bioRxiv: the preprint server for biology

Richard Sever[1], Ted Roeder, Samantha Hindle, Linda Sussman, Kevin-John Black, Janet Argentine, Wayne Manos, and John R. Inglis[1]

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York, USA

[1] Corresponding authors: sever@cshl.edu; inglis@cshl.edu

Abstract

The traditional publication process delays dissemination of new research, often by months, sometimes by years. Preprint servers decouple dissemination of research papers from their evaluation and certification by journals, allowing researchers to share work immediately, receive feedback from a much larger audience, and provide evidence of productivity long before formal publication. Launched in 2013 as a non-profit community service, the bioRxiv server has brought preprint practice to the life sciences and recently posted its 64,000th manuscript. The server now receives more than four million views per month and hosts papers spanning all areas of biology. Initially dominated by evolutionary biology, genetics/genomics and computational biology, bioRxiv has been increasingly populated by papers in neuroscience, cell and developmental biology, and many other fields. Changes in journal and funder policies that encourage preprint posting have helped drive adoption, as has the development of bioRxiv technologies that allow authors to transfer papers easily between the server and journals. A bioRxiv user survey found that 42% of authors post their preprints prior to journal submission whereas 37% post concurrently with journal submission. Authors are motivated by a desire to share work early; they value the feedback they receive, and very rarely experience any negative consequences of preprint posting. Rapid dissemination via bioRxiv is also encouraging new initiatives that experiment with the peer review process and the development of novel approaches to literature filtering and assessment.

Introduction

Dissemination of scientific manuscripts has traditionally occurred only after the research has been formally evaluated by scientific journals. In the print era, the high marginal costs associated with distribution favored this coupling of evaluation and dissemination; only manuscripts that passed a certain bar set by the journal were published and incurred printing costs. The resulting delays to dissemination have often prompted scientists to share draft manuscripts informally among close colleagues and more organized mechanisms for sharing preprints widely were piloted as early as the 1960s (Cobb, 2017).

Over the years, concerns about delayed dissemination have become more acute. The routine requirement among journals for external peer review has become universal only in the past few decades and authors increasingly feel that the demands made by reviewers and editors are lengthening the publication process still further (Vale, 2015). Moreover, the timescale of journal publication, which can take months to years (Royle, 2015), is increasingly at odds with the timescales on which scientists, in particular early career researchers, must demonstrate productivity when evaluated for appointments, tenure and grants (Sarabipour et al., 2019).

The advent of the Web offered an opportunity to decouple the dissemination of papers from their subsequent evaluation and certification by journals. The costs of dissemination online are significantly lower, reducing the financial argument for disseminating only peer-reviewed papers; online dissemination is almost immediate; and anyone with an Internet connection can view the work. The arXiv preprint server, launched in 1991 and currently hosted by Cornell University, has demonstrated the effectiveness of this approach (Ginsparg, 2011). Researchers in physics, computational science, mathematics, and various other disciplines routinely post manuscripts on arXiv prior to peer review and by mid-2019 the site had posted more than 1.5M papers. Several attempts were made to replicate the approach in the biological sciences (Marshall, 1999; Nature Publishing Group, 2012; Rawlinson, 2019). These were unsuccessful in part because of opposition from traditional publishers but also because there was little interest among biologists. More recently, however, the increasing pace of research, increasing dissatisfaction with delays caused by peer review, restricted availability of many published papers, and a general growth in enthusiasm for more openness and transparency in science communication have refocused attention on the potential for preprints in biology. bioRxiv was launched in 2013 in the hope that rapid sharing of biology preprints would eliminate delays to dissemination (Kaiser, 2013) and in doing so increase the pace of research itself (Quake, 2019). The purpose of this article is to summarize bioRxiv's progress and potential and provide a general reference for the project.

