Preprinting Microbiology

Patrick D. Schloss†

† To whom correspondence should be addressed: pschloss@umich.edu

Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI

1 ## Abstract

2 ## Importance

Background. Many scientists, including microbiologists, have begun to use preprints and other online venues such as social media, blog posts, and videos as methods to garner attention for their research and to engage the public. A preprint is an unpublished manuscript that is made publicly available without going through an official peer-review process. An author can post their manuscript to a preprint server for others to read, share, and make comments. Preprints were initially adopted among physicists in the 1960s as a method of sharing interesting research amongst colleagues. Eventually, this community evolved into what is now the *arXiv* (pronounced "archive") preprint server that was hosted at the Los Alamos National Laboratories from 1991 to 1999 and then at Cornell University. Among physicists and mathematicians, posting a preprint to *arXiv* has become the standard publication strategy followed by submission to a peer-reviewed journal. Within those communities researchers have been recognized with international awards for work that has only been posted to *arXiv* without formal review. Although *arXiv* has hosted a number of computational biology papers, the server has not drawn widespread attention from biologists. Among proponents of *arXiv*, preprints have aided in the development of research communication by accelerating the release of the science and helping it to achieve a wider audience for critique and reception. Considering the broadening adoption of preprints among microbiologists, we sought to explore the specific uses and concerns of preprints amongst microbiologists.

Landscape of preprint servers. In 2013, two preprint servers the *bioRxiv* (pronounced "bio-archive") and *PeerJ Preprints* were launched to parallel the success of *arXiv*. Both platforms offer similar features: preprint posting is free; each preprint receives a digital object identifier (DOI) that facilitates the ability to cite preprints in other scholarly work; if the preprint is ever published, the preprint is linked to the published version; the submission process for both options is relatively simple allowing authors to upload a PDF version of their preprint and supplemental materials; preprints are typically publicly available in under 24 hours; they have built in venues for authors to discuss their research with people who leave comments on the manuscript; preprints undergo a basic screening process to remove submissions with offensive or non-scientific content; and the sites provide article-level metrics indicating the number of times an abstract has been accessed or the preprint has been downloaded. There are several important differences between the two options. First, *PeerJ Prints* is a for-profit organization and *bioRxiv* is a non-profit organization

sponsored by Cold Spring Harbor Laboratory. This difference can be meaningful to authors since some journals, including the ASM Journals, will only accept submissions that have been posted on preprint servers hosted by non-profit organizations. Second, preprints at *PeerJ Preprints* are posted under the Creative Commons Attribution License (CC-BY) and *bioRxiv* preprints can be posted under one of four CC-BY licenses or with no permission for reuse. This can be relevant for authors hoping to submit their work a journal as journals will not consider manuscripts posted as preprints under a CC-BY license (e.g. *Proceedings of the National Academy of Sciences*). A cosmetic, but important difference between the two is the layout and feel of the two websites. Compared to the *bioRxiv* site, the *PeerJ Preprint* site is more fluid, gives readers the ability to "follow" a preprint, and provides better access to article keywords and the ability to search preprints. With broader acceptance of preprints by traditional journals, many journals, including all of the ASM journals, have established mechanisms to directly submit manuscripts that are posted as preprints on *bioRxiv*. It is only possible to transfer a *PeerJ Preprint* for submission to *PeerJ*. In many ways, preprint servers have taken on the feel of a journal. As adoption of this approach to disseminating research expands, it is likely that the features of these sites will continue to improve. It is also likely that interfaces from third-parties will improve. For example, although Google Scholar includes preprints hosted at *bioRxiv* and *PeerJ Preprints* in their search results, PubMed and Web of Science do not. There is hope that the National Institutes of Health (NIH) will renew their interest in including preprints in PubMed search results.

