily digest” as these mailing list as busy with >20 emails a day and could quickly fill up your email inbox. The daily digest option, send 1 email a day which contains all the correspondence from that day.

R Help

My favorite beginners guide to R is from Emmanuel Paradis. Its available from http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

Introduction to R classes and objects <http://cran.r-project.org/doc/manuals/R-intro.html>

Tom Short's R reference card (<http://cran.r-project.org/doc/contrib/Short-refcard.pdf>) and other contributed are useful <http://cran.r-project.org/other-docs.html>

I have taught several R courses (Bio503, HSPH) and the course notes are online. I hope you find them useful also. The 2011 course is available at <http://sites.harvard.edu/icb/icb.do?keyword=k76038> but I will this the most recent course from my HSPH website homepage <http://www.hsph.harvard.edu/research/aedin-culhane/>

Bioconductor Help

I found the Bioconductor course and workshop examples very useful when I was learning. These are available at <http://www.bioconductor.org/workshops>

For more information on getting started in R and Bioconductor, please see Vince Carey's guide at <http://bosbioc.wordpress.com/>

Thomas Girke, UC Riverside has also written excellent online tutorials on R and bioconductor available from <http://manuals.bioinformatics.ucr.edu/>

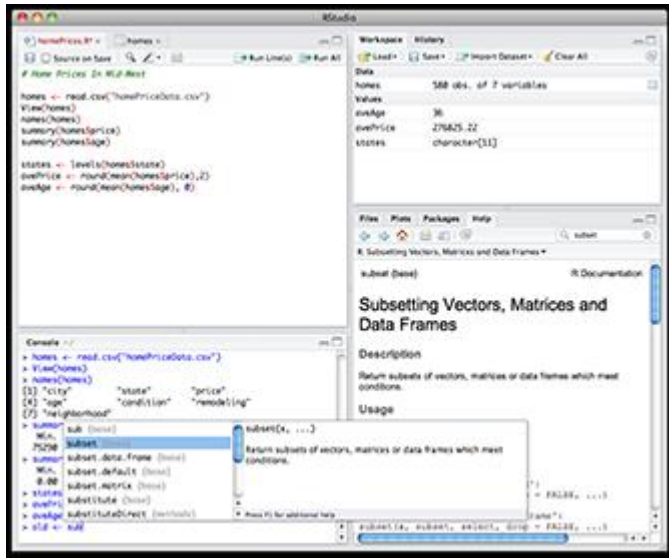
Interfaces for R

I have used all of these below at one stage or another and all color R code, highlighting brackets and are useful

Rstudio (my current favorite)

