Openness and Computational Reproducibility in plant pathology: where do we stand and a way forward

Adam H. Sparks1,2,✉, Emerson M. Del Ponte3, Kaique S. Alves3, Zachary S. L. Foster4, and Niklaus J. Grünwald5

1 Department of Primary Industries and Regional Development, Farming Systems Innovation, Perth WA 6000, Australia  
2 University of Southern Queensland, Centre for Crop Health, Toowoomba Qld 4350, Australia  
3 Departmento de Fitopatologia, Universidade Federal de Viçosa, Brazil  
4 Department of Botany and Plant Pathology, Oregon State University, Corvallis OR 97331, USA  
5 Horticultural Crops Research Unit, USDA Agricultural Research Service, Corvallis OR 97330, USA

✉ Corresponding author: [Adam H. Sparks <[Adam.Sparks@dpird.wa.gov.au](mailto:Adam.Sparks@dpird.wa.gov.au)>](mailto:Adam.Sparks@dpird.wa.gov.au)

**Keywords: Data Science, Computational Biology, Techniques**

# Abstract

Open research practices have been highlighted extensively during the last ten years in many fields of scientific study as essential standards needed to promote transparency and reproducibility of scientific results. Scientific claims can only be evaluated based on how protocols, materials, equipment and methods were described; data were collected and prepared; and, analyses were conducted. Openly sharing protocols, data and computational code is central for current scholarly dissemination and communication, but in many fields, including plant pathology, adoption of these practices has been slow. We randomly selected 450 articles published from 2012 to 2021 across 21 journals representative of the plant pathology discipline and assigned them scores reflecting their openness and reproducibility. We found that most of the articles were not following protocols for open science, and were failing to share data or code in a reproducible way. We also propose that use of open-source tools facilitates reproducible work and analyses benefitting not just readers, but the authors as well. Finally, we also 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 increasing our basic and applied knowledge of 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 the methods and results. These research findings include biological materials (host and pathogen genotypes), nucleic/protein sequences, experimental and simulated raw data annotations, drawings and photographs and statistical analysis code among other materials and data generated during the course of the research.

Recently, open science has been highlighted with many funding agencies expecting data to be available upon the conclusion of the research project (Government of Canada 2016; Noorden 2017; Australian Research Council 2018), journals in the field promoting the sharing of data (Del Ponte 2020), and other scientists interested in sharing and collaborating more widely (Wald 2010).

Reproducibility is one component under the umbrella of open science. By proactively practising open science, the work increases the chance to become more reproducible through the availability of data and code. That is, open science leads to reproducibility and replicability.

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 was recently highlighted by many authors (Tiwari et al. 2021; Eckert et al. 2020; Dienlin et al. 2020; Preeyanon et al. 2018; Wallach et al. 2018; 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.

In the biological sciences it is not always possible to use identical test material or perhaps the time or resources are not available for full reproducibility, *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 for 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 (2009). *Reproducible research and Biostatistics*. Biostatistics, 10 (3): 405-408.

Therefore our definition of reproducibility will be that the computer code and data are made freely available to others for verification and conducting alternate analyses or for use in instructional purposes. And that the software used are also easily obtained and preferably open source to avoid licensing or other issues related to accessibility for end-users related to costs or non-standard file formats, etc.

Plant pathologists routinely provide information on protocols and chemicals allowing for reproducibility. However, frequently biological specimens such as strains, cultures or cultivars are not available after publication. These cases do constitute a lack of reproducibility but will not be covered here.

# Materials and Methods

## Article Selection and Evaluation Events

In an effort to understand where as a discipline we stand as plant pathologists regarding open science and reproducible research, we surveyed a broad selection of articles to represent a broad swathe of publications to evaluate our status. We hand-picked 21 journals which we felt represented research publications in the field of plant pathology (Table 1) that encompassed a range of subject matter foci, applied and fundamental work, country of origin and ranking metrics, *e.g.* quartile range or citation index, to represent the discipline. The aim was to gather as complete an overview of the status of reproducibility in plant pathology journals as possible and avoid bias in the findings by skewing towards high-impact journals that may have a greater influence. From those 21 journals we randomly selected 450 articles published from 2012 to 2021. Using R (R Core Team 2022), two lists were created, the first was a list of the 21 journals and the second was a list of the evaluators that were evaluating the articles for reproducibility. Initially there were four evaluators, later a fifth was added and the list was recreated. There were three evaluation events that took place to increase the yearly coverage of the evaluations. Each evaluator was assigned 50 articles to evaluate during an evaluation event. A list of randomly generated numbers to represent page numbers from one to 150, sampled with replacement, was assigned to a randomized list of the 21 journals for . This was done because 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.

Articles were then manually selected by visiting the journals’ websites and selecting the articles within which the randomly assigned page numbers fell, *i.e.*, if the page number was 32, the article that started on page 28 and ended on page 35 was selected as it contained page 32. In cases where an article was not suitable, *e.g.*, a review or otherwise not related to plant pathology or the randomly assigned numbers for that journal and year fell within the same article, the next article in that journal was selected until a suitable article was found.