The launch of bioRxiv

bioRxiv is an initiative of Cold Spring Harbor Laboratory (CSHL), a non-profit research institute with a unique international reputation as both a leading research institute and a hub for scientific communication. CSHL has been a meeting place for scientists for more than 100 years and a center of professional scientific education for more than 50 years. The annual CSHL Symposium was central to the birth of molecular biology and genomics, and conferences at CSHL continue to attract thousands of scientists every year. The laboratory also has significant publishing expertise, as the originator of classic books and manuals and several academic journals. It was therefore a natural steward for a community preprint server for life sciences and the initiative received strong encouragement from the laboratory's leadership. bioRxiv was launched in 2013 following discussions with members of the academic community, librarians, and arXiv, many of whom would join the project as Advisory Board members and Affiliate scientists (see https://biorxiv.org/about-biorxiv). Notably, following consultations with representatives of arXiv,

the project was named "bioRxiv", not "bio-arXiv", to reduce the likelihood that users would mistakenly contact arXiv staff for bioRxiv technical support.

Technical basis for bioRxiv

Given the potentially vast number of biology preprints — several hundred thousand papers each year — it was clear that bioRxiv would require an industrial scale architecture that could process and display a high volume of submissions and stably accommodate millions of online readers with minimal downtimes. bioRxiv's hosting and manuscript management sites would have to include state-of-the-art features biologists had come to expect of online journals and be able to accommodate both existing and future integrations with other participants in the scholarly communication ecosystem (e.g. search engines, indexing services, journals, and manuscript submission systems). After defining the specifications required, we partnered with HighWire Press, a company developed within and part-owned by Stanford University that had a proven record of more than 20 years in online manuscript hosting and technology development for clients including the American Academy for the Advancement of Science (AAAS) and The National Academy of Sciences (NAS).

The submission side of bioRxiv is based on a BenchPress submission system adapted for preprint handling and automated transfer to the display site. The display side is based on modified HighWire Drupal technology. Additional customization by CSHL developers uses CSS, JavaScript and external databases to enhance and supplement the display on the site and provide additional feeds and services. In addition, the site is integrated with the third-party Disqus and Hypothesis commenting/annotation tools. A significant difference from traditional journals is that the architecture needs to accommodate the ability to upload revised versions of papers at any time (Fig. 1). All preprints are assigned a single digital object identifier (DOI). Each version of the preprint receives a unique URL, with the DOI for the preprint defaulting to the most recent version of the paper posted (see below). Articles can be cited by DOI or version-specific URL identifier.