Acceptance of preprints by journals. An early controversy encountered by researchers interested in posting their work as preprints as a stage in disseminating their research was whether it constituted prior publication. The broad consensus at this point is that preprints do not constitute prior publication. The current policies of journals that commonly publish microbiology research including those published by ASM, the Microbiology Society, International Society for Microbial Ecology, PLOS, the *Proceedings of the National Academy of Science*, *Science*, and *Nature* have a permissive stance towards prior posting of preprints prior to submission. Although journals published by *Cell* do not forbid authors from posting preprints prior to submission, they ask authors to consult an editor prior to posting and do not allow authors to post revised preprints that contain revisions that respond to editorial input. Considering the relatively fluid nature of many of these

policies and the journals' specific policies, prospective authors should be aware of the positions taken by the journals where they may eventually submit their work. Comprehensive lists of journals' attitudes towards preprints are available online and are regularly updated.

Preprints and peer-review. The use of preprints for citations in other scientific reports and grant proposals has been called into question because preprints upend the traditional peer-review editorial process. It is important to note that this process was adapted to the technologies and trends that have evolved over the past 100 years. The formal peer-review system that most journals currently use was not developed until the end of the 1800s with the advent of typewriters and carbon paper (doi:10.1016/S0167-7799(02)01985-6). Editorial decisions were typically made by a single person or a committee (i.e. the editorial board) who had an expertise that covered the scope of the journal. As people's science became more specialized, new journals would form to support and provide a source of validation to the new specialty. The growth in science in the mid 1900s resulted in a shift from difficulties finding sufficient numbers of manuscript to publish to having too many manuscripts submitted. It has been argued that the widespread adoption of decentralized peer-review was due to the increased specialization and to deal with the large number of manuscript submissions (JAMA 1990;263:1323-1329). Peer-review did not achieve widespread use at many journals, including the *Journal of Bacteriology*, until the 1940s and 1950s. Given the rapid advances in communication technology and even greater specialization, it is worth pondering whether the current scientific publishing system and peer-review system, in particular, need to continue to adapt with our science.

Communicating research has traditionally been done within research group meetings, departmental seminars, conferences, and as publications. Along this continuum, there is an assumption that the quality of the science has been improved because it has been vetted by more experts in the field. The public dissemination of one's research is a critical component of the scientific method. By describing their research, scientists subject their work to formal and informal peer review. Their research is scrutinized, praised, and probed to identify questions that help seed the next iteration of the scientific method. A common critique of more modern approaches to publishing has been an inability to assess the quality of the science without the validation of peer-review. Attached to assertions of the validity of the research has become assertions of the impact and robustness of the research. These are all quality assessments that many acknowledge are difficult to assess

by the traditional peer-review process. This has led to some journals, most notably *PLOS ONE*, calling for referees to place a reduced emphasis on the perceived impact or significance of the work. It has also led to the call for replacing or complementing pre-publication peer-review with post-publication peer-review using PubMed Commons, PubPeer, journal-based discussion forums, and other mechanisms. Alas if scientists are going to depend on post-publication peer-review or informal methods of peer-review for documents like preprints, they must be willing to provide constructive feedback on the work of others.

Preprints have the potential to change the advancement of science. Preprints are often viewed as existing in a state of scientific limbo. As noted above, they represent a formal communication, but are not officially published. As the use of preprints grows and our perceptions of preprints matures, there are a number of issues that will need to be addressed. First, a common concern is that if a researcher posts their work as a preprint, it will be "scooped" by another researcher and the original researcher will lose their ability to claim primacy or their ability to publish the work in a journal. Considering the preprint is a citable work with a DOI, it would, in fact, be the preprint author that scooped the second. A growing number of scientific societies and journals, including ASM view preprints as citable and as a legitimate claim to primacy. Some worry that with such protection a researcher can make a claim without valid data to support their claims. This is possible; however, it is also the responsibility of the scientific community to utilize the peer-review mechanisms that are available to comment on those preprints pointing out methodological problems or to indicate that they are speaking beyond the data. A second area of concern is whether a preprint can be used to support a grant proposal. Given the length limitations of many funding agencies, there is a push to cite previous work to indicate a research team's competence in an area or to provide preliminary data. Some fear that the use of preprints will allow some to circumvent page limits by posting preliminary manuscripts. We would hope that both consumers of preprints and grant proposal reviewers would be able to differentiate between someone trying to game the system and someone that is using preprints as a mechanism to improve their science. A third concern is what role preprints should have in assessing a scientist's productivity. Clearly use of publication metrics is a contentious topic without considering the place of preprints. Regardless, given the propensity for researchers to list manuscripts as being "in

