

Introduction to R and Bioconductor

Alex Sánchez

Unitat d'Estadística i Bioinformàtica (UEB)

June 5, 2015

Table of Contents

A bit of interaction

- ▶ What is your R knowledge, on a 0(beginner) to 2 (expert) scale?
- ▶ How deep is your knowledge with R packages related to NGS, on a 0(none) to 2 (good)scale?
- ▶ What analyses do you plan to do in R?

Objectives

- ▶ Quick review of *R* history and capabilities
- ▶ Overview of the Bioconductor project
- ▶ Bioconductor classes and methods for NGS

What is R?

1. an implementation of the S language (Bell Laboratories, Rick Becker, John Chambers and Allan Wilks)
2. R is an integrated suite of software for
 - ▶ data manipulation
 - ▶ calculation and
 - ▶ graphical display.

What is R?(c'ed)

1. R is a vehicle for newly developing methods of interactive data analysis
 - ▶ develops rapidly
 - ▶ is being extended by a large collection of packages
 - ▶ Comprehensive R Archive Network (CRAN)
 - ▶ Bioconductor
2. However, most programs written in R are essentially ephemeral, written for a single piece of data analysis

R specifics

- ▶ a suite of operators for calculations on arrays, in particular matrices.
- ▶ an “environment”:
 - ▶ a fully planned and coherent system,
 - ▶ it can be *saved*, *loaded*, *exchanged*.

R and statistics

- ▶ R is an environment...
 - ▶ originally not designed for statistics,
 - ▶ many classical and modern statistical techniques implemented,
- ▶ Difference with S, S-plus, SAS or SPSS...
 - ▶ minimal output
 - ▶ minimal number of object (in comparison with the).

R and the window system

- ▶ R comes with a graphical system on all platform
 - ▶ console like: Unix
 - ▶ GUI and console: Mac, Windows
- ▶ Integrated Developer Interfaces (IDE) have been developed
 - ▶ StatET plugin (<http://www.walware.de/goto/statet>) for eclipse
 - ▶ **Rstudio** (<http://rstudio.org>)

Using R interactively

- ▶ Most of the time R is used interactively
- ▶ R console is very similar to Unix/linux
 - ▶ `ls` command for listing,...
 - ▶ The syntax is only slightly different:
 - ▶ `ls()` instead of `ls`
- ▶ Documentation and help pages always available:
 - ▶ through the “?” command (*perfect match*)
 - ▶ through the “?” command (*fuzzy matching*)
 - ▶ through `help.start()` if you have a windows system
 - ▶ searchable through `help.search()`

CRAN

- ▶ `http://cran.r-project.org`
- ▶ The comprehensive R Archive
 - ▶ 5578 packages! (26 May 2014)
 - ▶ easy to install
 - ▶ R CMD INSTALL (cmd line)
 - ▶ `install.packages()` (from within the environment)

What is Bioconductor

- ▶ A software project for the analysis of genomic data
- ▶ **Open** source and **open** development.
- ▶ <http://bioconductor.org>
- ▶ A collection of R packages with *some* common structures
 - ▶ >1.100 packages (554 soft., 600 annot.)
 - ▶ >300 developers, >4.000 citations
- ▶ Gentleman et al. *Bioconductor: open software development for computational biology and bioinformatics*. Genome Biology (2004) vol. 5 (10) pp. R80

Bioconductor: history and overview

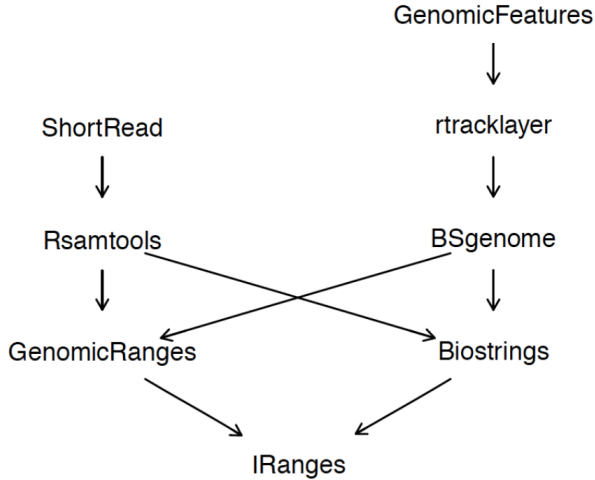
- ▶ Started in Harvard (2001) now hosted at Fred Hutchinson Cancer Research Center (FHCRC)
- ▶ Gentleman et al. *Bioconductor: open software development for computational biology and bioinformatics*. Genome Biology (2004) vol. 5 (10) pp. R80
- ▶ Focus on **Microarray** at first,
- ▶ and on **Next Generation Sequencing** as of 2008.

Bioconductor goals

- ▶ Provide access to powerful **statistical and graphical methods** for the analysis of biomedical and genomic data
- ▶ Facilitate the integration of **biological metadata** from WWW in the analysis of experimental data (e.g., GenBank, GO, LocusLink, PubMed)
- ▶ Allow the rapid development of **extensible, interoperable and scalable** software
- ▶ Promote **high-quality documentation** and **reproducible research**
- ▶ Provide **training** in computational and statistical methods

FHCRC, BIOC core packages

- ▶ Input and Output
 - ▶ rtracklayer, **Rsamtools**, **ShortRead**
- ▶ Sequence manipulation
 - ▶ **Biostrings**
- ▶ Range-based manipulations:
 - ▶ **IRanges**, **GenomicRanges**
- ▶ Annotations
 - ▶ **GenomicFeatures**, AnnotationDbi, BSgenome



53 Contributed packages (Sep. 2012)

- ▶ Chip-seq(14)
 - ▶ BayesPeak, CSAR, ChIPpeakAnno, ChIPseqR, ChIPsim, PICS, chipseq,...
- ▶ RNA-seq(18)
 - ▶ DEGseq, DESeq, Genominator, baySeq, edgeR, srnaSeqMao, goseq, gage, easyRNASeq,...
- ▶ **Infrastructure:**
genomeIntervals, girafe, cqn
- ▶ **base calling:** Rolexa
- ▶ **Visualization:**
HilbertVis, HilbertVisGUI
- ▶ **motif:** MotIV, rGADEM
- ▶ **domain-specific:**
MEDIPS, OTUbase, R453Plus1Toolbox
- ▶ **database:** SRADB, oneChannelGUI
- ▶ **smRNA:** segmentSeq

Installation

Two step installation

- ▶ First install R software: download from CRAN (www.cran.r-project.org)
- ▶ Install bioconductor
 - ▶ Download installer from Bioconductor website
 - > `source("http://bioconductor.org/biocLite.R")`
 - ▶ Make default installation (installs some basic and some popular packages)
 - > `biocLite()`
 - ▶ Add what you specifically need
 - > `biocLite(goProfiles)`