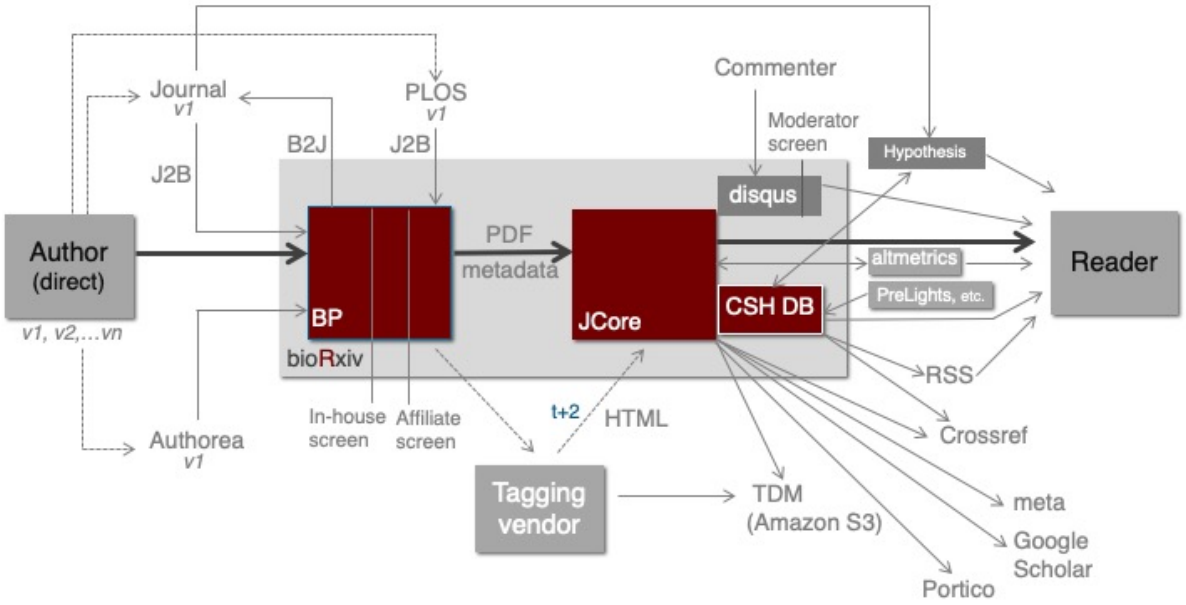

Fig. 1. bioRxiv workflow. Authors submit papers through a BenchPress (BP) module either directly or via J2B from journals and other services. Manuscripts are then displayed publicly via the HighWIre Drupal module (JCore). Cold Spring Harbor Laboratory databases (CSH DB) augment display and pass information to third parties. Papers are screened in BP and can also be sent to journals via B2J. Tagged HTML is generated by a compositor.

bioRxiv is committed to permanency of the content posted. All content is therefore also deposited with the archiving service Portico, a not-for-profit organization committed to long-term preservation of scholarly material.

Preprint screening

A defining feature of bioRxiv is that it does not perform peer review. Nevertheless, there is a need to screen papers to minimize the chance of

posting of inappropriate material and maximize the content's utility to readers. The bioRxiv screening process acts as a coarse filter for non-scientific/pseudoscientific content, non-biological/biomedical content, and potentially harmful content, as well as manuscripts solely comprising isolated data elements, and non-research articles such as recipes, textbook excerpts, narrative reviews and speculative theory. The decision to decline articles other than research papers, no matter how worthy, was a pragmatic one aimed at maximizing screening efficiency. It reduces subjectivity in screening and recognizes the reality that it is research rather than review/didactic content that suffers the distribution delays bioRxiv is intended to address (Sever, 2019).

bioRxiv screening is a two-stage process performed in a highly customized BenchPress environment. Papers first undergo an internal screen by bioRxiv staff, which includes automated plagiarism checks using Similarity Check software and search engines, as well as manual checks for spam and clearly inappropriate or incomplete content. Submissions are then further screened by a distributed group of bioRxiv Affiliates, all of whom are experienced scientists with principal investigator or equivalent positions. This ensures that every article posted on bioRxiv has been viewed by a scientist. It is important to emphasize, however, that the screening process is a coarse, quick filter intended to minimize the likelihood that readers will encounter content that is not *bona fide* biological research. It does not guarantee or certify the content in any way, and readers must use their own judgment in assessing its validity as science.

Initially bioRxiv was intentionally restricted to basic biology: any clinical work was excluded. This restriction was partially lifted in 2015/6 with the introduction of a pilot in which clinical research could be posted in two specific areas: epidemiology and registered clinical trials. Such papers had a specific screening process involving a group of medically qualified bioRxiv Clinical Affiliates. In 2019, the success of this pilot resulted in the launch of a dedicated preprint server for clinical research, medRxiv (Bloom, 2019; Rawlinson, 2019), and the bioRxiv Epidemiology and Clinical Trials subject categories stopped accepting new papers.

Preprint features

During the submission process authors upload either a complete article as a PDF file or a combination of Microsoft Word and figure files, which are then automatically converted into a single PDF. Manually entered article information generates the HTML metadata that is viewable when the article first appears online (see below). Authors may also upload additional supplemental files, such as movies or supplementary figures and tables. DOIs are assigned after the authors have approved the PDF for posting. Starting in late 2019, the DOI suffix for new postings will include this approval date, so the date the preprint was first approved can easily be viewed within citations (akin to a journal year and volume number). Article screening and posting typically takes 24–48 hours, barring any issues that need to be addressed by the authors before posting or occasional delays due to weekends or holiday periods. Papers initially post as author PDFs together with author-entered metadata and supplementary material. Full-text HTML generated by an outside compositor is added 24–48 hours later and includes in-line figures and linked references (Fig. 2).

