Teaching R/Bioconductor in Graduate Classes

BioC 2016

Thomas Girke

June 24, 2016

Type of Class

- Topic: *Data Analysis in Genome Biology*
- Objective: learn how to analyze NGS data independently and understand theory
- Format: lectures, tutorials, homework, student presentations and course projects
- Students are from different grad programs in genetics/bioinformatics, biosciences, statistics and CS
- Lectures cover theory
- Tutorials cover Linux command-line, computer cluster, R and Bioconductor
- Course projects focus on a NGS analysis problem with a final report written in R Markdown hosted in private GitHub repos

Teaching R and Bioconductor

- Basics of R straightforward to learn with easy to comprehend learning outcomes for students
- Teaching how to make good use of R packages is more complex task. Why?
  - Learning to work with objects of cores packages (*e.g. Genomic**) is challenging
  - Difficulty for students to solve errors on their own
  - Documentation and help system often too complex, too technical and overwhelming for beginners
  - Duplications among packages
- Clear learning benefit of introducing core packages: solving complex analysis problems
- Sometimes there are competitors that are easier to use. E.g. “why bother learning *GenomicRanges* if bedtools is all I need?”

Command-line vs. GUI Environment

- To work on remote systems efficiently, students have to learn how to work in a pure command-line interface.
- Too much time is spent on learning this basic skillset
- For command-line R, emacs/ESS and vim-r/tmux are still the best solutions but come with a steep learning curve.
- A command-line version of RStudio could provide some help?

Computations on a Computer Cluster

- Cluster computing also has a steep learning curve.
- BatchJobs/BiocParallel very helpful for students to learn parallel computing from within R
- Still students need to learn basics of managing job submissions from command-line to become proficient outside of R

R Markdown

- Easy to learn
- Valuable skillset for future research (*e.g.* efficiency and reproducibility)
- Also great help for instructor for developing teaching material (*e.g.* tutorials)
- A classroom package deploying classes on GitHub would be useful. Options could be:
    - Extension to BiocStyle (could support Bioc conference/workshop material)
    - R Markdown Websites
    - Various Jekyll-based solutions

GitHub is Another Game Changer

- Works well together with any data science course
- Very useful for hosting teaching material online and running a highly functional website for class
- Great help for student group projects
- Learning social coding skills
- Running entire class on GitHub works well with some help from third party solutions:
    - Google Sheets: for account sharing, even grading (with help from Jenny Bryan’s amazing googlesheet package), etc.
    - Piazza: online chat room for class