The each of the five authors were each assigned to rate a randomised list of journal articles using scoring criteria devised for the purposes of this research (Table 3) in three different scoring events as their time allowed for their participation. In the first event, three authors (AS, EDP, and ZF) evaluated 200 articles, in the second event two authors evaluated 100 articles (EDP, KA) and in the third round, five authors (AS, EDP, KA, ZF, NG) evaluated 150 articles for a total of 450 articles.

## Scoring Criteria

Each journal was classified as to whether they were completely open (TRUE), behind a paywall (FALSE) or a combination (BOTH). The five-year impact factor for 2022 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 an Open Document Spreadsheet (ODS) file and is left-joined with the article notes using a left\_join() from the ‘dplyr’ package (version 1.0.9) (Wickham et al. 2020) for analysis when a custom function for this analysis, import\_notes(), was 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 judgement of the article.

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

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 upper-case to standardise the capitalisation and alleviate any issues with capitalisation used between evaluators. A custom function, import\_notes(), was written to import the data, format the columns properly and calculate the overall reproducibility score in R (R Core Team 2022).

## Data Cleaning

Efforts were made to standardise the scoring as much as possible. In some cases the software\_avail, or the openness of the software used, score did not agree with the software that was recorded and these scores were modified after discussion to match the agreed on ratings based on the record of the software that was used in each respective paper. For example, a paper was rated as having a software\_avail rating of 3, but the evaluator noted that Excel had been used; here, the software\_avail rating was changed to 2 to bring it into line with the agreed upon guidelines. In other cases, papers did not adequately describe all the software used, e.g., GenAlEx, an Excel add-on, was referred to but Excel was not mentioned in the text. Here, Excel was added to the software\_used column and the software\_avail score was adjusted accordingly.

## Statistical Analysis

The statistical analysis was performed using R version R version 4.2.0 (2022-04-22) (R Core Team 2022) on an Apple MacBook Pro (13-inch, M1, 2020). Inter-rater differences were evaluated using percent agreement and Fliess’ Kappa [ZZZZ cite emerson here] The software availability, computational methods availability and data availability scores were converted from 0-3 to 1-4 ordinal score values and Bayesian mixed-effect models were fit using the contributed package ‘brms’ (version 2.17.0) (Bürkner 2017, 2018, 2021) to evaluate the fixed effects of the publication (journal title), five-year journal impact factor, journal open access policy, journal data and code deposition instructions. The models included ‘year’, ‘journal’ and ‘assignee’ as random effects except in the year of publication analysis where ‘year’ was a fixed effect, publication analysis where the ‘journal’ was a fixed effect and five-year impact factor where correlations between the journal title and impact factor affected the software availability model.

Non-informative priors were selected from a normal distribution with a mean of 0 and a standard deviation of 1.5 were selected for all models.

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

Models were fit using the cumulative(link = "logit") family function for ordinal scale data. Each model was run for 4000 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 a good model fit to the data and that the chains mixed well.

Model finesses were evaluated using model summaries and diagnostic plots from ‘brms’ (version 2.17.0) (Bürkner 2017, 2018, 2021) and posterior fits using pp\_check() (Figure 1) from the contributed R package ‘bayesplot’ (version 1.9.0) (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’ (version 0.12.1) (Makowski et al. 2019) and the Region of Practical Equivalence (ROPE) was set to -0.1 to 0.1.

Data were visualised using ‘ggplot2’ (version 3.3.6) (Wickham 2016), ‘ggpubr’ (version 0.4.0) (Kassambara 2020) and patchwork (version 1.1.1) (Pedersen 2020).

All methods are described in greater detail with the code necessary to reproduce the work in Sparks et al. (2022).

# Results

The differences between the co-authors’ classifications of the article class as applied or fundamental had no differences in the five articles that were evaluated by all of the co-authors (Figure 2A). However, there were several differences in evaluators’ interpretations of whether or not an article had a molecular focus (Figure 2B). All authors agreed on the evaluation of the computational methods and data availability of all five articles that were scored by each author (Figure 3A, 3B).

The analysis detected that the journal title did have an effect on the reproducibility score but there was no effect of article class or impact factor.

No journals surveyed were completely closed, all offered at least an option for open access but some were completely open. The analysis was unable to detect any differences in the effect of access policy on the openness and reproducibility of the evaluated articles, but in journals where the policy was to allow both open and pay-for-fee access, the articles were less likely to be open and reproducible when compared to journals that were completely open.

We were unable to detect any observable differences in the years that publications were submitted. The year 2016 appears to be have a higher score than prior and following years, but the analysis was undecided on the effect of all years on the publications’ openness and reproducibility.

Most journals in our discipline fail to mention the availability of the computational methods to support the article as published. Only two journals, *Plant Pathology* and *Tropical Plant Pathology* have full instructions for authors regarding both data and code deposition in repositories for sharing with readers. Three had no instructions at all, *Forest Pathology*, *Phytopathologia Mediterranea* and *Revista Mexicana de Fitopatología*, with the remaining 16 journals having varying levels of instruction for authors to deposit data but not code. While the analysis did not detect an effect of the journals’ instructions for sharing or depositing data and or code, the more detailed the instructions, the higher the article scored.