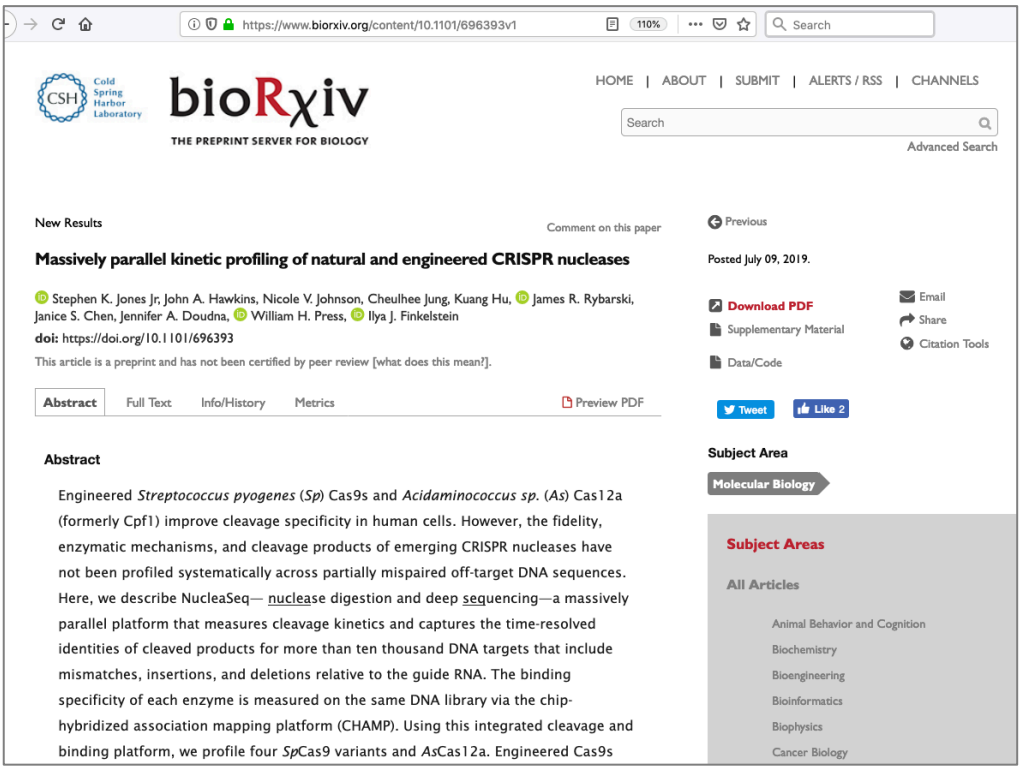

Fig. 2. Screenshot showing a bioRxiv article HTML view. Content is displayed in full-text HTML, along with a download link for the authors' PDF file and a variety of additional information.

Other elements displayed include the single subject category and article type (New Results, Contradictory or Confirmatory Results) selected by the authors, the article history (links to prior versions), and the authors' choice of terms under which they wish to make the article available. These include various Creative Commons licenses, 'all rights reserved', CC0 Public Domain dedication, and a specific US government Public Domain option required for NIH employees. In addition to standard article metadata, authors may also provide ORCIDs and links to

externally hosted data sets or code within a dedicated field. For revised versions of articles, they can also include a revision summary (version note) describing the changes they have made during revision. Additional elements viewable alongside articles include links to the final journal version of record when this appears (Fig. 3), accepted article notifications for participating publishers, article-level metrics and altmetrics, online comments, and links to third-party coverage elsewhere on the Web (see below).

