aRchive: enabling reproducibility of Bioconductor package versions

Nitesh Turaga[1], Enis Afgan[1], Eric Rasche[2], Dannon Baker[1] and The Galaxy Team

[1]Johns Hopkins University, Baltimore, MD, USA. Email: nturaga1@jhu.edu
[2] Center for phage Technology, TAMU, College Station, TX, USA.

Project Website: bioaRchive.github.io
Source Code: https://github.com/bioarchive/aRchive_source_code
License: MIT

The Bioconductor suite provides bioinformatics tools in the form of R packages, which have frequent version upgrades. Once an upgrade takes place in a Bioconductor package, it is hard to retrieve previous versions from the source repository. One of Galaxy's primary goals is enabling the reproducibility of any analysis, without the user having to consider it. A major component enabling this is the Galaxy Tool Shed, which provides a host of tools and dependencies that can be installed in Galaxy instances to provide precise versions of tools to Galaxy users. The inability to retrieve specific previous versions of Bioconductor packages makes reproducibility of Bioconductor-based analysis difficult, if not impossible, in Galaxy (or elsewhere). Integrating support for multiple versions of Bioconductor packages within Galaxy would yield immediate improvement for reproducibility while using Bioconductor packages wrapped as Galaxy tools. To that end, we have implemented a method that provides this level of reproducibility for Bioconductor tools in the context of the Galaxy Tool Shed. To do this, we started with a copy of the publicly available, read-only subversion repository of all Bioconductor packages. We then traced through the commit history of each package, extracting released versions and stored them as independent documents. All the versions of all the packages have now been made available in a newly formed, public *aRchive* from where they can easily be retrieved. The *aRchive* is automatically maintained via a cron job that updates the repository with new package versions as they are released. This makes it possible for developers to easily obtain any version of a Bioconductor package required by a pipeline, and make the tool available in the Tool Shed. The *aRchive* ensures previously performed analyses can be reproduced down to the exact version of the initial software used. Thinking to the future, the *aRchive* could be integrated into Bioconductor itself via the *biocLite()* function, with an additional argument specifying a version number of the package. This talk will describe implementation details, results, and future work related to *aRchive* and how it bridges the gap between Bioconductor and Galaxy.