The majority of the articles, 235, were classed as fundamental, with the remaining 215 classed as applied.

Most articles did not make any computational methods available in any fashion with 4 (0.009%) classified as ‘3’, which was the highest score available (Figure 4A). One (0%) were classed as ‘0’ and ten articles (2%) appeared to not use any computational methods.

More articles did a good job of using software that was reasonably available to anyone with 248 (55%) scoring ‘2’ or ‘3’ (Figure 4B) for software availability.

Software citations were also reasonably well reported in the literature with 214 (48%) scoring ‘2’ where the main software was cited with version numbers and 81 (18%) scoring ‘3’ where full citations including R packages or the SAS PROC used were cited where appropriate (Figure 4C).

However, data availability was mostly not available with 361 (80%) scoring ‘0’ where the data was not available or mentioned in the article. However, more articles, 45 (10%) scored ‘3’ than scored ‘1’ or ‘2’ combined, 42 (9%) with two articles not producing shareable data (Figure 4D).

In the top ten most frequently cited software, the most frequently cited program was MEGA, where different versions were not distinguished. The next three programs were statistical software, SAS (2nd), SPSS (3rd) and R (4th) with two other statistical programs also frequently cited, GenStat and Statistica (tied 8th). The remainder of the top ten programs we found were related to sequence analysis, *i.e.*, Clustal (5th), BLAST(6th), and BioEdit (10th) or phylogenetic analysis MrBayes (7th). The most general-purpose software cited in the top 10 was Excel® (9th). The Python programming language was cited only twice, not appearing in the top ten.

# Discussion

MEGA was the most widely cited software in this evaluation reflecting the wide spread use of sequence analysis in the field of plant pathology. The second, third and fourth, and eighth most frequently cited software packages were all statistical programs. SAS remains firmly entrenched in the discipline as the choice of software for statistical analysis followed by SPSS but R is not far removed from SPSS in fourth place with Statistica and Genstat tied for eighth. The remainder of the top ten software packages we found were related to sequence analysis, *i.e.*, BLAST, Clustal and BioEdit, phylogenetic analysis MrBayes or the general-purpose spreadsheet program Excel®. It is interesting to note however, that with the popularity of Python that it only appears in the survey twice and in 2020 in both articles, not enough times to be in the top 10 most-cited.

We recognise that the scoring methods were developed by the authors specifically for this evaluation and are not entirely objective and acknowledge that we have based our scoring decisions on our own experiences in the field of plant pathology. However, without assigning values, it is not possible to attempt to qualify or quantify where our discipline stands in regards to computational reproducibility.

In the authors’ view, while many software packages may be widely used or “freely” available if one has an institutional license, these types of software often hinder partners and interested parties that do not have access to them. Thus, the emphasis was given to using open source alternatives where possible over commercially available, non-free software packages. The authors believe that this is something to be encouraged as it allows the broadest possible audience to benefit from the work at the lowest possible monetary cost.

## Are There Good Examples?

While the majority of the articles that were surveyed did not score well, we are aware of and would like to highlight that there are several good examples from the plant pathology discipline that can be used as exemplars.

There are different ways to make research more open and reproducible, which can be thought of as levels of reproducibility. The first level is including tables of raw data or code with the paper as a supplemental file. This is suitable for studies that may have a small data set or simple analysis or for demonstrations purposes as Madden et al. (2015) demonstrate in their discussion regarding the use of values in statistical analyses where they supply an *e-****X****tra*\* with reproducible examples for readers to refer to. This allows the reader to quickly view the extra materials and a DOI is assigned as a part of the article itself.

At the second level, in our search of the literature, we found that Hill et al. (2019) shared code and example data to reproduce the results of the paper as supplementary materials, citing all packages used and the versions. The scripts provided used an R package, *checkpoint* (Ooi et al. 2022), to provide a mechanism to help ensure reproducibility by installing the package versions used by the authors that are necessary to reproduce the work rather than defaulting to the latest versions. However, the scripts did not run unhindered, as the supplementary materials suggested that they should, without changes to the data, but with some modifications to the data, and file name changes, the scripts ran allowing examples of the research to be reproduced using the definition of Peng (2009).

