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 *interpretted* – convenient and forgiving of user errors
- Coherent, extensively documented

R

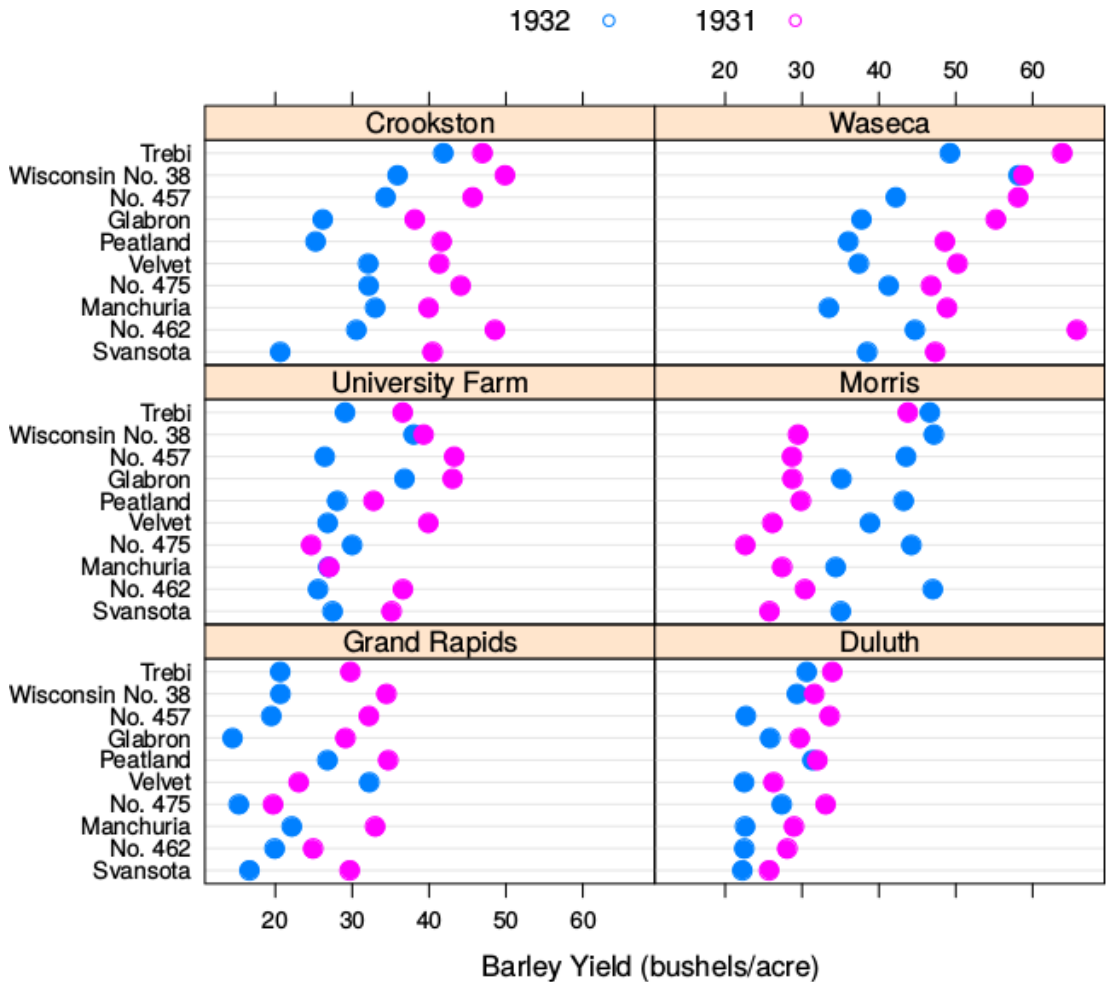
R is a language and environment for **statistical computing** and graphics

