GCC 2018 Reproducible Research Short Course

Keith Baggerly

2018-06-04

Contents

1 Overview

Weclome to the 2018 GCC Short Course on Reproducible Research (RR)!

In this course, we'll discuss some of the motivation behind recent efforts to improve the reproducibility of published research, and work through a case study in detail to produce an instance of a reproducible analysis. To improve the sharing of code as well as analyses, we'll work through construction of our own R packages, introducing functions which can make our jobs easier.

This course presumes a working knowledge of R, and will make heavy use of R and RStudio.

PLEASE DOWNLOAD AND/OR INSTALL REQUIRED SOFTWARE, PACKAGES, AND DATA BEFOREHAND!

This will save time and (possibly) frustration.

2 What You'll Need

This course will involve much live demo on my part and experimentation on your part, so you'll need to be able to follow along to get the most out of it.

2.1 Equipment

You'll need a laptop with recent versions of the relevant software installed (please bring your power cords too!). You should have write permission to create files and folders on this laptop.

2.2 Software

2.2.1 Base

Recent versions of R and RStudio, downloadable from CRAN and the RStudio download page.

I'm currently running

- R version 3.5.0 (2018-04-23)
- RStudio 1.1.447

on my MacBook Pro laptop (OS X v10.13.4).

2.2.2 Packages

I'll be making use of the following packages (in alphabetic order, my version numbers are shown)

- `devtools` 1.13.5
- `downloader` 0.4
- `GEOquery` 2.47.18 - this is just for parsing one of the datasets I'll use for illustration, and isn't as directly germane to RR per se.
- `here` 0.1
- `knitr` 1.20
- `lattice` 0.20.35 - for example only
- `magrittr` 1.5
- `readr` 1.1.1
- `rmarkdown` 1.9
- `roxygen2` 6.0.1
- `tidyr` 0.8.1

A broader "package of packages" I may use without extensive discussion:

- tidyverse 1.2.1

All of the above are available from CRAN, with the exception of `GEOquery`.

`GEOquery` is available from Bioconductor.

2.2.3 Windows and RTools

For those of you running Windows machines, you'll also need to install

- Rtools

in order to get new packages you create to compile properly. Jeff Leek gives a slightly more expansive description of installation here.

2.2.4 Bonus

2.2.4.1 git and github

As we go along, I'll be saving my work using the git version control system, and posting materials to the course's repository on github so you'll all be able to access everything after the course. We won't discuss these tools in the course itself, but you may want to consider installing git on your machine and setting up an account on github if you find the illustrated usages interesting.

2.2.4.2 MikTeX / TeX for pdf output

RStudio will let us export reports in a variety of formats (e.g., html, docx, md), but some types of output require additional addins. In particular, producing pdf output makes use of pdflatex, which in turn requires that you have some version of TeX installed. The type of distribution to install varies by operating system, so I'd use what the LaTeX page suggests.

The use of pdflatex is discussed a bit more in RStudio posts on Customizing LaTeX Options and Using Sweave and knitr

2.3 Data

I'm going to use a few microarray datasets for illustration purposes, involving

- the NCI60 panel of cell lines,
- patient samples from a breast cancer study, and
- example data purported to be useful in predicting response to treatment

2.3.1 NCI60

The NCI60 is a panel of cancer cell lines which has been maintained by the National Cancer Institute (NCI) for the past 40-50 years. There are 59 cell lines in the panel (there were initially thought be 60, but it was later recognized that one cell line was an inadvertent duplicate of another). One of the things that's special about this panel is that drug sensitivity information (e.g., what concentration of a drug is required to inhibit normal growth of the cell line by 50%) is publicly available for all cell lines in the panel for almost all cancer chemotherapeutics now in use.

These cell lines have been profiled with several types of molecular assays over the years, results of several of which are available from the NCI's Developmental Therapeutics Program (DTP) Molecular Target Data Page.

We're interested in a set of microarray data produced by Novartis using Affymetrix U95 microarrays (near the bottom of the page above). Novartis profiled all of the cell lines in triplicate (mostly; a few were profiled 2 or 4 times). We want the data reported for the individual arrays, **"WEB_DATA_NOVARTIS_ALL.ZIP"**, which is about 27Mb (or 145Mb uncompressed).

2.3.2 Patient Samples

We'd like to relate the cell line data to patient data, to see if drug sensitivity in cell lines can be used to predict drug sensitivity in patients. We're going to use microarray data from a study of 24 breast cancer patients who were treated with single-agent docetaxel; these data were initially described by Chang et al, Lancet 2003. Patients were dichotomized into those who had responded to therapy ("Sensitive") and those who had not ("Resistant"). The microarray data are available as two gene set experiments (GSEs) from the public Gene Expression Omnibus (GEO) repository.

- GSE349 collects profiles from resistant patients
- GSE350 collects profiles from sensitive patients

We want to work with the "SOFT formatted family files" from each,

- "GSE349_family_soft.gz"
- "GSE350_family_soft.gz"

Both of these are about 4.5Mb.

2.3.3 Example Data

In late 2006, Potti et al claimed to have found a way to use drug sensitivity information and genomic profiles of cell lines to infer likely patient response to treatment from a patient's genomic profile. For a given drug, they used the drug sensitivity information to sort the cell lines according to sensitivity, and then they contrasted the microarray profiles of the most sensitive and most resistant lines.

We asked the authors if they could be more specific about precisely which genomic profiles were being used, and which cell lines were being treated as sensitive and resistant for each of the 7 drugs they examined.

The first file we got back in response was "chemo.zip", available here:

https://figshare.com/s/66603862d770b4c73146

Using this information, we attempted to infer the identities of the cell lines involved.

3 Course Outline

- A motivational case study - why I'm a zealot
- Literate programming, coupling code and data, and Sweave
- Markdown and R Markdown
- Starting a new analysis project / anchoring
- Directory structure
- README
- Gathering and describing raw data
- Processing data, R/Rmd mappability
- Analysis and reporting
- Report structure
- Automating steps and R packages
- devtools and roxygen2
- DESCRIPTION and LICENSE
- Code and coding conventions
- Templating
- What's hard? The *habit*!

4 Readings and References

There are quite a few discussions of aspects of RR, ranging from single pages to monographs.

4.1 Short List

Baggerly and Coombes, "Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology", Ann App Statist, 2009. This is the source for most of my motivating examples at this point, and shows how bad things can get.

Wilson et al, "Good Enough Practices in Scientific Computing", PLoS Comp Biol 2017. This is an article-length discussion of analysis workflows covering many of the issues I'll discuss in the course.

Bryan, "Excuse Me, Do You Have a Moment to Talk about Version Control?", PeerJ, 2017. Jenny Bryan has done a lot of thinking about the nuts and bolts of how to actually get things done. In addition to coverage of git and github, this article also discusses file management and which files should be shared.

Gandrud, Reproducible Research with R and RStudio, 2e. This monograph covers many of the topics we touch on as well as a few that we won't (e.g., the use of Make and Makefiles).

Xie, Dynamic Documents with R and knitr, 2e. This book covers knitr, rmarkdown, code chunks, and the use of html as an output format in far more detail than we'll be able to.

Wickham, R Packages. This is the reference for everything I'll present on the use of R Packages.

Cheat Sheets from RStudio. RStudio has a whole bunch of well done 1-or-2 page cheatsheets hitting the highlights of key packages and/or resources. Ones most relevant to this course include those for R Markdown, the RStudio IDE, and Package Development.