The third level is the use of proper code repositories such as GitHub and data repositories like FigShare, Zenodo or OSF.io among others, allow for the deposition and updating of code, figures, data preprints or any other materials that support the article itself while providing a DOI and citation for a standalone citation. Sparks et al. (Sparks et al. 2011, 2014) used FigShare to provide models, data and code, (Sparks 2016) necessary to replicate model development and the subsequent study on the effects of climate change on potato late blight. Similarly, Carleson et al. (2019) hosted the code for reproducing a population genomic analysis of *Phytophthora plurivora* on GitHub, while providing all data on OSF.io (the Open Science Framework). Lehner et al. (2017) used GitHub to host a code repository of their research compendium website with data and a reproducible report that explains in detail all steps of the analysis and the R code for conducting a meta-analysis for assessing heterogeneity in the relationship between white mold incidence and soybean yield and between incidence and soybean yield (<https://emdelponte.github.io/paper-white-mold-meta-analysis/>). The website clearly demonstrates the analysis to readers and uses R so that anyone can easily replicate the work. Likewise, Duku et al. (2016) provide scripts for models, data and code for graphs, via a website hosted by GitHub, <http://adamhsparks.github.io/MICCORDEA/>, that can be used to replicate their study modelling the effects of climate change on rice bacterial blight and rice leaf blast in Tanzania. However, even this example is not totally open as it requires ArcGIS, an expensive proprietary software, to run the provided Python scripts for the RICEPEST model (Savary and Willocquet 2014). This highlights the difficulty in making work completely open and reproducible, further work could be undertaken to make the Python scripts rely totally on free-open-source-software but the work as it it provided here does allow for the complete reproducibility of the paper as written provided with the proper tools.

The fourth level includes packaging the full analysis in a containerized software application, *e.g.*, Docker. Docker is an open source containerization platform that enables users to package several applications and an operating system into containers, thereby standardising the executable components by combining application source code (or analyses) with the operating system required to run that analysis on any computer. Using an open source language like R, Julia or Python allows us to share our work in a fashion where we know that the analysis will run exactly the same on every computer. As an example of this approach, Khaliq et al. (2020a) provided a research compendium as a Docker container with a DOI and a full R package that enables readers to fully replicate their analysis of *Ascochyta rabiei* conidia dispersal in chickpea using the data collected and stepping through other points where weather data were investigated and various models were fit before deciding on the best fit and recreate any figures as published in the article (Khaliq et al. 2020b). When this level is employed, tools such as Binder, <https://mybinder.org>, can be used that allow readers and reviewers to launch an interactive session in their web browser and interact with the data and rerun the analysis in an RStudio instance or Jupyter notebooks as Mioroni et al. used (Miorini et al. 2018, 2019). Kamvar et al. (2015) took a slightly different approach by including all files necessary for the analysis and most output files in a repository (Kamvar et al. 2014) that also included an installable R package that was used for the original analysis. While there are many other methods, these two approaches illustrate some of the best-practices where the data and other files were deposited in repositories with DOIs and reproducibility issues were addressed by using R packages to handle dependencies and other versioning issues making the work more portable.

## Best Practices to Enhance Reproducible Research

Journals can encourage authors to deposit their data and code by providing instruction for how to deposit the data and code in a repository and mint a DOI to accompany these resources.

A general work flow 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 to the evolution of computational methods and reproducible practices (Figure 5).

The 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 files) 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 paywall. 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 the 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.

When making your science more open and reproducible, methods, software used (this includes items such as R, Julia 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 can 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, Julia, SAS or Python enables 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 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.

Text files for saving small sets of data are preferable. 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) or tab separated (TSV) files. This also helps ensure that the data are reusable. Larger data sets may warrant the use of a proper database like GenBank, MariaDB or PostgreSQL which provide users with several benefits but two important benefits to mention here are (i) data redundancy, ensuring no records are duplicated and (ii) data consistency, ensuring that all records in a data set are recorded in the same format for every observation. While personal 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. Databases such a GenBank are preferable for molecular data ensuring data integrity and machine readability.

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 and this will help support discoveries and support further research. In turn, sharing data will lead to new citations for your work as others discover and use it. To make your data the most widely discoverable and usable, ensure that it has a persistent identifier. A digital object identifier (DOI) is the most common (<https://www.doi.org/>) but the Handle.Net Registry (HNR) (<https://handle.net/>) is also an option. There are different options for generating a DOI for your data and other materials. FigShare, Zenodo and OSF.io all offer persistent archives along with a services 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/>.

Once you have determined how to best manage your source code and the data sets for analysis 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 tracking of changes 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 as these are not an optimal way to preserve and share your work over the long term. Doing either of these leaves the work in an unstable state where future users may be unable to access the work as they are fraught with link-rot and other issues. It is a best practice to always ensure that you have deposited the data with a provider such as GenBank, Zenodo, FigShare or OSF.io 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. If readers are uncertain, we suggest to also consult with local librarians about possible resources. Most universities and other research-focused workplaces provide a facility for staff to deposit papers and other academic materials, but this may extend to software development repositories as well. The important aspect is that data once deposited cannot be modified any more.

In this work we have evaluated the state of computational reproducibility in the plant pathology literature and presented suggestions for areas of improvement. As we prepared this letter, we became more aware of the urgent need to spread and establish an open science attitude and culture among plant pathologists. To assist in fostering this sort of change in our discipline, Open Plant Pathology (OPP), <https://www.openplantpathology.org>, an institution-independent and non-funded initiative, was founded in January 2018 by two of this letter’s co-authors, Del Ponte and Sparks, in 2018 with the following vision: “foster a diverse community culture that values openness, transparency and reproducibility of scientific research data and methods in our field”. We started OPP with a minimal infrastructure and support from other enthusiastic colleagues that allows members to interact sharing and gathering ideas on how we can improve the openness and reproducibility in our discipline.

