

---

## Lecture 1:6 BioConductor, Microarrays and Genomics

Day 1: August 16, 2004: 15:45 – 17:15

Ben Bolstad, Biostatistics, University of California, Berkeley

---

### Key Concepts

- Introduce the BioConductor project
- Describe the Affymetrix microarray technology
- Outline some methods for pre-processing Affymetrix microarrays

### What you will be able to do at end of this section

- Gain a basic understanding of BioConductor and the tools it provides

### Overview

This lecture will introduce the BioConductor project, which is an open source, collaborative effort to produce an environment for the analysis of Genomic data. It is based on the open source statistical programming language R. An overview of the project along with some examples from specific packages will be given.

The Affymetrix GeneChip® is a commercially microarray system produced by Affymetrix. This talk will give a brief overview of the technology. More details can be found in the publications listed at <http://www.affymetrix.com/community/publications/foundation.affx>

The talk will conclude with some discussion of how Affymetrix GeneChip® microarrays may be pre-processed using the BioConductor tools.

### Appendix

#### 1. Resources

##### i) Original Papers

- R. C. Gentleman, V. J. Carey, D. J. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F.

Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. K. Smyth, L. Tierney, Y. H. Yang, and J. Zhang (2004) [Bioconductor: Open software development for computational biology and bioinformatics](#). *Genome Biology*. In press (pre-print: <http://www.bepress.com/bioconductor/paper1/>)

**ii) Software**

- BioConductor
  - o The open source R based environment for analysis of Genomic Data
- dChip
  - o Another program for the analysis of Affymetrix data. <http://www.dchip.org>
  - o Free for non commercial users

**iii) Text books:**

- Statistical Analysis of Gene Expression Microarray Data (2003) Edited by Terry Speed Chapman & Hall/CRC
- The Analysis of Gene Expression Data (2003) Edited by Giovanni Parmigiani et al
- Introductory Statistics with R (2002) by Peter Dalgaard
- Modern Applied Statistics with S (2002) by William Venables and Brian Ripley

**iv) Web Sites:**

- <http://www.bioconductor.org>
  - o The main BioC website.
- <http://www.r-project.org>
  - o This is the main website for the R-project. This is the open source statistical language/program upon which Bioconductor is based.

**2. Presentation Overheads**

**3. Instructions for installing R and BioConductor on your own system**

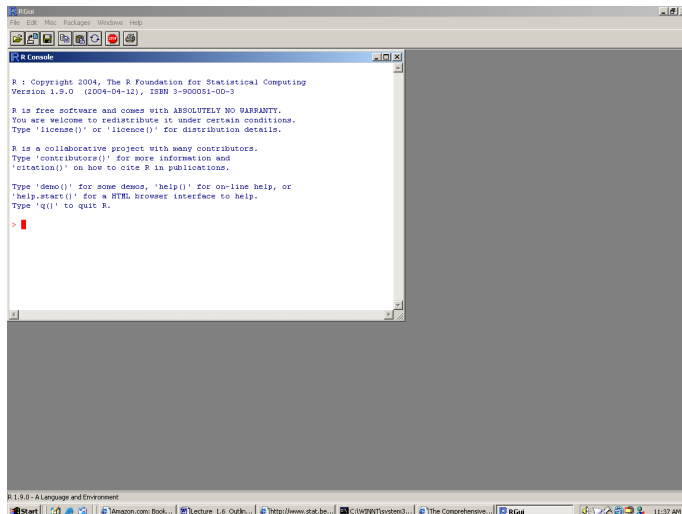
This set of instructions is for your later reference. You should not need to follow these instructions at the current time.

## 3.1 Downloading and installing R

The R webpage is <http://www.r-project.org>. At this website you will find information about the R program and many additional packages.

### 3.1.1 Windows Systems

1. You need to download the Windows installer from CRAN. Go to either <http://cran.r-project.org> or one of the US mirrors <http://cran.us.r-project.org/> <http://cran.stat.ucla.edu/>, or the Canadian mirror <http://probability.ca/cran/>.
2. Click on Windows (95 and later)
3. Click on Base
4. Click on rw1091.exe and save the download at a location you can find.
5. When it has finished downloading, double click rw1081.exe to start the setup program.
6. Follow the on screen setup instructions. This should pretty much be a matter of clicking next through a series of screens.
7. Find R either via the start menu or the icon on the desktop.
8. You should now see something similar to the following on your screen. Congratulations you are done installing R.



```
R : Copyright 2004, The R Foundation for Statistical Computing
Version 1.9.0 (2004-04-12), ISBN 3-900051-00-3

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.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

>
```

### 3.1.2 Linux Systems

On Linux you have two options. One is to download a pre-compiled version for your specific distribution. I will give details for the second option, which is to download and compile the source code for yourself.

1. You need to download the source code from CRAN. Go to either <http://cran.r-project.org> or one of the US mirrors <http://cran.us.r-project.org/> <http://cran.stat.ucla.edu/> or the Canadian mirror <http://probability.ca/cran/>
  2. Click on R-1.9.1.tgz to download the source code and save it somewhere.
  3. Go to the location you saved the downloaded file and type `tar xzvf R-1.9.1.tgz`. This will uncompress and extract the R source code.
  4. Change into the source code directory `cd R-1.9.1/`
  5. You might want to read the INSTALL file, but this is not completely necessary.
  6. If you have root access on your machine just type `./configure`. If don't have root access and you have sufficient disk space somewhere type `./configure --prefix=/path/to/install/location`
- where of course you replace `/path/to/install/location` with your install location.
7. Now type `make`. It will start compiling. This might take awhile, depending on the speed of your machine.
  8. Type `make install`. This will install R.
  9. You may need to add the bin subdirectory of your install location to your path. Use `setenv` (csh/tcsh) or `export` (bash) to do this.
  10. type `R` at the command-line.
  11. If all goes well you should have a working R installed.

### 3.1.3 Installing additional R packages

There may come a time where you want to install an additional package to your R installation because the base install does not have a function that you need. You can find many packages on CRAN.

On Windows you use the “packages” menu. You have two options:

1. Either download the file from CRAN (make sure you get the file with the “.zip” extension) and use the “install from local zip file” option
- or
2. Choose the “Install package(s) from CRAN” and select the package you want (it will be downloaded and installed automatically).

On Linux/UNIX machines you use `R CMD INSTALL` to install packages. eg

```
R CMD INSTALL packagename 1.0.0.tar.gz
```

You may need to set the R `LIBS` environment variable.

### 3.2 Installing BioConductor

Once you have installed R you need to install BioConductor. While it is possible to download each package individually and manually install the packages one by one the recommended procedure is to use an installation script that will handle all the dependencies for you automatically. You can download and execute the installation script entirely within R.

At the R prompt type

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

which will download and load the `getBioC` script into your current R session. Next it is time to execute the script and get it to download and install BioConductor. Typing

```
getBioC("all")
```

will download and install all the BioConductor packages from the current release. Note that this is a lot to download and may take a long time depending on the speed of your connection. If you want to restrict your installation to a smaller subset of packages you could instead use either

```
getBioC("cdna")
```

or

```
getBioC("affy")
```

which will only the core packages for the analysis of cDNA or Affymetrix data (respectively).

Once you have installed BioConductor you start it by loading the package(s) you wish to use. For example to load the *affy* package

```
library(affy)
```