Introduction to Bioconductor and object oriented programming in R

Héctor Corrada Bravo
based on slides developed by
Jim Bullard, Kasper Hansen, Margaret Taub

PASI, Guanajuato, México
May 3-4, 2010

To follow along...

- Slides and handouts here:
  http://www.biostat.jhsph.edu/~hcorrada/PASI_2010/
- Start R, and run the following:

```
> source("http://bioconductor.org/biocLite.R")
> biocLite(c("affy", "genefilter",
+     "SpikeInSubset", "limma"))
```

Why

- ▶ Biological experiments are continuously generating more and more data!
- ▶ It has become impossible to analyze a high-throughout biological experiment without statistics and bioinformatics.
- ▶ Different research groups are constantly rewriting essentially the same software for only slightly different purposes.
- ▶ With some exceptions (e.g. BLAST), we often find ourselves writing software which has probably already been written (usually at least twice).
- ▶ Bioconductor is an open-source / open-development set of tools which can be widely employed in a number of genetic and biomedical settings.

Bioconductor: Overview

- ▶ Bioconductor was started in the Fall of 2001
- ▶ The core maintainers and hosts of the Bioconductor websites are located at Fred Hutchinson Cancer Research Center
- ▶ A new version of Bioconductor is released twice-anually and is linked to the release of R
- ▶ Bioconductor makes heavy use of the S4 class system of R. This often makes it more complicated than simpler R packages to work with.
- ▶ Bioconductor has been most successful in the analysis of Microarray data – but has features for other types of data such as flow cytometry and emerging tools for next-gen sequencing.
- ▶ R is written almost exclusivey by Statisticians; although it is a very general environment for computing most often we think of R as a statistical programming environment.

Bioconductor: Goals

1. Provide access to statistical and graphical tools to perform analysis in a number of bioinformatics settings.
2. Provide a framework for extending the components of the system to customize the environment for particular settings.
3. Provide a comprehensive set of documents describing the system and how to extend/interact with the components.
4. Provide tools to interact with publicly available databases as well as other sources of meta-data. [1]

---

[1] These goals were adapted from W. Huber's slides

Website

Lets take a small website tour to see what's there. `bioconductor.org`

Installation

- *Always* install using biocLite. It will check your R version etc.

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

- This will install package Biobase. To install a standard set of packages, including most of those used today, use

  ```
  > biocLite()
  ```

- Compared to most CRAN packages, Bioconductor packages typically has a much more complicated set of dependencies.

Resources

- Mailing lists "BioC" (medium volume), "Bioc-devel" (very low volume), "Bioc-sig-seq", all of them friendly.
- If you email about a problem, *always* include the output of `sessionInfo` in a session where the error/problem occurred. If there is an error, a `traceback` is often useful. Search the archive first. Sometimes it helps to know things such as 32/64 bit, was the package installed from source or binary, what specific OS. Read the posting guide!
- Vignettes are great! Use vignette (base) or (my preference) openVignette (Biobase). Sometimes they are *very* extensive.

Some books

- Gentleman (2008): "R programming for Bioinformatics".
- Hahne, Huber, Gentleman, and Falcon (2008): "Bioconductor Case Studies".
- Gentleman, Carey, Huber, Irizarry, and Dudoit (2005): "Bioinformatics and Computational Biology Solutions Using R and Bioconductor".

Bioconductor books tends to age quickly.
There are several good R books out there. Springer has a (cheap) series of softcovers on R.