Approaches to Package Management – *Bioconductor*

Martin Morgan (Martin.Morgan@RoswellPark.org)
Roswell Park Cancer Institute
Buffalo, NY, USA

3 July, 2017

Bioconductor

"Analysis and comprehension of high-throughput genomic data"

- Established 2002; 1383 packages (core team and contributed).
- Well-respected, cited (20k PubMedCentral full-text citations), used (>350k unique IP addresses / year).
- https://bioconductor.org; https://support.bioconductor.org
- CRAN-style repository. Cloud front content delivery (plus a few mirrors maintained for local purposes).
- Primarily supported through US NIH.

Release cycle

Six-month releases

- 'Devel': new packages and features.
- 'Release': end-users.

Which *R*? The one end-users see.

- Now: release and devel both on *R*-3.4.
- October: release on *R*-3.4, devel on *R*-devel.

Comments

- Cohesive packages – deep dependency graph.
- Enables change – breakage in devel tolerated.
- *BiocInstaller*`::biocLite()` to manage repositories seen by `install.packages()`.
- *R*-devel not always exposed to *Bioconductor* packages.

Package management

Version-controlled packages.

- All packages under SVN; individual developer accounts.
- Versioning scheme `x.y.z`. `y` even in release, odd in devel. Each commit bumps `z`.
- Will discuss GIT in a second...

Comments

- Mostly package developer commits, but core team can step in.
- Eases incorporation of breaking (ours, CRAN, *R*) changes

Nightly builds

- `R CMD build` / `check`;
- Cross-platform; release & devel.
- SVN snapshot; all packages.
- Successful builds get pushed to public repositories.

Comments

- 'Continuous integration', sort of.
- Sometimes 'impossible' public repositories
  - A introduces feature that breaks B. A pushed but old B still available.
  - B depends on feature in newest A. A builds and installs (so used by B) but fails check (so not pushed). B builds & checks so pushed.

New packages I

Submission – open and reviewed.

- Maintainer posts a public Github issue.
- Moderated (manual) – is it a legitimate package?
- Built and checked. Usually, maintainers iterate until 'OK'.
- Assigned reviewer (core team; implementation), plus community input (implementation, science).
- Goal: incremental improvement, rather than absolute standard.

New packages II

Comments

- Wide range of quality.
- Time consuming and sometimes uninspiring; hard to standardize across reviewers.
- Maybe 80% use *roxygen2* (and probably *devtools*).
- Common issues: *Bioconductor* interoperability; documentation; *R* code.

New packages III

Common issues: *R* code

- Generally, iteration instead of vectorization (tell-tale sign: use of parallel evaluation).
- Robustness
  - `1:n` (vs. `seqlen(n)`).
  - `if (<scalar binary logical>) {}` (challenging!)
- 'Copy-and-append' `x = numeric(); for (i in 1:n) x <- c(x, i)`
- Vocabulary `apply(x, 2, sum)` vs. `colSums(x)`.
- Hoisting constant expressions out of loops.
- Cyclomatic complexity.

Software management

Currently...

- SVN repository
- git / svn 'bridge' – sync git repositories with git.
- About 1/2 commits via git / svn bridge.

Migrating to git

- `git clone https://git.bioconductor.org/packages/BiocGenerics`

Challenges

- Cannonical location / distributed support & social environment.

Acknowledgments

- Core: Valerie Obenchain, Hervé Pagès, Lori Shepherd, Marcel Ramos, Nitesh Turaga, Daniel van Twisk.
- The research reported in this presentation was supported by the National Cancer Institute and the National Human Genome Research Institute of the National Institutes of Health under Award numbers U24CA180996 and U41HG004059. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

https://bioconductor.org,
https://support.bioconductor.org