Open Science, Open Plant Pathology

Adam H. Sparks1,✉, Emerson Del Ponte2, Kaique dos Santos Alves2, Zachary Foster3, and Niklaus J Grünwald4

24 July, 2020

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 300 articles published from 2012 to 2018 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.

1 University of Southern Queensland, Centre for Crop Health, Toowoomba, Queensland, Australia  
2 Departmento de Fitopatologia, Universidade Federal de Viçosa, Brazil  
3 Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA  
4 Horticultural Crops Research Unit, USDA Agricultural Research Service, Corvallis, OR, USA

✉ Correspondence: [Adam H. Sparks <[Adam.Sparks@usq.edu.au](mailto:Adam.Sparks@usq.edu.au)>](mailto:Adam.Sparks@usq.edu.au)

# 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 plant disease 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 and the ability of others to make use of your methods and results. These research findings include biological materials (isolates or strains), nucleic/protein sequences, experimental and simulated raw data annotations, drawings and photographs and statistical analysis code among other materials and data generated as a course of conducting research. That is, open science leads to reproducibility and replicability.

## Definitions

In order for us to easily discuss the topic, we first must define what we mean so that we may clearly communicate. Many of the terms used in this area have varying definitions that may or may not agree with each other. For instance, reproducible research is recently highlighted by many authors (Baker 2016; Iqbal et al. 2016; Editors at Nature 2016; Patil et al. 2016; Weissgerber et al. 2016; Brunsdon 2015; Sweedler 2015; Fitzjohn et al. 2014; Ioannidis 2014; Fidler and Gordon 2013; Stodden et al. 2013) as an important issue. The International Union of Pure and Applied Chemistry provides definitions of repeatability in the “Compendium of Chemical Terminology – Gold Book” as:

“The closeness of agreement between independent results obtained with the same method on identical test material, under the same conditions (same operator, same apparatus, same laboratory and after short intervals of time). The measure of repeatability is the standard deviation qualified with the term: `repeatability’ as repeatability standard deviation. In some contexts repeatability may be defined as the value below which the absolute difference between two single test results obtained under the above conditions, may be expected to lie with a specified probability”, (McNaught and Wilkinson 1997).

and reproducibility is defined as:

“The closeness of agreement between independent results obtained with the same method on identical test material but under different conditions (different operators, different apparatus, different laboratories and/or after different intervals of time). The measure of reproducibility is the standard deviation qualified with the term ‘reproducibility’ as reproducibility standard deviation. In some contexts reproducibility may be defined as the value below which the absolute difference between two single test results on identical material obtained under the above conditions, may be expected to lie with a specified probability. Note that a complete statement of reproducibility requires specification of the experimental conditions which differ”, (McNaught and Wilkinson 1997).

However, while these definitions are clear and well accepted in physical sciences, in biological sciences it is not always possible to use identical test material or perhaps the the time or resources are not available for full reproduction, *e.g.*, field trials that span years and locations or complex glasshouse experiments. Therefore we will follow Peng’s (2009) definition that provides clear guidelines a minimum standard of ``reproducible research’’:

“The replication of scientific findings using independent investigators, methods, data, equipment, and protocols has long been, and will continue to be, the standard by which scientific claims are evaluated. However, in many fields of study there are examples of scientific investigations that cannot be fully replicated because of a lack of time or resources. In such a situation, there is a need for a minimum standard that can fill the void between full replication and nothing. One candidate for this minimum standard is ``reproducible research’’, which requires that data sets and computer code be made available to others for verifying published results and conducting alternative analyses”. Peng, R. (2009). *Reproducible research and Biostatistics*. Biostatistics, 10 (3): 405-408.

### Open Science

Open science has become highlighted lately with many donors expecting data to be available (Government of Canada 2016; Noorden 2017; Australian Research Council 2018) and other scientists interested in sharing and collaborating more widely (Wald 2010).

# 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 (Figure 1).

A first level of reproducibility involves openly available research materials such as isolates or 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) (Wilkinson et al. 2016) 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, this includes items such as R or Python packages that were directly used in the analysis or production of the paper, etc. should be cited properly. Deposit and annotate biological materials with an herbarium or other repository. Provide a full description of all equipment used, *e.g.* a Spectrum Technologies Watchdog 2700 weather station was used to record wind speed, direction, rainfall, temperature and relative humidity at one hour intervals. 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.

