Open Science, Open Plant Pathology

Adam H. Sparks

Emerson Del Ponte

Sydney Everhart

Zachary Foster

Niklaus J Grünwald

15 February, 2019

Open research practices have been highlighted extensively during the last ten years in many fields of study as essential standards needed to promote transparency and reproducibility of scientific results. In fact, scientific claims can only be evaluated based on how protocols, materials, equipment and methods were described; data were collected and prepared; and analyses conducted. Openly sharing protocols, data and computational code are central for current scholarly dissemination and communication, but in many fields, including plant pathology, adoption of these practices has been slow. We randomly selected 200 articles published from 2012 to 2016 across 21 journals representative of the pathology discipline, and assigned them scores reflecting their openness and reproducibility. We found that most of the articles were not very open, failing to share data or code in a meaningful way. We propose that using open source tools for producing open and reproducible work and analysis is advantageous, benefiting not just readers, but the authors as well and provide ideas and tools to promote open, reproducible research practices among plant pathologists.

# Introduction

Modern plant pathological research has many facets given the array of disciplines and sub-disciplines currently involved. Collectively, they contribute to increase our basic and applied knowledge on several aspects of pathogen biology and disease development to ultimately improve management. Scientific research in the field varies from the purely observational or descriptive nature to inferential based on experimental or simulation-derived data sets. Whatever the case, research findings are verifiable based on how much of the research materials, processes and outcomes are made available beyond what is reported in the scientific article. These include biological materials (strains), nucleic/protein sequences, experimental and simulated raw data annotations, drawings and photographs and statistical analysis code among other data.

## Reproducible Research

Reproducibility in scientific research has once again been highlighted recently (Nature 2016; Baker 2016) as an important issue. Patil et al. (2016) have provided several definitions to clarify the concepts surrounding reproducibility. For the purposes of this paper we follow the definitions as given by Patil et al. (2016).

## Open Science

Open science has become highlighted lately with many donors expecting data to be available (XXXX) and other scientists interested in sharing and collaborating more widely (XXXX).

# A General Workflow

A general workflow for producing academic research involves clearly defining a research question, obtaining data for testing the hypothesis, summarizing/analyzing and presenting data and results, and writing the manuscript. Here we defined three levels of reproducibility which are also related with the evolution of computational methods and reproducible practices 1.

A first level of reproducibility involves openly available research materials such as strains and/or nucleic acid sequences in public collections and citations for methods used. A second level involves providing raw data and code as binary files (PDF or other non-text file) in supplemental materials, which do not allow prompt access to the data and running the code because of use of expensive commercial software or a pay-wall. The highest level includes efforts to annotate structured raw data (FAIR – Findable, Accessible, Interoperable and Re-usable) and fully document the analysis using open source code which are deposited in public repositories and can be run by anyone following download of data and code. The first level, as reported, is an essential step that is not substituted by the other practices and eventually researchers fail to provide sufficient description or correct citations. In the next section we present standards and tools that can be used to ensure reproducibility.

### Methods

When making your science more open and reproducible, methods, software used (including things like R packages), etc. should be cited properly. This allows end users to identify what was used and identify the methods used more accurately. Just as importantly, this acknowledges the contributions of others whose works were instrumental in your research. This also helps ensure that you as a researcher are able to reconstruct what you have done since you will have good notes and documentation and be able to identify if something changes, *e.g.* a package version, what effect it had on your research.

* deposit and annotate biological materials
* provide full description for equipments, etc.
* rOpenSci Reproducible Research: <http://ropensci.github.io/reproducibility-guide/sections/introduction/>

### Data

* Data formatting (flat files; use Comma Chameleon, Table Tool, others?)
* Data annotation
* Data storage (don’t edit raw data files; use file permissions to prevent changes to raw data files, use data bases where possible and appropriate; etc.)

### Source Code

* The problem of commercial software and mouse-based routines
* Why to avoid binary files as supplements?
* Writing and documenting using open source software
* Availability in public repositories

### Repository

* Using GitHub, Gitlab, Bitbucket for code (and small data?)
* Using Figshare, Dataverse or some other data repository for large data
* Zenodo or OSF to generate DOIs
* Why not use a lab website (DOIs, other reasons)
* Most ideally, deposit in university library or other proper long-term data archival service

# Status in Plant Pathology

* Madden et al. (2015) supply an *e-****X****tra*\* with reproducible examples for readers.
* Duku et al. (2016) provide models, data and code, (<http://adamhsparks.github.io/MICCORDEA/>) necessary to replicate the entire study modelling the effects of climate change on rice bacterial blight and rice leaf blast in Tanzania.
* Sparks et al. (2011, 2014) provide models, data and code, (<http://adamhsparks.github.io/Global-Late-Blight-MetaModelling/>) necessary to replicate model development and the subsequent the study on the effects of climate change on potato late blight.
* Del Ponte provides data and a reproducible report that explain in details all steps of the analysis and the R codes for conducting a meta-analysis for assessing heterogeneity in relationship between white mold incidence and soybean yield and between incidence and soybean tied.
* Example from Grünwald lab:
  + paper <http://apsjournals.apsnet.org/doi/full/10.1094/PHYTO-12-14-0350-FI>
  + github repo <https://github.com/grunwaldlab/Sudden_Oak_Death_in_Oregon_Forests>
* Other examples from plant pathology providing e-Xtras or supplemental material

Twenty-one plant pathology discipline journals were selected by the authors as representations of discipline-based journals target by the plant pathology research community. Among them, both fundamental and/or applied as well as journals covering specific group of pathogens/plants or broad areas were included. Two hundred articles were randomly selected from issues published from 2012 to 2016. A list of randomly selected pages was assigned to a randomized list of the 21 journals (Sparks et al. 2017) where the page number fell within an article for the given journal. In cases where an article was not suitable, *e.g.*, a review or otherwise not related to plant pathology, the next article was selected until a suitable article was found. Notes regarding the selection of articles can be found in the file, XXXX, available in this paper’s repository. The pages list was numbered from page one and went to 150. This was done since some journals restart their numbering with each issue and also ensures that the journal is more likely to have a page number corresponding to the randomly generated value. This also assumes that there is no effect or bias on reproducibility based on the time of year that an article was published, since most journals start with page number one at the beginning of the year. The list of journals was saved as a comma separated value (CSV) file and imported into R (R Core Team 2018).