preparation" or "in review" on an application or curriculum vitae, listing them instead as preprints that can be reviewed by a committee could be seen as a significant way to enhance an application. In fact, several funding agencies are starting to encourage fellowship applicants to include preprints in their materials. Others are mandating that researchers post preprints for all of their work prior to submitting the work to a journal. Beyond these concerns, preprints are also causing some to change their publication goals. Some authors are explicitly stating that a preprint will not be submitted to a journal. Although these authors may be a minority of those who post preprints, such an approach may be attractive to those that want to have a mechanism for people to cite a report of a brief research communication, a critique of another publication, or negative results. It is clear that the adoption of preprints will challenge how scientists interact and evaluate each other's work. There is great potential to empower researchers by controlling when a citable piece of work is made public.

Microbiology anecdotes. The peer-review editorial process can be lengthy and adversarial. In contrast, preprints represent a rapid and potentially collaborative method for disseminating research. Several anecdotes from the microbiology literature are emblematic of benefits of the rapid release cycle that is inherent in the use of preprints. First, preprints have proven useful for rapidly disseminating results for disease outbreaks and new technologies. Prior to the recent Zika virus outbreak there were approximately 50 publications that touched on the biology and epidemiology of the virus; the number of Zika virus-related publications is now over 1,700. During this time, more than 110 Zika virus-related preprints have been posted at *bioRxiv*. Any manuscript that was published went through several month delays in releasing information to health care workers, the public, and scientists needing to learn new methods to study a previously obscure virus. In contrast, those that posted their work as a preprint were able to disseminate their methods and results instantly. Over the last several years rapid advances in DNA sequencing technologies have fundamentally changed how microbial science is performed. One notable technology, the MinIon sequencing platform from Oxford Nanopore has received considerable attention from researchers who post new methods and results to preprint servers. For such a rapidly developing technology, the ability to share and consume methods from other scientists has created a feed forward effect where the technology has likely advanced at a faster rate than it otherwise would have. Second,

preprints have proven useful for rapidly correcting the scientific literature. On February 9, 2015, Cell Systems posted a study by Afshinnekoo et al. online. The study collected and analyzed metagenomic sequence data from the New York City subway system and reported finding *Yersinia pestis* and *Bacillus anthracis*. Because of the focus on these two bioterrorism agents, this study generated a considerable amount of press attention. On April 25, 2015, Petit et al. posted a preprint to Zenodo demonstrating that there was no evidence for *B. anthracis* in the dataset. On July 29, 2015, a critique was published by Cell Systems along with a response from the original authors offering a correction to their manuscript. A second anecdote surrounds the publishing of a draft tardigrade genome in *PNAS*. On November 23, 2015 a study of Boothby et al. was first published online. The authors posited that 17.5% of its genes came from bacteria, archaea, fungi, plants, and viruses. Another group had been analyzing sequence data from a parallel tardigrade genome sequencing project and did not observe the same result. By December 1, 2015, the second group had posted a preprint comparing the two genome sequences and demonstrating that the exciting claims of horizontal gene transfer were really the product of contaminants; this analysis would eventually be posted online by the original journal on March 24, 2016 followed by a rebuttal by the original authors on May 31, 2016. Two other analyses of the original data were published in May 2016 and a third was posted as a preprint on February 2, 2016. These anecdotes underscore the value of having a rapid posting cycle to correcting errors in the scientific literature and that results posted to preprint servers were able to correct the record within weeks of the initial publication while the traditional path took six months in both cases. A final notable case where preprints have accelerated the correction of the scientific record was a preprint posted by Bik et al. reporting numerous cases of image manipulation in peer reviewed studies. Their preprint was posted on April 20, 2016 and published in *mBio* on June 7, 2016. Instead of using preprints to react to published papers that have been through peer review, it would be interesting to consider how the editorial process for these examples and the Arsenic life paper would have been different had they initially been posted as preprints.