The use of programming or scripting languages such as R, SAS or Python enable you as a scientist to keep very detailed records of what was computationally performed. This is as opposed to using software such as spreadsheet programs like Excel, Google Sheets, Numbers, Calc or others which can be used for simple statistical analyses and visualization (but should not be) or other point-and-click software packages that do not enable you to keep an accurate record of the steps taken to import, format, visualise and analyse data.

### Data

Prefer text files for saving small sets of data. Data that are saved in binary formats such as PDF files are difficult to reuse because they are not easily machine readable. In many cases, data sets are small enough and curated in spreadsheets, which should be saved as a plain text file, *e.g.* comma separated (CSV) files. This also helps ensure that the data are reusable.Larger data sets may warrant the use of a proper database like MariaDB or PostgreSQL which provide users with several benefits but two important ones to mention here are data redundancy, no duplicated records; and data consistency, all of your records are sure to be recorded in the same way for every observation. While databases may offer many advantages, the trade-off is that they are more complex to set up and administer, especially for a small data set that might be only a few rows and columns in a database.

Ideally, once the data are complete, best practices for keeping your data as you perform your work include treating the raw data as read-only and using file permissions to prevent changes to the raw data files. It should be noted, that the use of a database management system also allows for both of these at the expense of added complexity. Saving files in proprietary formats such as .xls(x) can also lead to issues in the future when opening using newer (or older) software versions. Unexpected changes to values in the data (Ziemann et al. 2016) may also occur when using proprietary formats.

If steps are followed to make the data FAIR then it will be readable by humans and machines alike this will help support new discoveries and support further research. In turn sharing data will lead to new citations for your work as other discover and use it.

To make your data the most widely usable, ensure that it has a persistent identifier. A digital object identifier (DOI) is the most common (<https://www.doi.org/>) but handle is also an option (<https://handle.net/>). There are different options for generating a DOI for your data and other materials. FigShare, Zenodo and OSF all offer long-term archival along with a service to generate a DOI for your materials. The use of a persistent identifier works to ensure that even if the data are moved, they can still be located using that unique identifier.

For more on FAIR data, visit Go-Fair <https://www.go-fair.org/fair-principles/>.

### Sharing Your Research

Once you have determined how to best manage your source code for analysis and the data sets the next step is to consider how to share your data. Providers such as FigShare, Dataverse, OSF.io and Zenodo allow for you to deposit your data and generate a DOI for sharing your project once you are finished with it.

Other providers exist that allow for you to not only track changes but also to share the data openly, these include GitHub, Gitlab and Bitbucket. GitHub is arguably the most popular and widely used software development platform currently. Data that is encoded in CSV or other plain text formats can easily be deposited in a repository along with code for analysis to enable changes to be tracked and other users to download and replicate the work.

We would advise against the practice of depositing data on a laboratory website or a site such as GitHub only though. Doing either of these leaves the work in an unstable state where future users may be unable to access the work. It is a best practice to always ensure that you have deposited the data with a provider such as Zenodo, FigShare or OSF and generated a DOI for the materials to help ensure continued accessibility. Many of these providers provide rather easy ways to link the project with a software development repository to help ensure that the data are available in perpetuity.

Readers should also consult with their local librarians about local resources. Most universities provide a facility for staff to deposit papers and other academic materials, but this may extend to software development repositories in some cases up as well.

Readers are encouraged to avoid using lab websites and other personal pages for sharing projects over the longer term as they are fraught with link-rot and are not an optimal way to share your projects.

# Status in Plant Pathology

## Article Selection

In order for us to determine where we are as a discipline with regards to sharing openly and reproducible papers and data, we randomly selected 300 articles published from 2012 to 2018 across 21 journals (Table 1) representative of the pathology discipline, and assigned them scores reflecting their openness and reproducibility. The journals represented both fundamental and applied areas as well as pathogen or other group specific publications. Three hundred articles were randomly selected from issues published from 2012 to 2018. A list of randomly selected pages was assigned to a randomized list of the 21 journals 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 page 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 from Google’s sheets as a comma separated value (CSV) file for later use and distribution in a research compendium.