Cross platform- available on Windows, Linux, MacOS etc

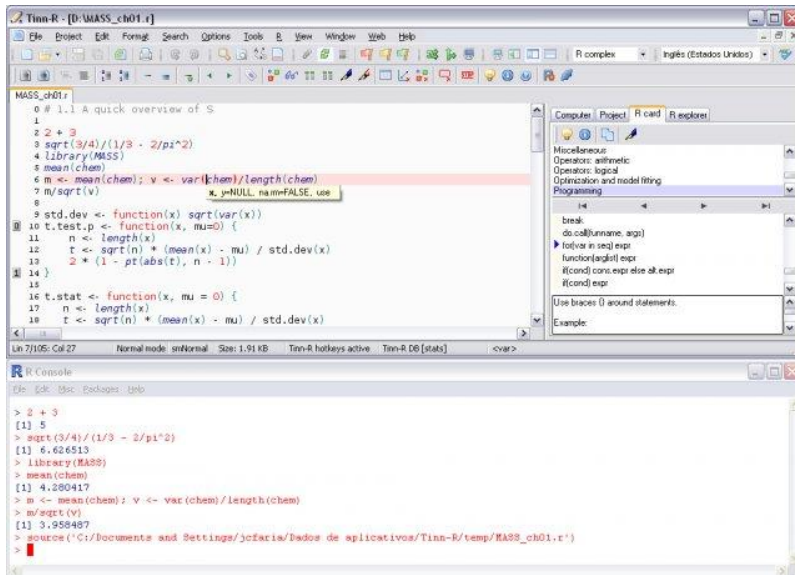
<http://www.rstudio.org/>



TinnR

Only available on Windows

<http://www.sciviews.org/Tinn-R/>

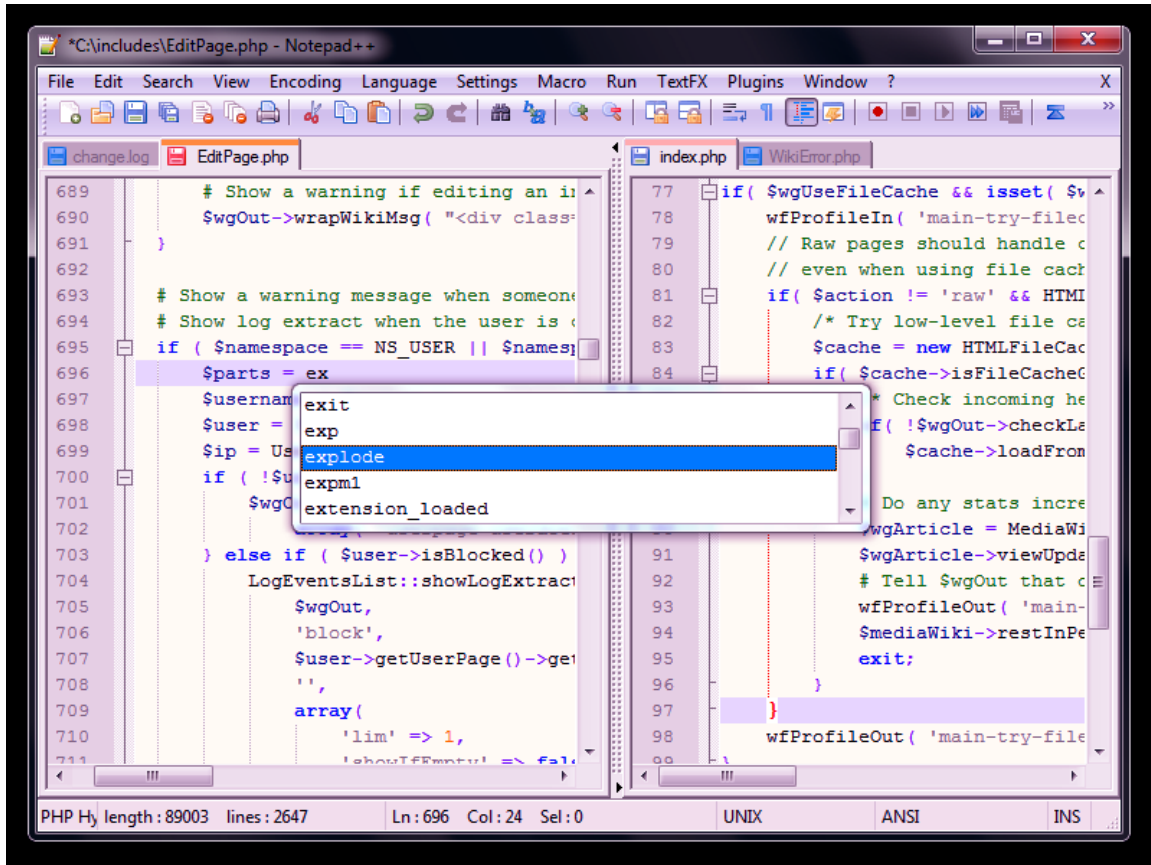


NotePad++ (with the plugin NpptoR)

Only available on Windows

<http://notepad-plus-plus.org/>

<http://sourceforge.net/projects/npptor/>



On Linux I have also used **Kate** the KDE text editor which has inbuilt highlighting of R code. On Mac, many users simply use the standard R GUI which is pretty good and has some features that are better than the version that comes on Windows/Linux.