Original Online: 9 February 2015 http://www.sciencedirect.com/science/article/pii/S2405471215000022
Uh, no online: 29 July 2015 http://www.sciencedirect.com/science/article/pii/S2405471215000162
Correction: 29 July 2015 http://www.sciencedirect.com/science/article/pii/S2405471215000150

Read preprint: April 25, 2015 https://zenodo.org/record/17158#.WIprNbGZOiI

Bremm PNAS: May 12, 2016. http://www.pnas.org/content/113/22/E3054.full Arakawa PNAS: May 31, 2016. http://www.pnas.org/content/113/22/E3057.full Delmont PeerJ: February 2, 2016. https://peerj.com/preprints/1695/

Specific challenges for microbiology. Although preprints offer an efficient and novel venue for disseminating microbiology research, there are several considerations that the scientific community and those that oversee preprint servers must consider. It is critical that assurances be given that policies are in place to address these issues. First, a significant amount of attention has been given to the potential dual use of microbiology research for individuals seeking to engage in terrorist activities. Second, for researchers engaging in research that involves human subjects it is critical that assurances be made that institutional review boards have been consulted and have approved of the research. Third, there is significant concern regarding researchers hiding potential conflicts of interest that could affect a project's experimental design, analysis, and interpretation of results. Finally, recent expansions in scientific publishing have revealed numerous cases of plagiarism. Microbiology journals have policies in place to address these issues that should be easily implemented by preprint servers.

Metrics for microbiology-affiliated preprints. To analyze the use of preprints, we downloaded the *bioRxiv* on December 31, 2016. We chose to analyze the use of *bioRxiv* preprints because these preprints are amenable for submission to ASM journals and there were 7,434 preprints compared to the 2,650 available at *PeerJ Preprint*. Among the 7,434 preprints on bioRxiv, 329 were assigned by the authors into the Microbiology category. One limitation of the *bioRxiv* interface is the inability to assign manuscripts to multiple categories or to tag the content of the preprint. For example, this manuscript could be assigned to either the Microbiology or the Scientific Communication and Education categories. To counter this limitation, we developed a more permissive approach that classified preprints as being microbiology-affiliated if their title or abstract had words containing *yeast*, *fung*, *viral*, *virus*, *archaea*, *bacteri*, *microb*, *microorganism*, or *pathogen*. We identified 1,228 additional manuscripts that we considered microbiology-affiliated. These microbiology-affiliated preprints were primarily assigned to the Evolutionary Biology (N=221), Genomics (N=184), or

Bioinformatics (N=182) categories. As the total number of preprints has grown exponentially since the creation of *biorxiv*, submission of microbiology-affiliated preprints has largely followed this growth (**Figure 1A**).

Although preprints are still relatively new, the collection of microbiology-affiliated preprints indicates widespread experimentation with the format and considerable geographic diversity. Reflecting the still relatively novelty of preprints, 1,132 (86.15%) corresponding authors who submitted a microbiology-affiliated manuscript (N=1,314 total) have posted a single preprint and 3.58% have posted 3 or more preprints. Corresponding authors that have posted microbiology-affiliated preprints are from 60 countries and are primarily affiliated with institutions in the United States (50.80% of microbiology-affiliated preprints), United Kingdom (11.95%), and Germany (4.17%). As the preprint format matures, it will be interesting to see whether the fraction of authors that post multiple preprints increases and whether the geographic diversity amongst those authors is maintained.