## Scoring criteria

Four individuals were assigned to rate a randomised list of journal articles.

Each journals was classified as to whether they were completely open (TRUE), behind a pay wall (FALSE) or a combination (BOTH). The five-year impact factor for 2018 for each journal was retrieved from InCites Journal Citation Reports, Clarivate Analytics and entered in a separate sheet in the Google sheets file. This was downloaded, saved as a CSV file and is left-joined with the article notes using dplyr::left\_join() for analysis when the import\_notes() command is executed.

Articles were classified as fundamental or applied and a note was made whether the article’s primary focus was using or developing molecular techniques (TRUE/FALSE). This classification was based on the evaluator’s own judgement of the article.

Where possible, the software used in the preparation of the publication was recorded in the notes if it was cited or otherwise specified in the article text.

Articles were scored in four areas for reproducibility on a scale of 0 to 3:

* **computational methods available**
  + 0 - Not available or not mentioned in the publication
  + 1 - Uses expensive proprietary software that only institutions would typically purchase (*i.e.*, ArcGIS standard is $7000)
  + 2 - Uses proprietary software that most individuals can afford (*e.g.*, Excel, Matlab? ($500))
  + 3 - Uses entirely open source and free software (*e.g.*, R, Python)
  + NA - No software was used in the research that can be determined as the article is written
* **software available**
  + 0 - Not available or not mentioned in the publication
  + 1 - Uses expensive proprietary software that only institutions would typically purchase (*i.e.*, ArcGIS standard is $7000US)
  + 2 - Uses proprietary software that most individuals can afford (*e.g.*, Excel) or SAS that is now available free for use with a univerisity login
  + 3 - Uses entirely open source and free software (*e.g.*, R, Python).
  + NA - No software was used in the research that can be determined as the article is written
* **software cited**
  + 0 - not mentioned
  + 1 - Software mentioned by name only
  + 2 - Software cited with version number
  + 3 - All software components (SAS PROCs, R or Python packages) etc., cited
  + NA - No software was used in the research that can be determined as the article is written.
* **data available**
  + 0 - Not available or not mentioned in the publication
  + 1 - Available upon request to author
  + 2 - Online, but inconvenient or non-permanent (*e.g.*, login needed, pay wall, FTP server, personal lab website that may disappear, or may have already disappeared)
  + 3 - Freely available online to anonymous users for foreseeable future (*e.g.*, archived using Zenodo, dataverse or university library or some other proper archiving system)

Values for software packages were checked for spelling consistency and corrections were made manually where necessary. When working with these data in R, all strings of software character values were converted to fully uppercase to standardise the capitalisation and alleviate any issues with capitalisation used between evaluators. A custom R function, Reproducibility.in.Plant.Pathology::import\_notes(), was written to import the data, format the columns properly and calculate the overall reproducibility score in R (R Core Team 2020).

A total reproducibility score was calculated as the percent of total possible for a given paper, *e.g.* some papers did not use specialized software therefore they were not scored for this category.

We recognise that the scoring methods were developed by the auhors specifically for this exercise and are not entirely subjective. We have based our scoring decisions on our own experiences in the field of plant pathology. While many software packages may be widely used or “freely” available if one has an institutional licence, these types of software often hinder partners and interested parties that do not have these luxuries. Thus, emphasis was given to using open source alternatives where possible over paid software packages. The view of the authors is one that this is something to be encouraged as it allows the broadest possible audience to benefit from the work at the least cost.

## Analysis

Bayesian mixed effect models were fit using the contributed package ‘brms’ (Bürkner 2018) in the R environment (R Core Team 2020) to evaluate the fixed effects of the publication (journal title), article class, molecular focus, five year journal impact factor, journal open access policy on the reproducibility index and year to test for changes over time in the reproducibility index.

Relatively flat priors were selected with a mean 0 and a standard deviation 10 for the fixed effects.

The random effects were year and assignee for all models excepting the model which tested for year effects. In this model the random effects were the publication title and assignee.

Models were fit using the hurdle\_gamma() family function due to zero-inflated continuous data. Each model was run for 3000 iterations in four chains.

The adapt\_delta() and max\_treedepth() values were adjusted as necessary on a case-by-case basis for each of the models to ensure that the chains mixed well.