We believe that adopting an attitude of open and collaborative science and using the best reproducibility practices in our daily work, directly benefits us as researchers. For example, between complicated analyses, reviews and revisions and questions years later about the data that was collected or analysis that was conducted, it is extremely beneficial to be able to easily reproduce your work quickly and easily. This letter was drafted over the course of several years as the authors had time to devote to it. Having everything in a reproducible framework made it easy to resume work and set-aside as necessary without losing information and having everything well-documented made it more simple to do this. Second, it benefits the reviewer by aiding their understanding of the work done and gives them more materials to use to make suggestions for improvements when reviewing and the end-user or reader is better able to verify the validity of the methods used and recreate the analysis. Perhaps more importantly, sharing these details helps with knowledge transfer by showing other interested parties how something was done rather than simply describing it. Lastly, openly sharing your work and making it discoverable can lead to new collaborations and synergistic ideas. One of the most important messages that we would like to share is that there is more to the work than just the paper. Sharing materials detailing the analysis that was performed and documenting the data provide citable products and enhance the manuscript providing the reader with a richer set of information with which to understand the work that was performed. This open sharing of code and data leads to greater impact as work is cited if resources such as code or data are reused. With mandates for openly sharing data becoming more common with funding agencies, we can and should embrace this and move the discipline forward and have a greater impact with our work.

# Authors’ Contributions

A.H.S, E.DP., Z.F. and N.J.G. conceived of the presented idea and scoring system. E.DP. provided five-year impact factors. A.H.S., E.DP., Z.F. K.A. and N.G. evaluated articles for scoring. A.H.S. designed the computational framework for analysis and created the research compendium with feedback from E.DP. and Z.F. A.H.S. wrote the manuscript in consultation with E.DP. All authors provided critical feedback and helped contribute to the final version of the manuscript.

# Data and Code Availability

## Data

The raw data for this work are documented and available from DOI [https://doi.org/10.5281/zenodo.4941722](10.5281/zenodo.4941722).

## Code

All code used in the analyses and data visualization and associated materials have been made available as research compendium available from DOI [https://doi.org/10.5281/zenodo.1250664](10.5281/zenodo.1250664). A webpage version of the compendium is available from <https://openplantpathology.github.io/Reproducibility_in_Plant_Pathology/>.

# Acknowledgements

The authors are grateful for insightful comments from Dr. David Ferris and the USQ Centre for Crop Health Advisory Group on the final manuscript.

# Literature Cited

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 reproducibility crisis](https://doi.org/10.1038/nature.2016.20076). Nature News.

Brunsdon, C. 2015. [Quantitative methods I: Reproducible research and quantitative geography](https://doi.org/10.1177/0309132515599625). Prog. Hum. Geogr.

Bürkner, P.-C. 2018. [Advanced Bayesian multilevel modeling with the R package brms](https://doi.org/10.32614/RJ-2018-017). The R Journal. 10:395–411.

Bürkner, P.-C. 2021. [Bayesian item response modeling in R with brms and Stan](https://doi.org/10.18637/jss.v100.i05). Journal of Statistical Software. 100:1–54.

Bürkner, P.-C. 2017. [brms: An R package for Bayesian multilevel models using Stan](https://doi.org/10.18637/jss.v080.i01). Journal of Statistical Software. 80:1–28.

Carleson, N. C., Fieland, V. J., Scagel, C. F., Weiland, J. E., and Grünwald, N. J. 2019. Population structure of *Phytophthora plurivora* on rhododendron in Oregon nurseries. Plant disease. 103:1923–1930.

Del Ponte, E. M. 2020. [Towards a more open and transparent plant pathology research](https://doi.org/10.1007/s40858-020-00388-z). Tropical Plant Pathology. 45:361–362.

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., et al. 2020. [An agenda for open science in communication](https://doi.org/10.1093/joc/jqz052). Journal of Communication. 71:1–26.

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](https://doi.org/10.1007/s10584-015-1580-2). Climatic Change. 135:569–583.

Eckert, E. M., Cesare, A. D., Fontaneto, D., Berendonk, T. U., Bürgmann, H., Cytryn, E., et al. 2020. [Every fifth published metagenome is not available to science](https://doi.org/10.1371/journal.pbio.3000698). PLOS Biology. 18:e3000698.

Editors at Nature. 2016. [Reality check on reproducibility](https://doi.org/10.1038/533437a). Nature. 533:437.

Fidler, F., and Gordon, A. 2013. Science is in a reproducibility crisis: How do we resolve it? Phys.org. Available at: <https://phys.org/news/2013-09-science-crisis.html>.

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](https://doi.org/10.1111/rssa.12378). 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].

Hill, G. N., Beresford, R. M., and Evans, K. J. 2019. [Automated analysis of aggregated datasets to identify climatic predictors of botrytis bunch rot in wine grapes](https://doi.org/10.1094/phyto-10-17-0357-r). Phytopathology. 109:84–95.