Indexing and discovery

DOIs assigned to bioRxiv articles are all deposited with the DOI-registration agency Crossref on the day of posting. Once the article is published in a journal, bioRxiv adds a link to the formally published version alongside the preprint and updates the Crossref DOI record with this information, which is subsequently available via bioRxiv (api.biorxiv.org) and Crossref (api.crossref.org) APIs. bioRxiv identifies preprint–journal article matches through a variety of scripts that search PubMed and Crossref databases for title and author matches. Matched authors are then alerted and have the opportunity to remove the link if the match is incorrect and/or supply matches for articles that have not been identified by bioRxiv scripts. bioRxiv extends this approach to articles that have been retracted from journals, so this information can also be displayed alongside relevant preprints.

bioRxiv includes numerous built-in search and alert features and is indexed by a variety of third-party discovery tools. Readers can browse the site by subject category or using the Solr-powered search feature within the hosting site. Personalized email alerts for specific search terms can also be generated, and subject-category-specific RSS feeds and Twitter accounts provide additional mechanisms for content alerts. Additional personalization is planned for the near future. bioRxiv is indexed by generic search engines, as well as the dedicated literature-discovery engines Google Scholar and Microsoft Academic. It is also indexed by Europe PubMed Central, the AI-powered biomedical discovery tool Meta, and the Rxivist (Abdill & Blekhman, 2019). A variety of APIs are planned to further facilitate additional search and alert services by third parties, along with a dedicated text and data mining (TDM) repository.

Manuscript transfer

To reduce the burden on authors who wish to submit to both bioRxiv and journals—and to further encourage preprint posting—we have developed bioRxiv-to-journal (B2J) and journal-to-bioRxiv (J2B) streams that allow authors to transfer articles between bioRxiv and journal submission systems. This means that authors need only upload files and manually enter core metadata once, saving them significant time and effort, although some journals require additional journal-specific metadata following B2J that must be entered separately at the journal submission site. B2J and J2B use the standard File Transfer Protocol (FTP) to transmit a ZIP archive containing XML metadata and manuscript files in a way that can easily be generated/ingested by journal submission systems. B2J and J2B pre-date, and in some ways inspired, the Manuscript Exchange Common Approach (MECA; Sack, 2018), a recommended new approach for transferring submissions between journals. Work is currently underway to adapt B2J and J2B to the MECA protocol.

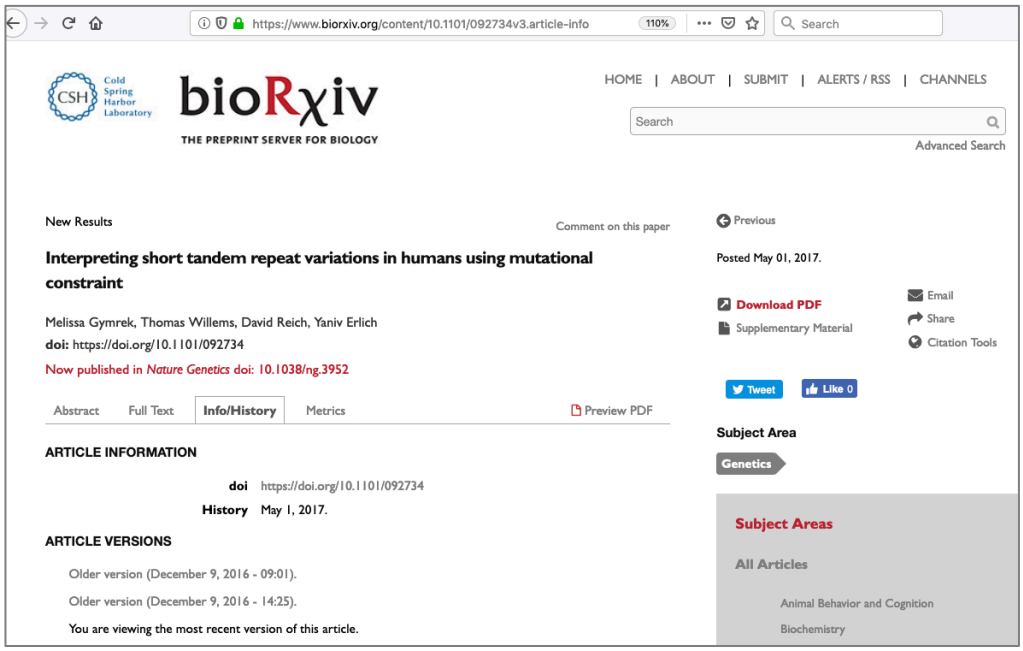

Fig. 3. Screenshot showing the bioRxiv article Info/History tab. Links to earlier versions of the article show the evolution of the article over time. Once a formal journal version of record appears, this is prominently linked in red below the bioRxiv DOI.