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

R

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

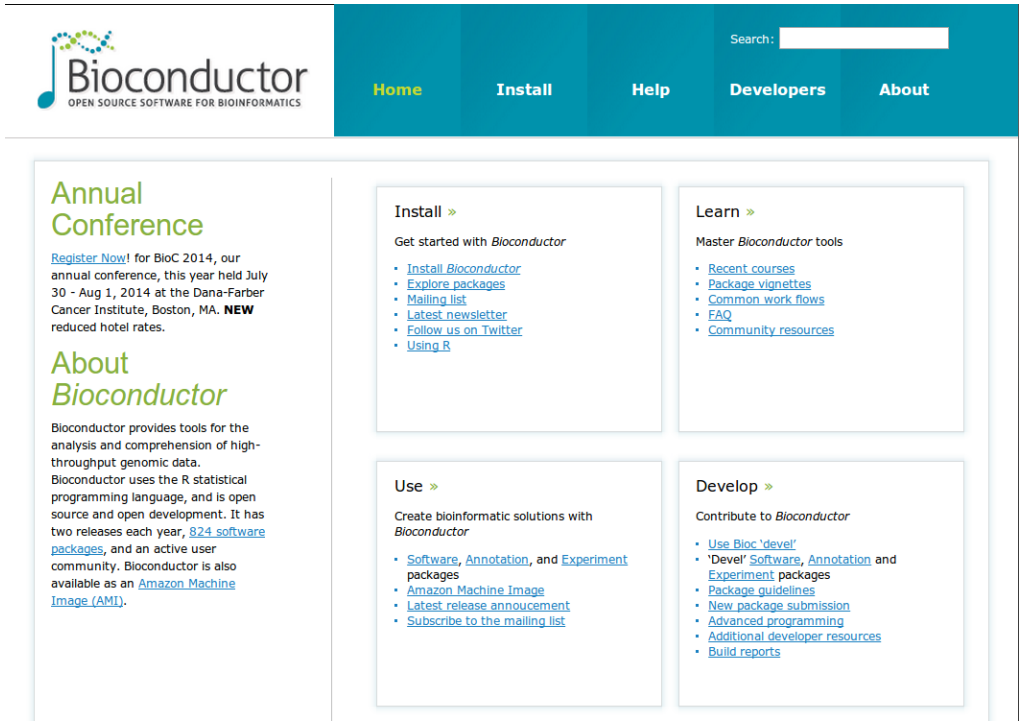
Used by...

- 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`



- Estabished 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

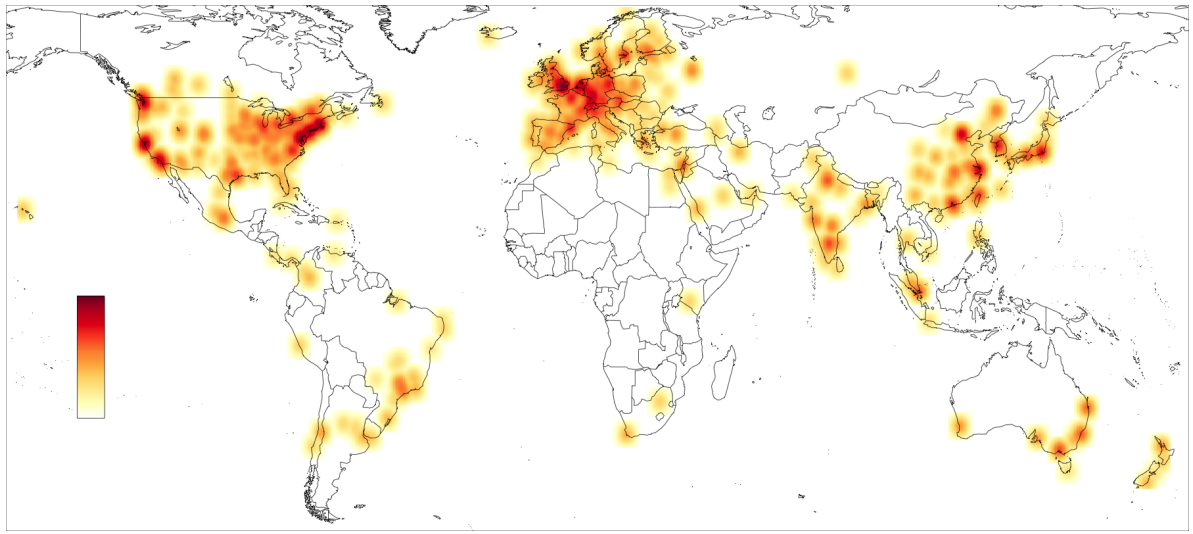
- Extensive
- **Respected**
- Well-used
- Accessible

PubMedCentral full-text citations

		Citations
Bioconductor		9070
RNA-seq		
edgeR	Diff. expression	647
DESeq	Diff. expression	648
Microarray		
affy	Pre-processing	2318
limma	Diff. expression	4503
GOstats	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
- $\sim$ 600 mailing list posts from $\sim$ 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`