Ioannidis, J. P. A. 2014. [How to make more published research true](https://doi.org/10.1371/journal.pmed.1001747). 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](https://doi.org/10.1371/journal.pbio.1002333). PLoS Biology. 14:1–13.

Kamvar, Z. N., Larsen, M. M., Kanaskie, A. M., Hansen, E. M., and Grünwald, N. J. 2015. [Spatial and temporal analysis of populations of the sudden oak death pathogen in Oregon forests](https://doi.org/10.1094/PHYTO-12-14-0350-FI). Phytopathology. 105:982–989.

Kamvar, Z. N., Larsen, M. M., Kanaskie, A. M., Hansen, E. M., and Grünwald, N. J. 2014. [Sudden Oak Death in Oregon Forests: Spatial and temporal population dynamics of the sudden oak death epidemic in Oregon Forests](https://doi.org/10.5281/zenodo.13007).

Kassambara, A. 2020. *Ggpubr: ’ggplot2’ based publication ready plots*. Available at: <https://CRAN.R-project.org/package=ggpubr>.

Khaliq, I., Fanning, J., Melloy, P., Galloway, J., Moore, K., Burrell, D., et al. 2020a. [The role of conidia in the dispersal of *Ascochyta rabiei*](https://doi.org/10.1007/s10658-020-02126-2). European Journal of Plant Pathology. 158:911–924.

Khaliq, I., Melloy, P., Burrell, D., Fanning, J., Galloway, J., Moore, K., et al. 2020b. *ChickpeaAscoDispersal: A research compendium to accompany “the role of conidia in the dispersal of Ascochyta rabiei”*. Available at: <https://doi.org/10.5281/zenodo.3810826>.

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](https://doi.org/10.1111/ppa.12590). 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?](https://doi.org/10.1094/PHYTO-07-15-0165-LE) 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](https://doi.org/10.21105/joss.01541). Journal of Open Source Software. 4:1541.

Miorini, T. J. J., Kamvar, Z. N., Higgins, R. S., Raetano, C. G., Steadman, J. R., and Everhart, S. E. 2019. [Differential aggressiveness of *Sclerotinia sclerotiorum* isolates from North and South America and partial host resistance in Brazilian soybean and dry bean cultivars](https://doi.org/10.1007/s40858-018-00273-w). Tropical Plant Pathology. 44:73–81.

Miorini, T. J. J., Kamvar, Z. N., Higgins, R., Raetano, C. G., Steadman, J., and Everhart, S. E. 2018. Data and analysis for variation in pathogen aggression and cultivar performance against *Sclerotinia sclerotiorum* in soybean and dry bean from the U.S. And Brazil. Available at: <https://osf.io/2x7fc/> [Accessed July 29, 2020].

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

Ooi, H., Vries, A. de, and Microsoft. 2022. *checkpoint: Install packages from snapshots on the Checkpoint server for reproducibility*. Available at: <https://CRAN.R-project.org/package=checkpoint>.

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>.

Pedersen, T. L. 2020. *Patchwork: The composer of plots*. Available at: <https://CRAN.R-project.org/package=patchwork>.

Peng, R. D. 2009. [Reproducible research and biostatistics](https://doi.org/10.1093/biostatistics/kxp014). Biostatistics. 10:405–408.

Preeyanon, L., Pyrkosz, A. B., and Titus Brown, C. 2018. [Reproducible bioinformatics research for biologists](https://doi.org/10.1201/9781315373461-7). In Chapman; Hall/CRC, p. 185–217.

R Core Team. 2022. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.

Savary, S., and Willocquet, L. 2014. Simulation modeling in botanical epidemiology and crop loss analysis. The Plant Health Instructor. Available at: <https://www.apsnet.org/edcenter/disimpactmngmnt/topc/BotanicalEpidemiology/Pages/default.aspx>.

Sparks, A. 2016. Global-late-blight (meta)modelling. Available at: <https://figshare.com/articles/dataset/Global_Late_Blight_Modelling/1066124/6>.

Sparks, A. H., Del Ponte, E. M., Alves, K. S., Foster, Z., and Grünwald, N. J. 2022. Compendium of r code and data for ‘status and best practices for reproducible research in plant pathology. Available at: <https://doi.org/10.5281/zenodo.1250665>.

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](https://doi.org/10.1890/ES11-00128.1). 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](https://doi.org/10.1094/PDIS-04-11-031). 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](https://doi.org/10.1371/journal.pone.0067111). PLoS One. 8:2–9.

Sweedler, J. V. 2015. [Striving for reproducible science](https://doi.org/10.1021/acs.analchem.5b04300). Analytical Chemistry. 87:11603–11604.

Tiwari, K., Kananathan, S., Roberts, M. G., Meyer, J. P., Shohan, M. U. S., Xavier, A., et al. 2021. [Reproducibility in systems biology modelling](https://doi.org/10.15252/msb.20209982). Molecular Systems Biology. 17.

Wald, C. 2010. Scientists embrace openness. Available at: <https://www.science.org/content/article/scientists-embrace-openness> [Accessed May 1, 2022].

