RNA-seq analysis using Bioconductor

Unit *nn*:

Alex Sánchez asanchez@ub.edu

Statistics Department. University of Barcelona (UB) *and*
Statistics and Bioinformatics Unit. Vall d'Hebron Institut de Recerca (VHIR)

June 3, 2014

UNITAT
D'ESTADÍSTICA I
BIOINFORMÀTICA

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?

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
   - devolops rapidly
   - is being extenden 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
  - can be saved, loaded,exchanged

R and statistics

- R is an environment
  - not designed for statistics
  - many classical and modern statistical techiniques implemented
- R is an environment
  - Differnece with S,S-plus,SAS and SPSS
  - minimal ouput
  - minimal number of object

R and the window system

- R comes with a graphical system on all plataform
  - console like: Unix
  - GUI and console: Mac, Windows
- Integrated Devoeloper Interface (IDE) have been developed
  - StatET plugin (`http://www.walware.de/goto/statet`) for eclipse
  - Rstudio (`http://rstudio.org`)

Using R interactively

- R enviroment is vey similar to Unix
  - ls command for listing,...
  - The syntax is only slightly different:
    - ls () instead of ls
- Documentation and help pages always avaliable:
  - through the "?" command (perfect match)
  - through the "?" command (fuzzy matching)
  - through hel.start() if you have a windows system
  - searchable through help.search()

CRAN

- The comprehensive R Archive
  - 5578 packages! (26 May 2014)
  - easy to install
    - R CM INSTALL (cmd line)
    - install.packages (from within the environment

Bioconductor: history and overview

- Gentleman et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology (2004) vol. 5 (10) pp. R80
  - Fred Hutchinson Cancer Research Center (FHCRC)
- A set og packages developed for the analysis and comprehension of high throughput genetic data
  - ¿1.100 packages (554 soft.,600 annot.)
  - ¿300 developers, ¿4.000 citations
- Focus on microarray at first, and on Next Generation Sequencing as of 2008.

FHCRC,BIOC core packages

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

- DIBUIX!

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

Outlook

- Numerous packages,bioC or third-party
- Complex, evolving infrastructure
- Very active community
  - an R release every year
  - a Bioc release every 6 months
  - bioconductor website
    - `http://bioconductor.org`
  - bion mailing lists
    - bioc-sig-sequencing
    - bioc-devel (heavy traffic)