The B2J and J2B manuscript transfer services are not just available to journals. B2J has been used to transfer papers from bioRxiv to the journal-independent peer review services Axios (now closed) and Peerage of Science, and J2B and B2J will be used by the new portable peer-review service Review Commons. Meanwhile, authors who drafted papers using the authoring platform Authorea have been able to submit to bioRxiv directly from this service via J2B, and this may represent a model for similar tools.

Withdrawals

Manuscripts posted on bioRxiv receive DOIs and thus are citable and part of the scientific record. In addition, they are indexed by third-party services, creating a permanent digital presence independent of bioRxiv records. Consequently, bioRxiv's policy is that papers cannot be

removed, except in exceptional cases for legal reasons or matters of biosecurity.

Authors can, however, have articles marked as "Withdrawn" if they no longer stand by the findings/conclusions or if they acknowledge fundamental errors in the article. In these cases, the default view becomes a withdrawal statement providing an explanation for the withdrawal, but the original article is still accessible via the article history tab. In rare cases, an article can be withdrawn by bioRxiv itself as a consequence of unethical behavior by an author or a technical error made by bioRxiv or its technology partners.

Withdrawn articles are clearly identified within the bioRxiv website. Ensuring that this signal is perpetuated within the ecosystem and that such withdrawals are effectively identified, indexed and displayed by third-party services is an area currently being investigated.

