Reproducible Research

Marco Chiapello

June 10, 2016

Center for Proteomics
University of Cambridge
mc983@cam.ac.uk

Overview

Introduction

Reproducible research Reasons

Reproducible research Rules

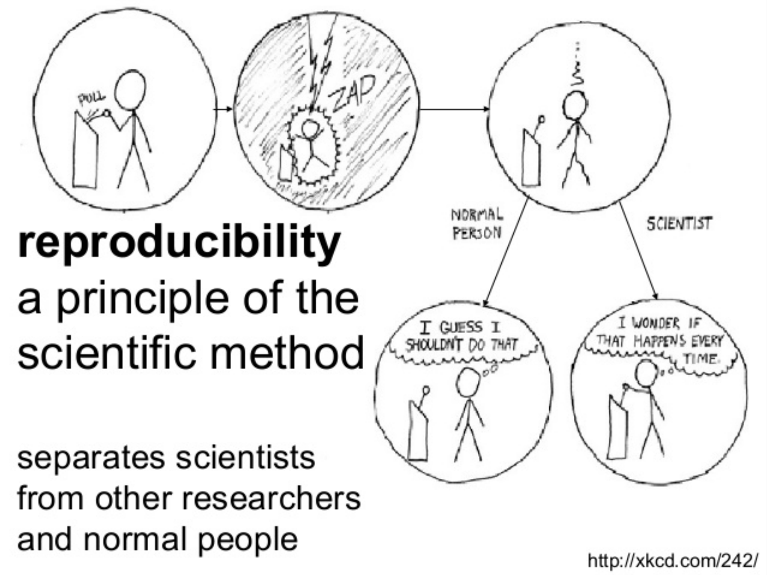
Reproducible research Tools

Conclusion

Introduction

Replication is the ultimate standard by which scientific claims are judged [2]
The fact that an analysis is reproducible does not guarantee the quality, correctness,
or validity of the published results.

What reproducible research is



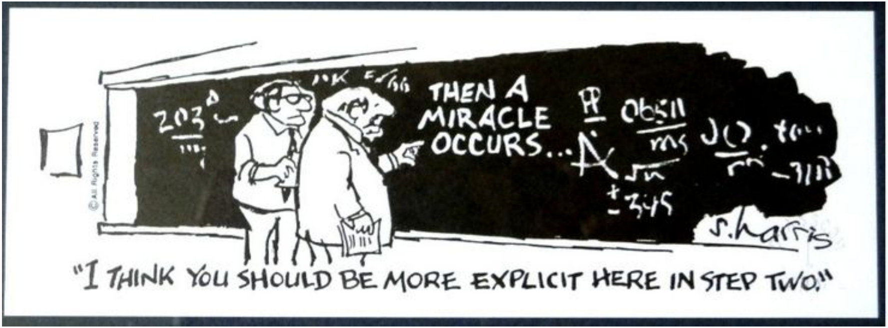
reproducibility
a principle of the scientific method

separates scientists from other researchers and normal people

PULL
ZAP
NORMAL PERSON
SCIENTIST
I GUESS I SHOULDN'T DO THAT
I WONDER IF THAT HAPPENS EVERY TIME.

http://xkcd.com/242/

What reproducible research is



- This is exactly how it seems when you try to figure out how authors got from a large and complex data set to a dense paper with lots of busy figures.

  Without access to the data and the analysis code, a miracle occurred.
- And there should be no miracles in science. [1]

What reproducible research is

$$\mathit{DATA} + \mathit{ANALYSIS} \rightarrow \mathit{RESULTS}$$

Common practice of writing statistical reports:

- We import a dataset into Excel
- Run a procedure to get all results
- Copy and paste selected pieces into a typesetting program
- Add a few descriptions
- Finish a report

What reproducible research is

There are obvious dangers and disadvantages in this process:

1. It is error-prone due to too much manual work;
2. It requires lots of human effort to do tedious jobs;
3. The workflow is barely recordable, therefore it is difficult to reproduce;
4. A tiny change of the data source in the future will require the author(s) to go through the same procedure again;
5. The analysis and writing are separate, so close attention has to be paid to the synchronization of the two parts.

What reproducible research is

What is Reproducible Research?

THE ABILITY TO REPRODUCE SOMEONE ELSE RESULTS

What do you need?