As stated above, preprints offer researchers the opportunity to improve the quality of their work by adding a more formal and public step to the scientific process. Among the microbiology-affiliated manuscripts, 146 (9.32%) had been commented on at least once and only 35 (2.23%) more than three times using the *bioRxiv*-hosted commenting feature. Although the hosted commenting is only one mechanism for peer review, this result was somewhat disturbing since the preprint model implicitly depends on people's willingness to offer others feedback. Regardless, authors do appear to be incorporating feedback from colleagues or editorial insights from journals as 404 (25.80%) of preprints were revised at least once. Among the preprints posted prior to January 1, 2016, 31.61% of the Microbiology category preprints, 35.12% of the microbiology-affiliated preprints, and 33.79% of all preprints have been published. As noted above, not all authors submit their preprints to journals. This would indicate that the "acceptance rates" are actually higher. Regardless, considering that these acceptance rates are higher than many peer-reviewed journals (e.g. approximately 20% at ASM Journals), these results dispel the critique that preprints represent overly preliminary research.

Measuring the impact and significance of scientific research is notoriously difficult. Using several metrics we sought to quantify the effect that microbiology-affiliated preprints have had on the work

of others. Using the download statistics associated with each preprint, we found that the median number of times an abstract or PDF had been accessed was 923 (IQR: 603 to 1445) and 303 (IQR: 167 to 568), respectively. These values represent two aspects of posting a preprint. First, they reflect the number of times people were able to access science before it was published. Second, they reflect the number of times people were able to access a version of a manuscript that is published behind a paywall. To obtain a measure of a preprint's ability to garner attention and engage the public, we obtained the Altmetric Attention Score for each preprint (**Figure 1B**). The Altmetric Attention Score measures the number of times a preprint or paper is mentioned in social media, mainstream media, peer reviews, and policy documents. A higher score indicates that a preprint received more attention. Microbiology-affiliated preprints have had a median Altmetric Attention Score of 7.28 (IQR: 3.25 to 16.3495) and those of all preprints hosted at *bioRxiv* have had a median score of 7.05 (IQR: 3 to 15.556). For comparison, the median Altmetric Attention Score for articles published in *mBio* published since November 2013 was 4.45 (IQR: 1.25 to 13.604). Of all scholarship tracked by Altmetric, the median Altmetric Attention Score for preprints posted at *biorxiv* ranks at the 86 percentile (IQR: 66 to 94. A more traditional and controversial metric of impact has been the number of citations an article receives. We obtained the number of citations for the published versions of manuscripts that were initially posted as preprints. To allow for a comparison to traditional journals, we considered the citations for preprints published in 2014 and 2015 as aggregated by Web of Science (**Figure 1C**). Among the preprints that were published and could be found in the Web of Science database, the median number of citations was 6.50 (IQR: 2-14; mean: 13.60). For comparison, for the papers published in *mBio* in 2014 and 2015, the median number of citations was 5 (IQR: 2-9; mean: 6.73). Although it is impossible to quantify the quality or impact of research with individual metrics, it is clear that preprints and the publications that result from them are broadly accepted by the microbiology community

Preprints from an author's perspective. Posting research as a preprint gives an author great control over when their work is made public. Under the traditional peer-review model, an author may need to submit and revise their work multiple times to several journals over a long period before it is finally published. In contrast, an author can post the preprint at the start of the process for others to consume and comment on as it works its way through the editorial process. A first