Model finesses were evaluated using model summaries and diagnostic plots from ‘brms’ (Bürkner 2018) and posterior fits using pp\_check() from the contributed R package ‘bayesplot’ (Gabry et al. 2019).

Probability of direction or maximum probability of effect was plotted, pd(), and a test for practical equivalence, equivalence\_test(), was performed for each model using the contributed R package ‘bayestestR’ (Makowski et al. 2019).

## Results

Forty-eight articles scored 0 % out of a possible 100 % for reproducibility. The mean value was 31.2 % and the median value was 33.3 %. One article scored 100 %, or was classed as fully reproducible.

No journals surveyed were completely closed, all offered at least an option for open access but some were completely open.

The majority of the articles were classed as fundamental, 160, with the remaining 138 classed as applied.

Most articles did not make any computational methods available in any fashion with two, 0.007 %, classing as “3”, which was the highest score available. Two hundred eighty seven, 96 %, were classed as “0” and 9 articles, 0.03 %, appeared to not use any computational methods.

More articles did a good job of using software that was reasonably available to anyone with 206 scoring “2” or “3” (Table 2).

Software citations were also reasonably well reported in the literature with 120 scoring “2” where the main software was cited with version numbers and 57 scoring “3” where full citations including R packages or the SAS PROC used were cited where appropriate (Table 3).

The most frequently cited software was SAS. The remainder were mostly specialised software with the exception of Microsoft Excel, tied for 9th (Table 4).

# Discussion

* 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.
* Lehner et al. (2017) 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 (<https://emdelponte.github.io/paper-white-mold-meta-analysis/>).
* Khaliq et al. [] offer a full research compendium with a Docker instance and DOI that enable readers to fully replicate the analysis using the data collected and step through other points where weather data were investigated and various models were fit before deciding on the best fit.
* 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

# Acknowledgements

##### pagebreak

# References

Australian Research Council. 2018. ARC Open Access Policy. Available at: <https://www.arc.gov.au/policies-strategies/policy/arc-open-access-policy> [Accessed April 20, 2020].

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

Brunsdon, C. 2015. Quantitative methods I: Reproducible research and quantitative geography. Prog. Hum. Geogr.

Bürkner, P.-C. 2018. Advanced Bayesian multilevel modeling with the R package brms. The R Journal. 10:395–411.

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.

Editors at Nature. 2016. Reality check on reproducibility. Nature. 533:437.

Fidler, F., and Gordon, A. 2013. Science is in a reproducibility crisis: How do we resolve it? Phys.org. Available at: [http://phys.org/news/2013-09-science-crisis.html?utm{\\_}content=bufferde383{\{}{\&}{\}}utm{\\_}source=buffer{\{}{\&}{\}}utm{\\_}medium=twitter{\{}{\&}{\}}utm{\\_}campaign=Buffer{\{}{\#}{\}}ajTabs](http://phys.org/news/2013-09-science-crisis.html?utm%7B\_%7Dcontent=bufferde383%7B\%7B%7D%7B\&%7D%7B\%7D%7Dutm%7B\_%7Dsource=buffer%7B\%7B%7D%7B\&%7D%7B\%7D%7Dutm%7B\_%7Dmedium=twitter%7B\%7B%7D%7B\&%7D%7B\%7D%7Dutm%7B\_%7Dcampaign=Buffer%7B\%7B%7D%7B\#%7D%7B\%7D%7DajTabs).

Fitzjohn, R., Pennell, M., Zanne, A., and Cornwell, W. 2014. Reproducible research is still a challenge. ROpensci.org. 2014 Available at: <http://ropensci.org/blog/2014/06/09/reproducibility/>.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. 2019. Visualization in Bayesian workflow. J. R. Stat. Soc. A. 182:389–402.

Government of Canada. 2016. Tri-Agency Open Access Policy on Publications - Science.Gc.ca. Tri-Agency Open Access Policy on Publications. Available at: <http://www.science.gc.ca/eic/site/063.nsf/eng/h_F6765465.html?OpenDocument> [Accessed April 12, 2019].

Ioannidis, J. P. A. 2014. How to make more published research true. PLoS Medicine. 11.