- Analytic data
- Analytic code
- **Documentation for data and code**

What reproducible research is

Reproducible vs Replicable

		DATA	
		Same	Different
CODE	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Ref: `https://github.com/KirstieJane/ReproducibleResearch`

What reproducible research is

Reproducibility/reproduce

A study is reproducible if there is a specific set of computational functions/analyses (usually specified in terms of code) that exactly reproduce all of the numbers in a published paper from raw data.

Replication/replicate

A study is only replicable if you perform the exact same experiment (at least) twice, collect data in the same way both times, perform the same data analysis, and arrive at the same conclusions.

Replicability requires new samples and new data[1], which introduces new variability, and additional risks of errors. **Reproducibility** is, to some extent, a technical challenge, while replication gives the results scientific validity.

Ref: `https://github.com/lgatto/TeachingMaterial/tree/master/open-rr-bioinfo-best-practice`

[1] in particular biological replicates

Reproducible research Reasons

How does working reproducibly help to achieve more as a scientist [1]

Reasons to work reproducibly

REPRODUCIBILITY [1]

Idealist:

1. It is the foundation of science!
2. The world would be a better place if everyone worked transparently and reproducibly!

Realist:

1. It helps to avoid disaster
    - You need to record in detail how you got there
    - Work reproducibly early on will save you time later
2. It makes it easier to write papers
    - To have very transparent data and code, it costs just few minutes to spot a mistake (if any)
3. It helps reviewers see it your way
    - Made the data and well-documented code easily accessible to the reviewers
4. It enables continuity of your work
    - How can you ensure the continuity of work in your lab if progress is not documented reproducibly?
    - No proof of reproducibility, no result!
5. It helps to build your reputation
    - To build a reputation for being an honest and careful researcher

Reproducible research Rules

– based on Sandve et al., 2013 [3]

Rule 1

For Every Result, Keep Track of How It Was Produced

- The full sequence of pre- and post-processing steps are often critical in order to reach the achieved result
- Every detail that may influence the execution of the step should be recorded
- Include the name and version of the program, as well as the exact parameters and inputs

As a minimum, you should at least record sufficient details on programs, parameters, and manual procedures to allow yourself, in a year or so, to approximately reproduce the results

Rule 2

Avoid Manual Data Manipulation Steps

- Manual procedures are not only inefficient and error-prone, they are also difficult to reproduce
- Manual modification of files can usually be replaced by the use of standard UNIX commands or scripts
- Manual tweaking of data files to attain format compatibility should be replaced by for- mat converters that can be reenacted and included into executable workflows
- Manual operations like the use of copy and paste between documents should also be avoided

If manual operations cannot be avoided, you should as a minimum note down which data files were modified or moved, and for what purpose

Rule 3

Rule 4

Rule 5

Rule 6

Rule 7

Rule 8

Rule 9

Rule 10

Reproducible research Tools

> *Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to humans what we want the computer to do.*
>
> *– Donald E. Knuth Literate Programming, 1984*

Tools

Literate programming is a methodology that combines a programming language with a documentation language

- Write program code to do computing
- Write narratives to explain what is being done by the program code

Tools

At the lowest level, working reproducibly just means avoiding beginners? mistakes. Keep your project organized, name your files and directories in some informative way, store your data and code at a single backed-up location. Don?t spread your data over different servers, laptops and hard drives. To achieve the next levels of reproducibility, you need to learn some tools of computational reproducibility [8]. In general, reproducibility is improved when there is less clicking and pasting and more scripting and coding. For example, do your analysis in R (https://www.r-project.org/) or Python (https://www.python.org/) and document your analysis using knitR (http://yihui.name/knitr/) or IPython notebooks (http://ipython.org/). These tools help you to merge descriptive text with analysis code into dynamic documents that can be automatically updated every time the data or code change. As a next step, learn

Tools

Learning the tools of the trade will require commitment and a massive investment of your time and energy. A priori it is not clear why the benefits of working reproducibly outweigh its costs.

Does reproducibility sound like extra work? It can be, particularly when one is first trying to do it, that is, to break one's own previous nonreproducible habits

Conclusion

Conclusions

My advice is: learn the tools of reproducibility (Box 1) as quickly as possible and use them in every project.

References

[1] Markowetz, F. (2016). Five selfish reasons to work reproducibly. *Genome biology*, pages 1–4.

[2] Peng, R. D. (2011). Reproducible research in computational science. *Science (New York, NY)*, 334(6060):1226–1227.

[3] Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology*, 9(10):e1003285–4.

Acknowledgements

- Mike Deery Manager of the Cambridge Center for Proteomics
- Laurent Gatto Computational Proteomics Unit

- Andrea Genre Universita' di Torino