# Discussion

# Acknowledgements

##### pagebreak

# References

Baker, M. 2016. Muddled meanings hamper efforts to fix reproduciblity crisis. Nature News.

Duku, C., Sparks, A. H., and Zwart, S. J. 2016. Spatial modelling of rice yield losses in tanzania due to bacterial leaf blight and leaf blast in a changing climate. Climatic Change. 135:569–583.

Madden, L. V., Shah, D. A., and Esker, P. D. 2015. Does the *P* value have a future in plant pathology? Phytopathology. 105:1400–1407.

Nature. 2016. Reality check on reproduciblity. Nature. 533:437.

Patil, P., Peng, R. D., and Leek, J. 2016. A statistical definition for reproducibility and replicability. bioRxiv. Available at: <http://biorxiv.org/content/early/2016/07/29/066803>.

R Core Team. 2018. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.

Sparks, A. H., Forbes, G. A., Hijmans, R. J., and Garrett, K. A. 2011. A metamodeling framework for extending the application domain of process-based ecological models. Ecosphere. 2:art90.

Sparks, A. H., Forbes, G. A., Hijmans, R. J., and Garrett, K. A. 2014. Climate change may have limited effect on global risk of potato late blight. Global Change Biology.:3621–3631.

Sparks, A. H., Ponte, E. M. D., Foster, Z., and Grünwald, N. J. 2017. Reproducible-research-in-plant-pathology. Available at: <https://github.com/adamhsparks/Reproducible-Research-in-Plant-Pathology>.

##### pagebreak

##### pagebreak

# Figures

![Figure 1 An open and reproducible research workflow.](data:application/pdf;base64,)

Figure 1 An open and reproducible research workflow.

##### pagebreak

# Colophon

This report was generated on 2019-02-15 15:32:02 using the following computational environment and dependencies:

#> R version 3.5.2 (2018-12-20)  
#> Platform: x86\_64-apple-darwin18.2.0 (64-bit)  
#> Running under: macOS Mojave 10.14.3  
#>   
#> Matrix products: default  
#> BLAS/LAPACK: /usr/local/Cellar/openblas/0.3.5/lib/libopenblas\_haswellp-r0.3.5.dylib  
#>   
#> locale:  
#> [1] en\_AU.UTF-8/en\_AU.UTF-8/en\_AU.UTF-8/C/en\_AU.UTF-8/en\_AU.UTF-8  
#>   
#> attached base packages:  
#> [1] stats graphics grDevices utils datasets methods base   
#>   
#> other attached packages:  
#> [1] rsvg\_1.3   
#> [2] DiagrammeRsvg\_0.1   
#> [3] DiagrammeR\_1.0.0   
#> [4] Reproducibility.in.Plant.Pathology\_0.0.1-1  
#>   
#> loaded via a namespace (and not attached):  
#> [1] tidyselect\_0.2.5 xfun\_0.4 Rook\_1.1-1   
#> [4] purrr\_0.3.0 V8\_2.0 colorspace\_1.4-0   
#> [7] htmltools\_0.3.6 viridisLite\_0.3.0 yaml\_2.2.0   
#> [10] XML\_3.98-1.17 rlang\_0.3.1 pillar\_1.3.1   
#> [13] glue\_1.3.0 RColorBrewer\_1.1-2 plyr\_1.8.4   
#> [16] stringr\_1.4.0 munsell\_0.5.0 gtable\_0.2.0   
#> [19] visNetwork\_2.0.5 htmlwidgets\_1.3 evaluate\_0.13   
#> [22] knitr\_1.21 curl\_3.3 highr\_0.7   
#> [25] Rcpp\_1.0.0 readr\_1.3.1 scales\_1.0.0   
#> [28] jsonlite\_1.6 rgexf\_0.15.3 gridExtra\_2.3   
#> [31] brew\_1.0-6 ggplot2\_3.1.0 hms\_0.4.2   
#> [34] digest\_0.6.18 stringi\_1.3.1 bookdown\_0.9   
#> [37] dplyr\_0.8.0 grid\_3.5.2 influenceR\_0.1.0   
#> [40] tools\_3.5.2 magrittr\_1.5 lazyeval\_0.2.1   
#> [43] tibble\_2.0.1 crayon\_1.3.4 tidyr\_0.8.2   
#> [46] pkgconfig\_2.0.2 downloader\_0.4 assertthat\_0.2.0   
#> [49] rmarkdown\_1.11 rstudioapi\_0.9.0 viridis\_0.5.1   
#> [52] R6\_2.4.0 igraph\_1.2.4 compiler\_3.5.2

The current Git commit details are:

#> Local: master /Users/U8004755/Development/Reproducibility\_in\_Plant\_Pathology  
#> Remote: master @ origin (https://github.com/adamhsparks/Reproducibility\_in\_Plant\_Pathology.git)  
#> Head: [751becc] 2019-01-23: Edits to paper text