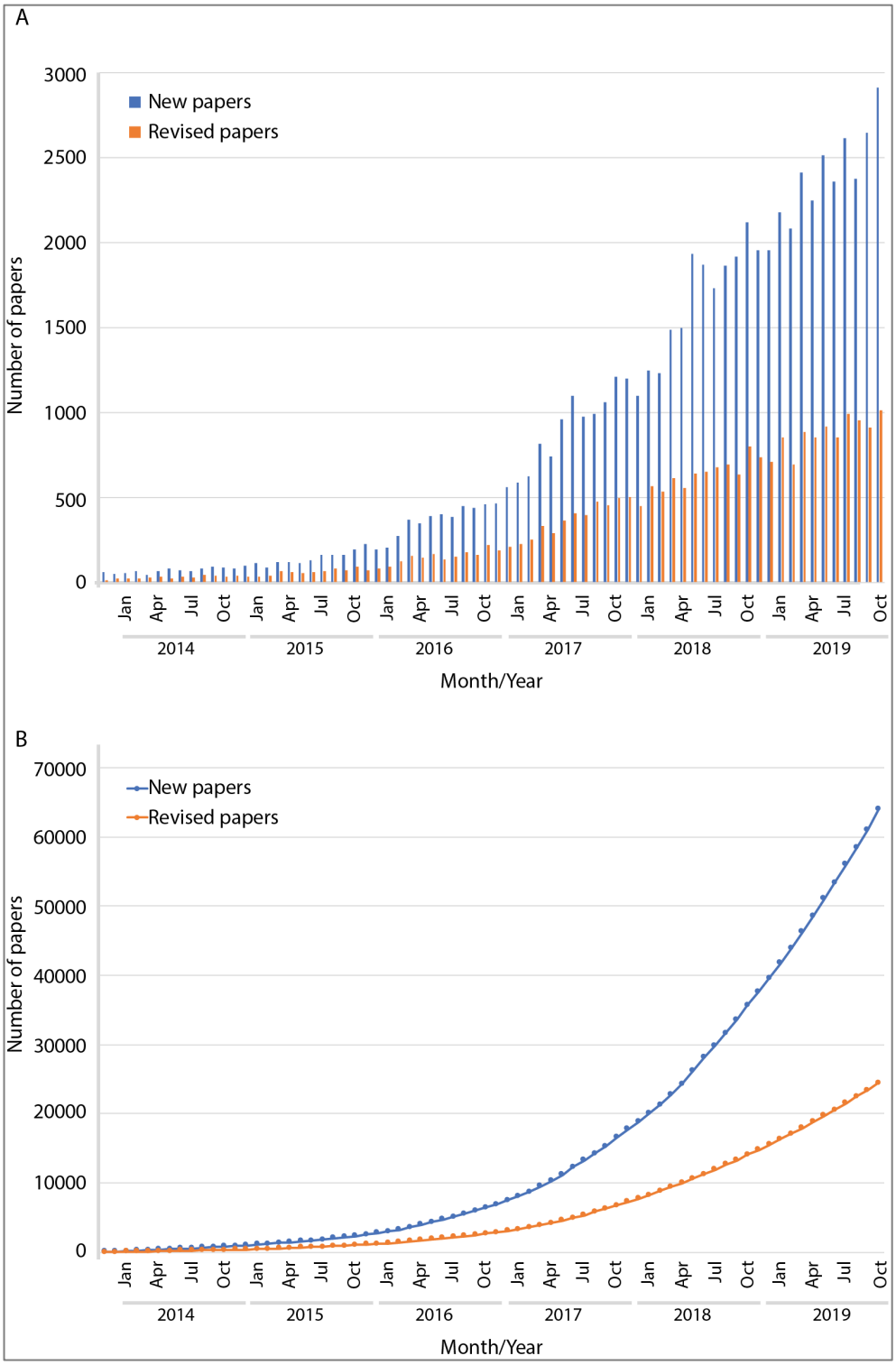

Fig. 4. The growth of bioRxiv. A. Monthly submissions to bioRxiv. New articles are in blue; revised articles are in red. B. Total number of papers on bioRxiv. New articles are in blue; revised articles are in red.

bioRxiv by the numbers

Below we summarize a series of data sets related to preprints posted on bioRxiv. The numbers are current at the time of writing, but we wish to alert readers to the fact that many of these metrics are updated in real time and available to interested readers at api.biorxiv.org.

Figure 4 shows the number of bioRxiv posts since 2013. Over this period submissions grew considerably from a handful to more than 2900 per month in 2019 (Fig. 4A). At the time of writing, the total number of first submissions to bioRxiv is more than 64,000 (Fig. 4B). The proportion of manuscripts that are revised has remained fairly constant at 25%–30%. Most papers are revised only once (if at all) but some are revised multiple times (8% have two revisions; 2% have three revisions; <1% have four or more revisions). Only 59 papers have been withdrawn to date; however, we note that the withdrawal option was introduced only in 2018 and so its existence may not be widely known among authors.

Table 1 shows the fractions of articles within different subject categories across a five-year period. Initially bioRxiv was dominated by papers in genomics, bioinformatics and evolutionary biology, but the percentages contributed by other subdisciplines have increased, most notably in neuroscience (Table 1). This is consistent with the experience of arXiv, which was initially dominated by high-energy physics but subsequently began to attract papers from other disciplines in large numbers (Ginsparg, 2011). bioRxiv preprints have been deposited by authors from 130 different countries, the most common being the USA, UK and Germany. The most prolific institutions are Stanford University, University of Oxford, and University of Cambridge (Table 2). The distribution of licenses chosen by authors has changed little over time: currently 35%