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., and Ioannidis, J. P. A. 2016. Reproducible research practices and transparency across the biomedical literature. PLoS Biology. 14:1–13.

Lehner, M. S., Pethybridge, S. J., Meyer, M. C., and Del Ponte, E. M. 2017. Meta-analytic modelling of the incidence–yield and incidence–sclerotial production relationships in soybean white mould epidemics. Plant Pathology. 66:460–468.

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.

Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. 2019. bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. Journal of Open Source Software. 4:1541.

McNaught, A. D., and Wilkinson, A. 1997. *Compendium of Chemical Terminology*. Second Edition. Blackwell Science Oxford.

Noorden, R. van. 2017. Gates Foundation demands open access. Nature. 541:270–270.

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>.

Peng, R. D. 2009. Reproducible research and biostatistics. Biostatistics. 10:405–408.

R Core Team. 2020. *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.

Stodden, V., Guo, P., and Ma, Z. 2013. Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. PLoS One. 8:2–9.

Sweedler, J. V. 2015. Striving for reproducible science. Analytical Chemistry. 87:11603–11604.

Wald, C. 2010. Scientists embrace openness.

Weissgerber, T. L., Garovic, V. D., Winham, S. J., Milic, N. M., and Prager, E. M. 2016. Transparent reporting for reproducible science. Journal of Neuroscience Research. 00:1–6.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 3:1–9 Available at: <https://www.nature.com/articles/sdata201618> [Accessed April 20, 2020].

Ziemann, M., Eren, Y., and El-Osta, A. 2016. Gene name errors are widespread in the scientific literature. Genome Biology. 17:177.

##### pagebreak

# Tables

Table: Table 1: Journal titles selected for inclusion and the number (n) of articles from each journal that were evaluated.

| journal | n |
| --- | --- |
| Australasian Plant Pathology | 8 |
| Canadian Journal of Plant Pathology | 13 |
| Crop Protection | 14 |
| European Journal of Plant Pathology | 11 |
| Forest Pathology | 14 |
| Journal of General Plant Pathology | 17 |
| Journal of Phytopathology | 14 |
| Journal of Plant Pathology | 15 |
| Molecular Plant Pathology | 18 |
| Molecular Plant-Microbe Interactions | 16 |
| Nematology | 11 |
| Physiological and Molecular Plant Pathology | 18 |
| Phytoparasitica | 19 |
| Phytopathologia Mediterranea | 9 |
| Phytopathology | 19 |
| Plant Disease | 15 |
| Plant Health Progress | 10 |
| Plant Pathology | 24 |
| Revista Mexicana de Fitopatología | 10 |
| Tropical Plant Pathology | 9 |
| Virology Journal (Plant Viruses Section) | 16 |

##### pagebreak

Table: Table 2: Article ratings for software availability where ‘0’, not available or not mentioned in publication to ‘3’, preferably open source and freely available.

| software availability | n | percent |
| --- | --- | --- |
| 0 | 57 | 0.19655172 |
| 1 | 27 | 0.09310345 |
| 2 | 145 | 0.50000000 |
| 3 | 61 | 0.21034483 |
| NA | 10 | NA |

##### pagebreak

Table: Table 3: Article ratings for software citations where ‘0’, not specified or mentioned in publication to ‘3’, fully cited with proper version numbers and all procedures detailed.

| software citation | n | percent |
| --- | --- | --- |
| 0 | 58 | 0.1986301 |
| 1 | 57 | 0.1952055 |
| 2 | 120 | 0.4109589 |
| 3 | 57 | 0.1952055 |
| NA | 8 | NA |

##### pagebreak

Table: Table 4: Top 10 softwares cited in all articles. Software used is any software package mentioned by name or properly cited in the evaluated articles, n is the number of times it was mentioned. All software names were converted to fully upper-case letters to avoid any inter-evaluator data entry issues.

| software used | n |
| --- | --- |
| SAS | 53 |
| MEGA | 42 |
| SPSS | 26 |
| R | 19 |
| BLAST | 16 |
| STATISTICA | 11 |
| GENSTAT | 10 |
| CLUSTAL W | 9 |
| CLUSTAL X | 7 |
| EXCEL | 7 |
| MRBAYES | 7 |
| BIOEDIT | 5 |
| PAUP | 5 |