Wallach, J. D., Boyack, K. W., and Ioannidis, J. P. A. 2018. [Reproducible research practices, transparency, and open access data in the biomedical literature, 20152017](https://doi.org/10.1371/journal.pbio.2006930) ed. Ulrich Dirnagl. PLOS Biology. 16:e2006930.

Weissgerber, T. L., Garovic, V. D., Winham, S. J., Milic, N. M., and Prager, E. M. 2016. [Transparent reporting for reproducible science](https://doi.org/10.1002/jnr.23785). Journal of Neuroscience Research. 00:1–6.

Wickham, H. 2016. *ggplot2*. Springer-Verlag GmbH. Available at: <https://www.ebook.de/de/product/26537420/hadley_wickham_ggplot2.html>.

Wickham, H., François, R., Henry, L., and Müller, K. 2020. *Dplyr: A grammar of data manipulation*. Available at: <https://CRAN.R-project.org/package=dplyr>.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. 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](https://doi.org/10.1186/s13059-016-1044-7). Genome Biology. 17:177.

# Tables

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

| Journal | n |
| --- | --- |
| Australasian Plant Pathology | 11 |
| Canadian Journal of Plant Pathology | 19 |
| Crop Protection | 23 |
| European Journal of Plant Pathology | 19 |
| Forest Pathology | 18 |
| Journal of General Plant Pathology | 24 |
| Journal of Phytopathology | 19 |
| Journal of Plant Pathology | 23 |
| Molecular Plant Pathology | 29 |
| Molecular Plant-Microbe Interactions | 22 |
| Nematology | 16 |
| Physiological and Molecular Plant Pathology | 24 |
| Phytoparasitica | 23 |
| Phytopathologia Mediterranea | 20 |
| Phytopathology | 28 |
| Plant Disease | 24 |
| Plant Health Progress | 16 |
| Plant Pathology | 30 |
| Revista Mexicana de Fitopatología | 19 |
| Tropical Plant Pathology | 18 |
| Virology Journal (Plant Viruses Section) | 25 |

Table 2: The top ten sofware programs that were found to be used in the papers that were surveyed.

| Software | Frequency |
| --- | --- |
| MEGA | 73 |
| SAS | 70 |
| SPSS | 48 |
| R | 41 |
| CLUSTAL | 28 |
| BLAST | 22 |
| EXCEL | 17 |
| MRBAYES | 16 |
| GENSTAT | 14 |
| STATISTICA | 14 |
| BIOEDIT | 10 |

Table 3: Articles were scored in four areas for reproducibility on a scale of ‘0’ to ‘3’ by five independent evaluators to rank articles in each of four areas, ‘Computational methods availability’, ‘Data availability’, ‘Software availability’ and ‘Software citation’.

| Category | Index Value | Definition |
| --- | --- | --- |
| Computational methods availability | 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, paywall, 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 |
|  | NA | No software was used in the research that can be determined |
| Data availability | 0 | Not available or not mentioned in the publication |
|  | 1 | Available upon request to the author |
|  | 2 | Online, but inconvenient or non-permanent, e.g., login needed, paywall, 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 |
| Software availability | 0 | Not available or not mentioned in the publication |
|  | 1 | Uses expensive proprietary software or requires getting a quote, >100USD/year, e.g., ArcGIS or Minitab |
|  | 2 | Uses proprietary software that most individuals can afford, <100USD/year, e.g., Excel, SAS with "SAS on Demand for Academics" version or free but not open source software, e.g., MEGA |
|  | 3 | Uses entirely free open source software, e.g., R, Julia, Python, QGIS, LibreOffice |
|  | NA | No software was used in the research that can be determined |
| Software citation | 0 | Not mentioned |
|  | 1 | Software mentioned by name only |
|  | 2 | Software cited with version number |
|  | 3 | All software components, e.g. SAS PROCs, R, Julia or Python packages, etc. cited |
|  | NA | No software was used in the research that can be determined as the article is written |

# Figures

![Figure 1: Posterior distribution visualisations for each of six models fitted to scoring data that were used to evaluate factors on reproducibility of 450 papers published in 21 plant pathology journals or plant pathology focused articles from other specialised journals. Six factors were evaluated, (A) Publication (Journal Title), (B) Article Class, (C) Five Year Impact Factor of Journal, (D) Journal Open Access Policy, (E) Year of Publication and (F) the Journals’ Reproducibility Instructions where y is the data’s density and yrep are ten random posterior draws. Model fitnesses were found to be good for all models and no effects of any of the factors were observed on the reproducibility score.](data:application/eps;base64,)

Figure 1: Posterior distribution visualisations for each of six models fitted to scoring data that were used to evaluate factors on reproducibility of 450 papers published in 21 plant pathology journals or plant pathology focused articles from other specialised journals. Six factors were evaluated, (A) Publication (Journal Title), (B) Article Class, (C) Five Year Impact Factor of Journal, (D) Journal Open Access Policy, (E) Year of Publication and (F) the Journals’ Reproducibility Instructions where y is the data’s density and yrep are ten random posterior draws. Model fitnesses were found to be good for all models and no effects of any of the factors were observed on the reproducibility score.

