EEB 603 – Reproducible Science: Syllabus

Sven Buerki

Fall 2018

Contents

Instructor

- Name: Sven Buerki
- Office: Science building, office 114 (ground floor).
- Email: svenbuerki@boisestate.edu
- Office hours: By appointment.

Class meets

Tuesday & Thursday from 9:00 AM to 10:30 AM in EDUC521.

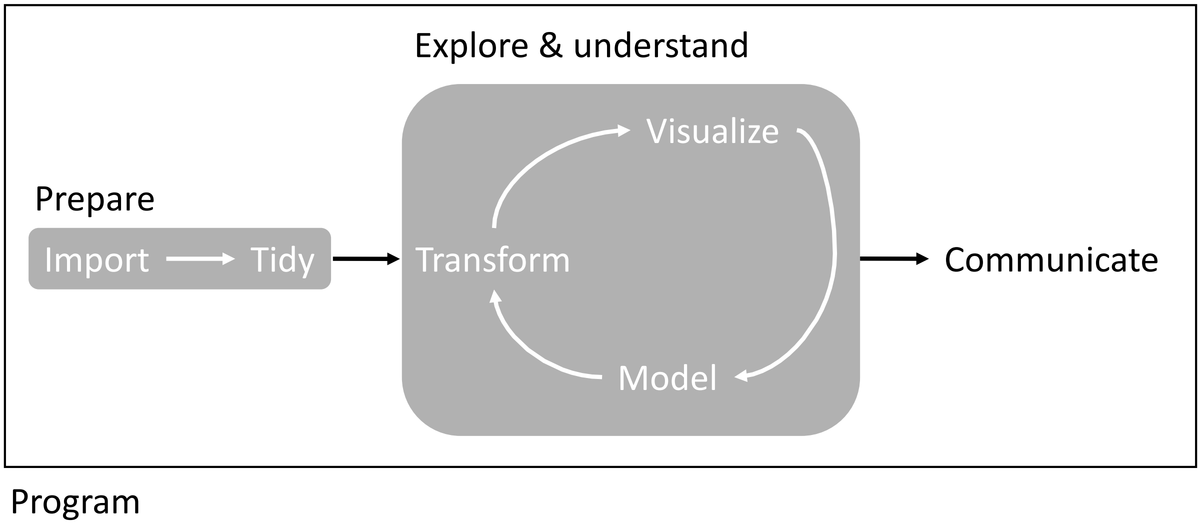

Figure 1: Summary of the bioinformatic workflow that will be taught in PART 2.

Aim

The scientific community widely acknowledges that we are in the midst of a reproducibility crisis (see e.g. Baker, 2016). This class starts by reviewing the evidence and causes supporting this crisis and aims at highlighting factors to boost reproducibility in science, especially in Ecology & Evolution. The overarching aim of this class is therefore to provide students with the theoretical knowledge and bioinformatic tools necessary to improve transparency, reproducibility and efficiency in scientific research. Across the course of the class, students will be taught to use open source software for their research such as R, RStudio and R Markdown (incl. knitr). Students will design bioinformatic tutorials and teach those to their peers. In addition, students will be tasked to develop individual projects aiming at developing a reproducible workflow specific to either their own thesis project or a scientific publication.

Ultimately, this class aims at **providing students with key knowledge to gather, store, share, prepare and analyse their data as well as communicate their results to the scientific community (Figure 1).**

Structure of the class

The class is subdivided into three parts:

- *PART 1: The Big Picture*
- *PART 2: Bioinformatics for Reproducible Science*
- *PART 3: Apply a Reproducible Approach to your Data*

Part 1 aims at providing students with key theoretical knowledge on this topic allowing them to successfully design a reproducible approach tailored to Ecology & Evolution. Part 2 provides students with opportunities to learn and apply the bioinformatic tools required to implement a reproducible workflow (see Figure 1). Here, each student (with the support of the instructor) will develop and teach a tutorial on a specific bioinformatic subject (the material will be taught across two sessions to provide enough time for students to assimilate each subject). Tutorials will be written in an open access format and distributed to the class. Finally, part 3 will consist in providing students with an opportunity to develop individual reproducible workflows tailored to their thesis project or to data presented in a publication. This will be done by applying knowledge gained during the previous parts and working in collaboration with the instructor (in some cases, we might seek support from thesis advisers).

