

Introduction to *R* and *Bioconductor*

Martin Morgan (mtmorgan@fhcrc.org)
Fred Hutchinson Cancer Research Center
Seattle, WA, USA

June 23, 2014

R

R is a language and environment for statistical computing and graphics

R

R is a **language and environment** for statistical computing and graphics

- ▶ Full-featured programming language
- ▶ Interactive and *interpreted* – convenient and forgiving of user errors
- ▶ Coherent, extensively documented

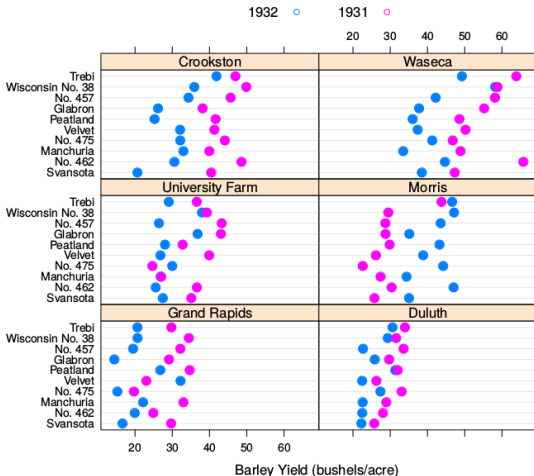
R

R is a language and environment for **statistical computing** and graphics

- ▶ Throughout the language, e.g., `factor` and `NA`
- ▶ Built-in statistical functionality
- ▶ Highly extensible via user-contributed *packages*

R is a language and environment for statistical computing and graphics

- ▶ Explore data
- ▶ Communicate results



R vectors, classes, and functions

- ▶ Vectors

- ▶ logical, integer, numeric, complex, character, raw (byte)
- ▶ factor: discrete levels
- ▶ Missing-ness, NA

- ▶ *data.frame*, *matrix*, and other *objects*

- ▶ Functions

- ▶ Operating on vectors, e.g., `log`, `lm` (fit a linear model)
- ▶ 'Higher order' functions – apply a function to several different vectors, e.g., `lapply(df, log)`

- ▶ Packages

None of this making sense? *R* introduction / refresher tutorial this afternoon

Using R

Documentation

- ▶ `help()`
- ▶ vignettes

Work flows

- ▶ Scripts...
 - ▶ Reproducible
 - ▶ Literate
- ▶ ... mature to *packages*
 - ▶ Coordinate data, analysis, and documentation
 - ▶ Share with others

Bioconductor project goal

Analysis and comprehension of high-throughput genomic data

Bioconductor project goal

Analysis and comprehension of high-throughput genomic data

Statistical analysis

- ▶ Reduce large data to manageable knowledge
- ▶ Cope with technological artifacts
- ▶ Rigorous exploration
- ▶ Designed experiments, e.g., treatment vs. control
- ▶ Leading-edge methods for leading-edge questions

Bioconductor project goal

Analysis and **comprehension** of high-throughput genomic data

- ▶ Understandable
- ▶ Reproducible
- ▶ Effective visualization
- ▶ Biological context, e.g., annotation
- ▶ Training

Bioconductor project goal

Analysis and comprehension of **high-throughput genomic data**

- ▶ Sequencing: RNA-seq, ChIP-seq, variants, copy number...
- ▶ Microarrays: expression, SNP, ...
- ▶ Flow cytometry
- ▶ Proteomics
- ▶ Images
- ▶ ...

What is *Bioconductor*?

Collection of packages in the *R* statistical programming language

- ▶ Developed by the *Bioconductor* core and international contributors
- ▶ Stable 'release' branch, and leading edge 'devel' branch
- ▶ Open source / open development

Used by...

- ▶ Individuals
- ▶ Academic labs & research groups
- ▶ Government agencies
- ▶ Pharma and other companies

How to learn & use *Bioconductor*

1. Install *R* (& *RStudio*?)
2. Identify and install packages
3. Write *R* scripts
 - ▶ Input & ‘massage’ data
 - ▶ Quality assessment
 - ▶ Statistical analysis
 - ▶ Visualization
 - ▶ Annotation
 - ▶ Reports & summaries
4. Share with colleagues, collaborators, and the community

<http://bioconductor.org>

The screenshot shows the Bioconductor website homepage. At the top, there is a search bar and a navigation menu with links for Home, Install, Help, Developers, and About. The main content area is divided into several sections:

