Status and Best Practices for Reproducible Research In Plant Pathology

A. H. Sparks, E. M. Del Ponte, N. J. Grünwald, and Z. Foster

Last update: 2017-08-15

First author: Centre for Crop Health, University of Southern Queensland, Toowoomba, Qld 4350, Australia; second author: Departamento de Fitopatologia, Universidade Federal de Viçosa, Viçosa, MG, Brasil; third and fourth authors: Horticultural Crops Research Laboratory, USDA Agricultural Research Service, Corvallis, OR 97330, USA

Accepted for publication:

# Abstract

Reproducible research practices have been highlighted extensively during the last ten years in many fields of study as essential standards needed to promote transparency and replication 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. Sharing of 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 reproducibility. We found that.. Protocols specific to plant pathology and open source tools for producing reproducible work and analysis is proposed to promote reproducible practices among plant pathologists.

# Main text

## Introductory

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 vary from purely observational/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, drawing and photographs, statistical analysis codes, among other data.

## 3 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).

## 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 (Fig. 1).

A first level of reproducibility involves making available research materials such as strains and/or nucleic acid sequences in public collection and citations for methods used. A second level involves providing raw data and codes as binary files (PDF or other non-text file) in supplemental materials which do not allow prompt access to the data and running the codes because of use of expensive commercial software or a paywall. A highest level includes efforts to annotate structured raw data and fully document the analysis using open source code which are deposited in public repositories and can be run more easily following download of data and codes. 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 (Figure 1).

### Methods

* Citation of methods, software, packages, etc.
* deposit and annotate biological materials
* provide full description for equipments, etc.

### 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 for code (and small data?)
* Using Figshare or Zenodo vs a lab website (DOIs, other reasons)

## 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 2017).

## Discussion

## Acknowledgments

## Literature cited

Baker, M. 2016. Is there a reproducibility crisis? Nature. 533:453–454.

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 Available at: <http://dx.doi.org/10.1007/s10584-015-1580-2>.

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. 2017. *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 Available at: <http://www.esajournals.org/doi/abs/10.1890/ES11-00128.1>.

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 Available at: <http://dx.doi.org/10.1111/gcb.12587>.

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> [Accessed ].