Content of the class

PART 1: The Big Picture

1. **Introduction: The reproducibility crisis & prospects to tackle it!**
   - Chapter 1: Introduction to R, RStudio & Markdown (incl. referencing).
   - Chapter 2: The reproducibility crisis.
   - Chapter 3: Reproducible science in the world of Ecology & Evolution: Key role of voucher specimens.
2. **Getting started: Overview of data workflow & used software**
   - Chapter 4: Data management & Reproducible code.
   - Chapter 5: Getting published & Peer review.
   - Chapter 6: Storing, collaborating, accessing files and versioning (GitHub & Dropbox).

PART 2: Bioinformatics for Reproducible Science

3. **Prepare your data**
   - Chapter 7: Organize and import data with R.
   - Chapter 8: Prepare/tidy data for analyses in R.
4. **Explore & understand your data**
   - Chapter 9: Statistical modelling and *knitr*.
   - Chapter 10: Visualize results with Tables.
   - Chapter 11: Visualize results with Figures (incl. phylogenies).
5. **Communicate your results**
   - Chapter 12: Presenting on the Web (HTML) and other formats (PDF, WORD) with *knitr*/R Markdown.

PART 3: Apply a Reproducible Approach to your Data

6. **Students develop and present reproducible workflows applied to a project**
   - Produce individual reports showcasing reproducible workflows tailored to thesis projects.
   - Oral presentations of individual students' projects.

Google Site

A Google Site was created to deposit material taught in class.

The website is restricted to students enrolled in this class and it can be accessed at this address: https://sites.google.com/boisestate.edu/eeb603/home

Publications & Textbooks

The reading material at the basis of this class is composed of a mixture of publications and chapters from two textbooks (Gandrud, 2015; Wickham and Grolemund, 2017). We will also study the *"Guides to"* published by the British Ecological Society. Please find below the references used in each chapter.

Chapter	Reference(s)
Chap. 1	Chapter 3 of Gandrud (2015)
Chap. 2	Baker (2016); Freedman et al. (2015); Munafo et al. (2017); Sarewitz (2016)

Chapter	Reference(s)
Chap. 3	Bone et al. (2015); Smith et al. (2016)
Chap. 4	British Ecological Society (2014a) & Chapter 4 of Gandrud (2015); British Ecological Society (2014d) & Chapter 2 of Gandrud (2015)
Chap. 5	British Ecological Society (2014b); British Ecological Society (2014c)
Chap. 6	Chapter 5 of Gandrud (2015)
Chap. 7	Chapter 6 of Gandrud (2015)
Chap. 8	Chapter 7 of Gandrud (2015) & Chapters 9-10 of Wickham and Grolemund (2017)
Chap. 9	Chapter 8 of Gandrud (2015)
Chap. 10	Chapter 9 of Gandrud (2015)
Chap. 11	Chapter 10 of Gandrud (2015), Chapters 1 and 22 of Wickham and Grolemund (2017) & Guangchuang et al. (2017)
Chap. 12	Chapters 11 to 13 of Gandrud (2015) & Chapters 21, 23, 24 of Wickham and Grolemund (2017)

The computing tools of Reproducible Science

Research is often presented in the form of slideshows, articles or books. These presentation documents announce a project's findings, but they are not the research, they are the advertisement part of the research project!

> *The research is the full software environment, code, and data that produced the results (Donoho, 2010).*

When we separate the research from its advertisement, we are making it difficult for others to verify the findings by reproducing them.

This class will give you the tools to dynamically combine your research with the presentation of your findings. The first tool will be a workflow for reproducible research weaving the principles of reproducibility throughout your entire research project, from data gathering to the statistical analysis, and the presentation of results. To reach this goal, you will learn how to use a number of computer tools that make this workflow possible.

The bioinformatic tools

The main bioinformatic tools covered in this class are:

- The **R** statistical language that will allow you to gather data and analyze it.
- The **LaTeX** and **Markdown** markup languages that you can use to create documents (slideshows, articles, books, webpages) for presenting your findings.
- The *knitr* and *rmarkdown* **packages** for R and other tools, including **command-line shell programs** like GNU Make and Git version control, for dynamically tyding your data gathering, analysis, and presentation documents together so that they can be easily reproduced.
- **RStudio**, a program that brings all of these tools together in one place.

Installing the main software

As shown above, **R** and **RStudio** are at the core of this class and will have to be installed on your computers. This can be easily done by downloading the software from the following websites:

- **R**: https://www.r-project.org
- **RStudio**: https://www.rstudio.com/products/rstudio/download/

The download webpages for these software have comprehensive information on how to install them, so please refer to those pages for more information.