##### pagebreak

Table: Table 5: Tally of reproducibility scores where 0 is not reproducible and 12 is fully reproducible, *e.g.* all software cited, software used are freely available to anyone, computational methods (statistical analysis or other computational code) are freely available for download and use, and raw data are available for download and use.

| score | n | percent |
| --- | --- | --- |
| 0.000000 | 53 | 17.7% |
| 8.333333 | 5 | 1.7% |
| 11.111111 | 2 | 0.7% |
| 16.666667 | 7 | 2.3% |
| 25.000000 | 44 | 14.7% |
| 33.333333 | 83 | 27.7% |
| 41.666667 | 52 | 17.3% |
| 44.444444 | 1 | 0.3% |
| 50.000000 | 22 | 7.3% |
| 55.555556 | 1 | 0.3% |
| 58.333333 | 14 | 4.7% |
| 66.666667 | 12 | 4.0% |
| 75.000000 | 2 | 0.7% |
| 91.666667 | 1 | 0.3% |
| 100.000000 | 1 | 0.3% |

##### 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 2020-07-24 21:12:32 using the following computational environment and dependencies:

#> R version 4.0.2 (2020-06-22)  
#> Platform: x86\_64-apple-darwin17.0 (64-bit)  
#> Running under: macOS Catalina 10.15.6  
#>   
#> Matrix products: default  
#> BLAS: /System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versions/A/libBLAS.dylib  
#> LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.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   
#> [7] base   
#>   
#> other attached packages:  
#> [1] rmarkdown\_2.3   
#> [2] rsvg\_2.1   
#> [3] DiagrammeRsvg\_0.1   
#> [4] DiagrammeR\_1.0.6.1   
#> [5] tidyr\_1.1.0   
#> [6] flextable\_0.5.10   
#> [7] janitor\_2.0.1   
#> [8] dplyr\_1.0.0   
#> [9] Reproducibility.in.Plant.Pathology\_0.0.1-1  
#> [10] here\_0.1   
#> [11] officer\_0.3.12   
#> [12] knitr\_1.29   
#>   
#> loaded via a namespace (and not attached):  
#> [1] tidyselect\_1.1.0 xfun\_0.15 purrr\_0.3.4   
#> [4] snakecase\_0.11.0 V8\_3.2.0 vctrs\_0.3.2   
#> [7] generics\_0.0.2 htmltools\_0.5.0 yaml\_2.2.1   
#> [10] base64enc\_0.1-3 rlang\_0.4.7 pillar\_1.4.6   
#> [13] glue\_1.4.1 gdtools\_0.2.2 RColorBrewer\_1.1-2  
#> [16] uuid\_0.1-4 lifecycle\_0.2.0 stringr\_1.4.0   
#> [19] zip\_2.0.4 visNetwork\_2.0.9 htmlwidgets\_1.5.1   
#> [22] evaluate\_0.14 curl\_4.3 fansi\_0.4.1   
#> [25] highr\_0.8 Rcpp\_1.0.5.1 readr\_1.3.1   
#> [28] backports\_1.1.8 jsonlite\_1.7.0 systemfonts\_0.2.3   
#> [31] hms\_0.5.3 packrat\_0.5.0 digest\_0.6.25   
#> [34] stringi\_1.4.6 bookdown\_0.20 rprojroot\_1.3-2   
#> [37] cli\_2.0.2 tools\_4.0.2 magrittr\_1.5   
#> [40] tibble\_3.0.3 crayon\_1.3.4 pkgconfig\_2.0.3   
#> [43] ellipsis\_0.3.1 data.table\_1.12.9 xml2\_1.3.2   
#> [46] lubridate\_1.7.9 assertthat\_0.2.1 rstudioapi\_0.11   
#> [49] R6\_2.4.1 git2r\_0.27.1 compiler\_4.0.2

The current Git commit details are:

#> Local: main /Users/adamsparks/Sources/GitHub/Analysis/Reproducibility\_in\_Plant\_Pathology  
#> Remote: main @ origin (git@github.com:openplantpathology/Reproducibility\_in\_Plant\_Pathology.git)  
#> Head: [57c2f2e] 2020-07-24: Update DOCKERFILE to use verse4.0.0 and script to knit docx