- Annual Conference:** A section titled "Annual Conference" with a sub-heading "Register Now!" for BIC 2014. It mentions the conference is held from July 30 - Aug 1, 2014 at the Dana-Farber Cancer Institute, Boston, MA, and notes that NEW reduced hotel rates are available.
- About Bioconductor:** A section titled "About Bioconductor" describing it as a platform for genomic data analysis and R-based bioinformatics. It mentions that Bioconductor uses the R statistical programming language and is open source and open development. It also notes that it has two releases each year, is available as an Amazon Machine Image (AMI), and has an active user community.
- Install:** A section titled "Install" with a sub-heading "Get started with Bioconductor". It lists links for "Install Bioconductor", "Custom packages", "Helping JAR", "Latest newsletter", "Follow us on Twitter", and "Using R".
- Learn:** A section titled "Learn" with a sub-heading "Master Bioconductor tools". It lists links for "Recent coverage", "Package vignettes", "Custom work flows", "FAQ", and "Community resources".
- Use:** A section titled "Use" with a sub-heading "Create bioinformatic solutions with Bioconductor". It lists links for "Software, Annotation, and Experiment packages", "Amazon Machine Image", "Latest release announcement", and "Subscribe to the mailing list".
- Develop:** A section titled "Develop" with a sub-heading "Contribute to Bioconductor". It lists links for "Use BioC 'dev'", "DevEl Software, Annotation, and Experiment packages", "Package guidelines", "New package submission", "Advanced programming", "Additional developer resources", and "Build reports".

- ▶ Established work flows, e.g., RNA-seq differential expression with *DESeq2*
- ▶ Flexible bioinformatic analysis, e.g., ...

Project strengths

- ▶ **Extensive**
 - ▶ Respected
 - ▶ Well-used
 - ▶ Accessible
- ▶ 824 software packages, 867 annotation packages, 202 experiment data packages
 - ▶ Sequencing, microarrays, flow cytometry, proteomics, image analysis, ...
 - ▶ All packages with vignettes and help pages
 - ▶ Tutorials, training material, national and international conferences

Project strengths

- ▶ Extensive
- ▶ **Respected**
- ▶ Well-used
- ▶ Accessible

“Community repositories that carry out testing are ideal. . . the genetics community is fortunately familiar with the Comprehensive R Archive Network and the principles of stewardship of modular software embodied in the Bioconductor suite. . . The journal has sufficient experience with these resources to endorse their use by authors.” – *Nature Genetics* 46, 1 (2014)

Project strengths

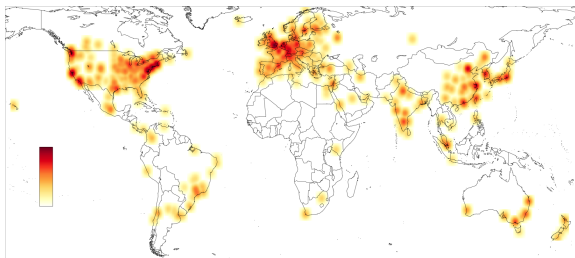
- ▶ Extensive
- ▶ **Respected**
- ▶ Well-used
- ▶ Accessible

PubMedCentral full-text citations

		Citations
<hr/>		
<i>Bioconductor</i>		9070
RNA-seq		
<i>edgeR</i>	Diff. expression	647
<i>DESeq</i>	Diff. expression	648
Microarray		
<i>affy</i>	Pre-processing	2318
<i>limma</i>	Diff. expression	4503
<i>GOstats</i>	GSEA	436

Project strengths

- ▶ Extensive
- ▶ Respected
- ▶ **Well-used**
- ▶ Accessible



- ▶ 225,000 unique IP addresses downloaded
9.3M packages
- ▶ 397,000 site visitors / year (27% increase)
viewed 2.8M pages
- ▶ ~ 600 mailing list posts from ~ 210
authors per month

Project strengths

- ▶ Extensive
 - ▶ Respected
 - ▶ Well-used
 - ▶ **Accessible**
- <http://bioconductor.org>
- ▶ Package vignettes & help pages
 - ▶ Work flows
 - ▶ Mailing list & 'guest posting' facility
 - ▶ Courses and other training
 - ▶ Annual Conference,
Boston July 30 – Aug 1.

Acknowledgements

- ▶ *Bioconductor* core: Vince Carey, Sean Davis, Kasper Hansen, Wolfgang Huber, Robert Gentleman, Rafael Irizzary, Michael Lawrence, Levi Waldron
- ▶ *Bioconductor* team: Sonali Arora (introductory material, copy number), Marc Carlson (annotation), Nate Hayden (pileup, C++), Valerie Obenchain (variants, ranges), Hervé Pagès (ranges, strings), Paul Shannon (systems biology), Dan Tenenbaum (web, build)
- ▶ The international *Bioconductor* community!
- ▶ Funding: US NHGRI / NIH U41HG004059; NSF 1247813.

More: <http://bioconductor.org>