Installing markup languages

If you are planning to create LaTeX documents, you will need to install a Tex distribution. Please refer to this website for more details: https://www.latex-project.org/get/

If you want to create Markdown documents you can separately install the *rmarkdown* package in R (see below for more details).

Installing R packages

We will be using a number of R packages especially designed to support reproducible research. Many of those packages are not included in the default R installation and will need to be installed separately. To install key packages used in class, copy the following code and paste it into your R console:

```r
install.packages(c("brew", "countrycode", "devtools", "dplyr", "ggplot2", "googleVis",
    "knitr", "rmarkdown", "tidyr", "xtable"))
```

Once you enter this code, you may be asked to select a CRAN "mirror" to download the packages from. Simply select the mirror closest to you.

Finally, it is highly likely that we will have to install additional packages. In this case, you can simply install it by using the same R function `install.packages()` or by using RStudio as follows: Select "Tools" -> "Install Packages ... " and then type the name of the package in the window (make sure to tick the "Install dependencies" box).

RStudio Cheat Sheets: A gold mine to design your bioinformatic tutorials

RStudio provides a suite of cheat sheets that can be accessed by going to the "Help" menu and selecting "Cheatsheets".

Five cheat sheets are especially relevant to chapters taught in this class:

- *RStudio IDE : Cheat sheet* (Chapter 6)
- *Data Manipulation with dplyr, tidyr* (Chapter 9)
- *Data Visualization with ggplot2* (Chapter 12)
- *R Markdown Cheat Sheet* (Chapters 6 & 13)
- *R Markdown Reference Guide* (Chapters 6 & 13)

These documents together with the material presented in publications & textbooks will provide the basis to design your bioinformatic tutorials.

Grading

There will not be any classical exams in this class, but we will rather focus on developing theoretical and bioinformatic skills and applying those to your research. In this context, each student will be asked to produce a bioinformatic tutorial and teach it to their peers (see PART 2). Each student will also be tasked to produce a report (tailored to their thesis project or a publication) and present their results and conclusions to the class.

Tests conducted during the class

Students will be graded based on the following four tasks:

- Produce a bioinformatic tutorial focusing on a chapter from PART 2. Depending on enrollment students may be working in pairs (150 points).
- Teach a bioinformatic lab (spread across 2 sessions; 100 points).
- Produce an individual report on thesis project/publication (200 points).
- One oral presentation on thesis project/publication (100 points).

Exams are summing to a total of **550 points** and Table 2 exhibits the grading scale applied in this class.

Table 2: Grading scale applied in this class.

Percentage	Grade
100-98	A+
97.9-93	A
92.9-90	A-
89.9-88	B+
87.9-83	B
82.9-80	B-
79.9-78	C+
77.9-73	C
72.9-70	C-
69.9-68	D+
67.9-60	D
59.9-0	F

Bioinformatic tutorial (150 points)

During week 1, students will be assigned a chapter of PART 2 to study and produce a bioinformatic tutorial. Based on enrollment, students might work individually or in pairs.

Tutorials will have to be written in the *knitr/rmarkdown* language as implemented in RStudio. Tutorials should be focused on developing a suite of exercises aiming at gaining key bioinformatic skills specific to each chapter (see PART 2). Students will be welcomed to use material presented in Gandrud (2015) and Wickham and Grolemund (2017) to develop their tutorials, but they can also use other sources as long as they are properly cited in their documents. See Publications & Textbooks and RStudio Cheat Sheets sections for more details.

Students should design their tutorials to be completed within 2 laboratory sessions (see below). **Tutorials should be submitted to the instructor 1 week in advance for correction and to be uploaded onto the Google site.**

Tutorial should include:

- A short introduction highlighting the theory and aims of the tutorial.
- A section on R package requirements with instructions on how to install those packages and their dependencies.
- A section introducing the data (dataset) used to support these exercises (and how to download those).
- A references section and links to manuals of R packages.
- Commented R code necessary to guide users through the exercises as well as some knowledge on expected outputs.

Teaching bioinformatic tutorial (100 points)

Students are expected to prepare a 10 minutes presentation providing general guidelines to complete the tutorial. Presentations will be uploaded onto the Google site and made accessible to all students. Students are expected to support their peers in completing tutorials by answering questions. The instructor will also be circulating in class and answering questions, but students are leading the teaching of the bioinformatic laboratories.

