labibi Documentation

Release 1.0

C. Titus Brown

March 16, 2016

Contents

CHAPTER 1

Repeatability vs reproducibility - a discussion

Learning goals: attendees will understand some of the basic arguments for computational repeatability in science.

While the name of the workshop is "reproducibility", what we're really going to be talking about here is what I will call "repeatability" - the ability to repeat, exactly, a computational analysis.

(There is a lot of epistemic confusion around the precise meanings of the words "reproducibility" and "replication", and even "repeatability". Sorry.)

In many computational analyses, repeatability should be pretty easy to achieve - given same input data,

Why is repeatability valuable? There are two reasons:

1. Efficiency, reuse, and remixing - you and others can repeat, edit, reuse, and repurpose work.
2. Reproducibility - if someone else cannot reach approximately the same results with similar input data, then *repeatability* is a requirement for tracking down the differences.

1.1 Two stories

PROBLEM: Amanda did variant calling on some genomic data, and wrote up the results for publication. One of the reviewers thinks her broader results are biased by the choice of aligner and asks her to redo the analysis with a different program.

SOLUTION: Amanda reruns her analysis to repeat her original results, then forks her workflow, edits three lines to replace the Bowtie2 aligner with the BWA-mem aligner, and reruns the analysis.

OUTCOME 1: Amanda discovers that indeed her analysis results are quite different with BWA.

OUTCOME 2: Amanda discovers that her analysis results are very similar whether she uses BWA or Bowtie.

PROBLEM: Amanda did variant calling on some genomic data, and published the results. After publication, another research group led by Julio (with access to different samples) says that they find a different global pattern of genotypic variants underlying the same phenotype. Because both research groups used different samples, data collection approaches, and data analysis approaches, they wonder where the true disagreement lies.

SOLUTION: Julio takes Amanda's pipeline and runs it on their data.

OUTCOME 1: They discover that Julio's data run with Amanda's pipeline gives Amanda's results, and now they can start tracking down what Julio did differently from Amanda in data analysis.

OUTCOME 2: They discover that Julio's data run with Amanda's pipeline gives Julio's results. This suggests that there is something different about the samples or data collection, while the data analysis itself is not the source of the differences.

1.2 Repeatability vs reproducibility: what's the target?

For science, *reproducibility* is the goal - if other scientists can't follow your process and reach roughly the same results, then the results aren't robust. (For more, read this excellent Wikipedia article, and you can also read Dorothy Bishop's excellent discussion of the reproducibility crisis in psychology.. Consider the second story above - if neither Amanda nor Julio's data analysis workflow had been easy to repeat, it would have been difficult to know whether data analysis was the source of the differences or not.

For computational folk, *repeatability* is useful for reproducibility, but also for other purposes - consider the first story, above, where Amanda can re-run her analysis quickly with a different aligner (or with different parameters). This is an increase in *efficiency*. This efficiency argument can be extended to many more scenarios –

- Tim, who collaborates with Amanda's lab, wants to do a similar analysis but with different data. He can use Amanda's pipeline as a starting point.
- Kim, Amanda's advisor, needs to repeat Amanda's analysis after Amanda leaves the lab. They can re-run Amanda's pipeline even without Amanda around.
- Amanda, in 3 years, needs to run a similar analysis on new data (or has a student that needs to run a similar analysis). Rather than trying to remember all the details of her own analysis from 3 years ago, she can start from a working analysis.
- Qin wants to do a meta-analysis of Amanda and Julio's data. They can start from Amanda's workflow and use it to run analyses on Amanda's data, Julio's data, etc. with some consistency.

The first three of these scenarios increase efficiency and time to publication for Amanda and her group; the fourth is a more general improvement for the field overall.

The question for you all to consider is this: how much time and effort should you put into making your workflows reproducible, given the expected benefits?

The *goal* of this workshop overall is to show you the *kind* of overall workflow and toolset that we, and many others, have converged upon.

CHAPTER 2

Indices and tables

- genindex
- modindex
- search