example illustrate the utility of preprints for improving access to research and the quality of its reporting. In 2014, the Schloss laboratory posted a preprint to *PeerJ Preprints* describing a method of sequencing 16S rRNA gene sequences using the Pacific Biosciences sequencing platform. At the same time, they submitted the manuscript for review at *PeerJ*. While the manuscript was under review, we received feedback from an academic scientist and from scientists at Pacific Biosciences that the impact of the results could be enhanced by using a recently released version of the sequencing chemistry. Instead of ignoring this feedback and resubmitting the manuscript to address the reviews, we generated new data and submitted an updated preprint a year later with a simultaneous submission to *PeerJ* that incorporated the original reviews as well as the feedback we received from Pacific Biosciences and the academic scientist. In other cases, our research group has simultaneously posted our manuscripts as preprints and submitted them to a journal. As a second example, this manuscript was posted to *bioRxiv* as a preprint on **XXXXXXX XX, 2017**. We then solicited feedback on the manuscript using social media. Two weeks later, we incorporated the comments and posted a revised preprint and submitted the manuscript to *mBio*. During that time, the abstract was read **XXXX** times and the PDF was accessed **XXXX** times. This process engaged **XXXX** commenters on *bioRxiv*, **XXXX** people on Twitter, **XXXX** on Facebook, and **XXXX** via email. We received useful feedback from **XXX** people. Compared to the two or three scientists that typically review a manuscript, this experience engaged a much larger and more diverse community than had we foregone the posting of a preprint. Although there are concerns regarding the quality of the science posted to a preprint server, we contend that responsible use of preprints as a part of the scientific process can significantly enhance the science.

Preprints from a publisher's perspective. A lingering question is what role traditional journals have in disseminating research if there is broad adoption of preprints. One perspective is that although authors value immediate accessibility to their work, they also value the editorial support from professional societies and other organizations. Furthermore, the professional copyediting, layout, and publicity that these publishers offer are also unique features of traditional journals. An alternative perspective is that preprints will eventually replace traditional journals. Certainly, this is a radical perspective, but it does serve to motivate publishers to address what they can offer preprint authors. By adopting preprint-friendly policies, journals can create an attractive environment for

authors. As discussed above, a growing number of journals have created mechanisms for authors to directly submit preprints to their journals. A new venture from *mSphere*, mSphereDirect, actively encourages authors to post their manuscripts as preprints as part of the author-driven editorial process. In addition to integrating preprints into the traditional editorial process, several professional societies that publish journals have also explicitly supported citation of preprints in their other publications and recognize the priority of preprints in the literature. These are policies that empower authors and make specific journals more attractive. Other practices have great potential to improve the reputation of journals. As measured above, preprints are able to garner attention on par with papers published in highly selective microbiology journals. Thus, it is in a journal's best interest to recruit these preprints to their journals. Several journals including **XXXXXX** and **XXXXXX** have publicly stated that they scout preprints for this purpose. Preprints can also be viewed as a lost opportunity to journals. A preprint that garners significant attention may be ignored when it is finally published, brining little additional attention to the journal. Going forward, it will be interesting to see whether publishers are able to develop innovative approaches so that they can benefit by incorporating preprints into their process or if their influence is reduced by the widespread adoption of preprints.

Conclusions. An increasing number of microbiologists are posting their unpublished work to preprint severs as an efficient method of disseminating their research prior to peer review. A number of critical concerns remain about how widespread their adoption will be, how they will be treated by traditional journals, and whether peer-review will adapt to new scientific trends and technologies. Regardless, preprints should offer a great opportunity for both scientists and journals to publish high quality science.

Acknowledgements

This work was supported in part by funding to PDS from the National Institutes of Health (P30DK034933). We appreciate the support of Altmetric, Inc and Thompson Reuters who provided advanced programming interface (API) access to their databases. The data, code, and an executable R-markdown version of this manuscript are available at

http://www.github.com/SchlossLab/Schloss_PrePrints_mBio_2017.

Figures

Figure 1. Summary of microbiology-affiliated preprints since the creation of *biorxiv.* The total number of preprints posted for each quarter ending December 31, 2016 has largely tracked the overall submission of preprints to *biorxiv* (A). The Altmetric attention scores of preprints posted to *biorxiv* are similar to those published in *mBio* since November 2013 indicating preprints engender a similar level of attention (B). The number of times preprints that were published in 2014 and 2015 have been cited is similar to the number of citations for papers published in *mBio* in 2014 and 2015 indicates that published preprints are frequently cited (C). Regions with common background shading in A are from the same year. The vertical lines in B and C indicate the median Altmetric impact score and the median number of citations.

References