Students will be graded according to their abilities to teach their tutorials and answer questions. The instructor might also use student's feedback to grade this test.

Individual report on thesis project/publication (200 points)

Students will work alongside the instructor to develop a reproducible workflow specific to their thesis project. In cases where students do not yet have a clear idea on their thesis project, they will work with instructor to identify a publication that can serve as basis for their individual project.

Reports will be written using the *knitr/rmarkdown* markup language as implemented in RStudio. The instructor expects students to provide a list of references supporting their reports. References will have to be cited in the text: it is not enough to through a bunch of references at the end of the report. This exercise aims at supporting methodological decisions taken in the report and increasing transparency.

Students investigate the following prior key elements

- What kind of data are already published/available (and which of those are relevant to your topic)?
- Where are those published data deposited?
- Can you reproduce the analyses based on published data?
- What types of data are or will be produced during your thesis project?
- What are the specificity of your data (in term of storage, sharing, etc.)?
- What are the publication standards in your field (see e.g. Donoho, 2010; Smith et al., 2016)?
- etc…

Students develop the following core processes

- A data management workflow specific to your research, which will cover the following stages of the data life cycle (see e.g. British Ecological Society, 2014a):
  - Create
  - Process
  - Document
  - Preserve
  - Share
  - Reuse
- A reproducible code to perform the following tasks to your data (see Fig. 1):
  - Import
  - Tidy/clean
  - Transform
  - Visualize
  - Model
  - Communicate

Oral presentation on thesis project/publication (100 points)

Each student will have to present their report during final week. The presentation should follow the same structure as the report and not exceed 15 minutes. There will be 5 minutes at the end of the presentation allocated for questions.

Cheating

You are responsible for knowing and understanding the BSU student conduct code. For a complete description see: https://deanofstudents.boisestate.edu/student-code-of-conduct/

References

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533: 452–454. Available at: https://doi.org/10.1038/533452a.

Bone, R.E., J.A.C. Smith, N. Arrigo, and S. Buerki. 2015. A macro-ecological perspective on crassulacean acid metabolism (cam) photosynthesis evolution in afro-madagascan drylands: Eulophiinae orchids as a case study. *New Phytologist* 208: 469–481. Available at: http://dx.doi.org/10.1111/nph.13572.

British Ecological Society ed.. 2014a. A guide to data management in ecology and evolution. British Ecological Society.

British Ecological Society ed.. 2014b. A guide to getting published in ecology and evolution. British Ecological Society.

British Ecological Society ed.. 2014c. A guide to peer review in ecology and evolution. British Ecological Society.

British Ecological Society ed.. 2014d. A guide to reproducible code in ecology and evolution. British Ecological Society.

Donoho, D.L. 2010. An invitation to reproducible computational research. *Biostatistics* 11: 385–388. Available at: http://dx.doi.org/10.1093/biostatistics/kxq028.

Freedman, L.P., I.M. Cockburn, and T.S. Simcoe. 2015. The economics of reproducibility in preclinical research. *PLOS Biology* 13: e1002165. Available at: https://doi.org/10.1371/journal.pbio.1002165.

Gandrud, C. 2015. Reproducible Research with R and RStudio. C. Gandrud [ed.], CRC Press.

Guangchuang, Y., S.D. K., Z. Huachen, G. Yi, and L.T.T. Yuk. 2017. Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8: 28–36. Available at: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12628.

Munafo, M.R., B.A. Nosek, D.V.M. Bishop, K.S. Button, C.D. Chambers, N.P. du Sert, U. Simonsohn, et al. 2017. A manifesto for reproducible science. *Nature human behaviour* 1: 0021.

Sarewitz, D. 2016. The pressure to publish pushes down quality. *Nature* 533: 147–147. Available at: https://doi.org/10.1038/533147a.

Smith, J.F., T.H. Parker, S. Nakagawa, J. Gurevitch, Ecology, and T.(. for T. in Evolution) Working Group. 2016. Promoting transparency in evolutionary biology and ecology. *Systematic Botany* 41: 495–497. Available at: http://www.bioone.org/doi/abs/10.1600/036364416X692262.

Wickham, H., and G. Grolemund. 2017. R for data science: Import, tidy, transform, visualize, and model data. 1st ed. O'Reilly Media, Inc. Available at: http://r4ds.had.co.nz.