![Figure 2: Evaluator classifications of (A) an articles focus being molecular techniques (TRUE/FALSE) and (B) an article’s class (Fundamental/Applied). Where all five articles were scored by all five authors to check for agreement between the evaluations.](data:application/eps;base64,)

Figure 2: Evaluator classifications of (A) an articles focus being molecular techniques (TRUE/FALSE) and (B) an article’s class (Fundamental/Applied). Where all five articles were scored by all five authors to check for agreement between the evaluations.

![Figure 3: Evaluator agreement for five articles scored for (A) computational methods (code) used, e.g. scripts or other self-written programs were freely and openly available and (B) data represented in the paper were freely and openly available. Note that one evaluator rated an article with computational methods available as ‘NA’, which is reflected in that article having only four evaluations for this criteria.](data:application/eps;base64,)

Figure 3: Evaluator agreement for five articles scored for (A) computational methods (code) used, e.g. scripts or other self-written programs were freely and openly available and (B) data represented in the paper were freely and openly available. Note that one evaluator rated an article with computational methods available as ‘NA’, which is reflected in that article having only four evaluations for this criteria.

![Figure 4: Aggregated article scores for each of the three categories evaluated, (A) presents ‘Software availability’, where ‘0’ was ‘Not available or not mentioned in the publication’; ‘1’ was ‘Expensive proprietary software that only institutions would typically purchase’; ‘2’ was ‘Used proprietary software that most individuals can afford (e.g., Excel) or SAS that is now available free for use with a university login’; and ‘3’ was ‘Used entirely open source and free software (e.g., R, Julia, Python)’; ‘NA’ indicates that no specialised software was necessary or used. (B) displays ‘Code availability’, where ‘0’ was ‘Not available or not mentioned in the publication’; ‘1’ was ‘Available upon request to the author; ’2’ was ‘Online, but inconvenient or non-permanent (e.g., login needed, paywall, FTP server, personal lab website that may disappear, or may have already disappeared)’; and ‘3’ was ‘Freely available online to anonymous users for foreseeable future (e.g., archived using Zenodo, dataverse or university library or some other proper archiving system)’; ‘NA’ indicates that no code was created to conduct the work that was detectable. (C) shows ‘Data availability’, where ‘0’ was ‘Not available or not mentioned in the publication’; ‘1’ was ‘Available upon request to the author; ’2’ was ‘Online, but inconvenient or non-permanent (e.g., login needed, paywall, FTP server, personal lab website that may disappear, or may have already disappeared)’; and ‘3’ was ‘Freely available online to anonymous users for foreseeable future (e.g., archived using Zenodo, dataverse or university library or some other proper archiving system)’; ‘NA’ indicates that no data were generated, e.g., a methods paper.](data:application/eps;base64,)

Figure 4: Aggregated article scores for each of the three categories evaluated, (A) presents ‘Software availability’, where ‘0’ was ‘Not available or not mentioned in the publication’; ‘1’ was ‘Expensive proprietary software that only institutions would typically purchase’; ‘2’ was ‘Used proprietary software that most individuals can afford (e.g., Excel) or SAS that is now available free for use with a university login’; and ‘3’ was ‘Used entirely open source and free software (e.g., R, Julia, Python)’; ‘NA’ indicates that no specialised software was necessary or used. (B) displays ‘Code availability’, where ‘0’ was ‘Not available or not mentioned in the publication’; ‘1’ was ‘Available upon request to the author; ’2’ was ‘Online, but inconvenient or non-permanent (e.g., login needed, paywall, FTP server, personal lab website that may disappear, or may have already disappeared)’; and ‘3’ was ‘Freely available online to anonymous users for foreseeable future (e.g., archived using Zenodo, dataverse or university library or some other proper archiving system)’; ‘NA’ indicates that no code was created to conduct the work that was detectable. (C) shows ‘Data availability’, where ‘0’ was ‘Not available or not mentioned in the publication’; ‘1’ was ‘Available upon request to the author; ’2’ was ‘Online, but inconvenient or non-permanent (e.g., login needed, paywall, FTP server, personal lab website that may disappear, or may have already disappeared)’; and ‘3’ was ‘Freely available online to anonymous users for foreseeable future (e.g., archived using Zenodo, dataverse or university library or some other proper archiving system)’; ‘NA’ indicates that no data were generated, e.g., a methods paper.

![Figure 5: An example of an open and reproducible research workflow. Starting with the question, determine the methodology, describe it, make it available and cite it. Data are used in analysis and any binary files or code are made available as supplements to the manuscript. Source code and raw data are made available in a public repository, preferably with version control for tracking changes through time, and a DOI for final released products.](data:application/eps;base64,)

Figure 5: An example of an open and reproducible research workflow. Starting with the question, determine the methodology, describe it, make it available and cite it. Data are used in analysis and any binary files or code are made available as supplements to the manuscript. Source code and raw data are made available in a public repository, preferably with version control for tracking changes through time, and a DOI for final released products.