Table 1 – bioRxiv papers by subject category

Subject Category	2013 (%)	2014 (%)	2015 (%)	2016 (%)	2017 (%)	2018 (%)	2019 (%)
Animal Behavior and Cognition*		1.15	1.14	1.55	1.62	1.71	1.51
Biochemistry	0.92	0.57	1.14	2.09	2.78	3.08	3.17
Bioengineering	0.92	0.57	1.31	0.79	1.48	2.06	2.52
Bioinformatics	12.8	18.3	17.8	15.1	11.33	9.28	7.91
Biophysics	3.67	2.53	2.11	3.85	4.08	4.48	4.42
Cancer Biology	4.59	1.72	2.68	2.06	3.13	3.45	3.6
Cell Biology	1.83	2.07	1.14	3.47	4.74	5.06	5.26
Clinical Trials**			0.06	0.04	0.17	0.23	0.13
Developmental Biology	2.75	2.53	1.77	1.87	2.93	3.38	2.99
Ecology	8.26	4.94	4.79	4.94	3.84	3.96	4.41
Epidemiology***				1.66	2.85	2.94	2.33
Evolutionary Biology	15.6	20.2	18.5	11.2	7.43	5.59	5.13
Genetics	8.26	8.15	10.7	7.72	6.68	5.08	4.86
Genomics	11.0	15.0	13.5	10.8	7.61	6.19	5.46
Immunology	1.83	1.49	0.8	1.09	1.57	2.66	3.24
Microbiology	2.75	1.61	3.3	5.53	6.39	8.95	9.07
Molecular Biology	1.83	1.61	1.14	1.96	2.57	3.44	3.64
Neuroscience	8.26	7.12	8.26	14.9	18.6	18.3	19.0
Paleontology	0	0.11	0.17	0.09	0.13	0.17	0.14
Pathology	0	0.11	0.28	0.28	0.3	0.55	0.63
Pharmacology and Toxicology*		0.23	0.28	0.45	0.73	0.8	1.09
Physiology	0	0.69	0.51	0.51	0.95	1.18	1.67
Plant Biology	4.59	1.26	1.94	2.26	2.76	2.63	3.24
Scientific Comm. & Education	0	0.57	0.46	0.74	0.6	0.7	0.82
Synthetic Biology	4.59	1.95	1.42	1.09	1.08	0.88	0.83
Systems Biology	3.67	4.71	4.16	3.7	3.19	2.79	2.4
Zoology	1.83	0.8	0.68	0.38	0.42	0.46	0.56

*introduced 2014
**introduced 2015
***introduced 2016

CC BY-NC-ND, 32% all rights reserved, 19% CC BY, 7% CC BY-NC, 6% CC BY-ND, 1% CC0/Public Domain.

bioRxiv usage has grown significantly over time (Fig. 5). The site currently receives >4 million abstract views per month and ~1.5 million PDF downloads per month. The growth has been consistent and is punctuated by occasional spikes due to articles of particular general interest – for example, a paper by the National Toxicology Program investigating the effects of cell phone radiation on carcinogenesis (Wyde et al., 2016). The numbers for full-text HTML views are currently around half the level of PDF downloads, but full-text HTML was introduced only in 2019 and is unavailable until 24–48 hours after PDF posting, so immediate feeds/alerts will favor the PDF.

Most bioRxiv preprints are ultimately published in traditional journals. Our matching algorithms find that ~70% of bioRxiv preprints are published by journals within two years, a period sufficient for most papers to have passed through review and revision cycles to acceptance. This fraction is consistent with findings for arXiv (Larivière et al. 2013). When a preprint is published in a journal, a prominent link to the publication is inserted above its abstract. Such a link may be absent because the title and/or authorship of the manuscript have changed sufficiently during publication to make it no longer identifiable by matching algorithms or because the paper is still under consideration at a journal. A 70% publication rate is therefore probably an underestimate.

Table 2 – Top ten preprinting institutions

Institution	Preprints
Stanford University	747
University of Oxford	589
University of Cambridge	554
University of Washington	419
University of Pennsylvania	380
University College London	376
University of Edinburgh	363
University of Michigan	360
Harvard University	355
Imperial College London	337

Articles that first appeared as preprints on bioRxiv have now been published in more than 2000 journals. Supplementary Table 1 shows the number of preprints for the 20-most-common destination journals at the time of writing. Comprehensive, updated numbers are available at api.biorxiv.org. The journals that publish bioRxiv preprints represent a wide spectrum of specialties and are both open access and subscription-based. Unsurprisingly, the mega-journals *PLOS ONE* and *Scientific Reports* are highly represented. Journals such as *eLife* that participate in both B2J and J2B also receive significant numbers of papers. Journals that cover subdisciplines highly represented in bioRxiv are more likely to receive relatively high numbers of papers compared with equivalent titles in less well represented subdisciplines. The interval between the posting of a preprint and its publication in a journal is influenced by

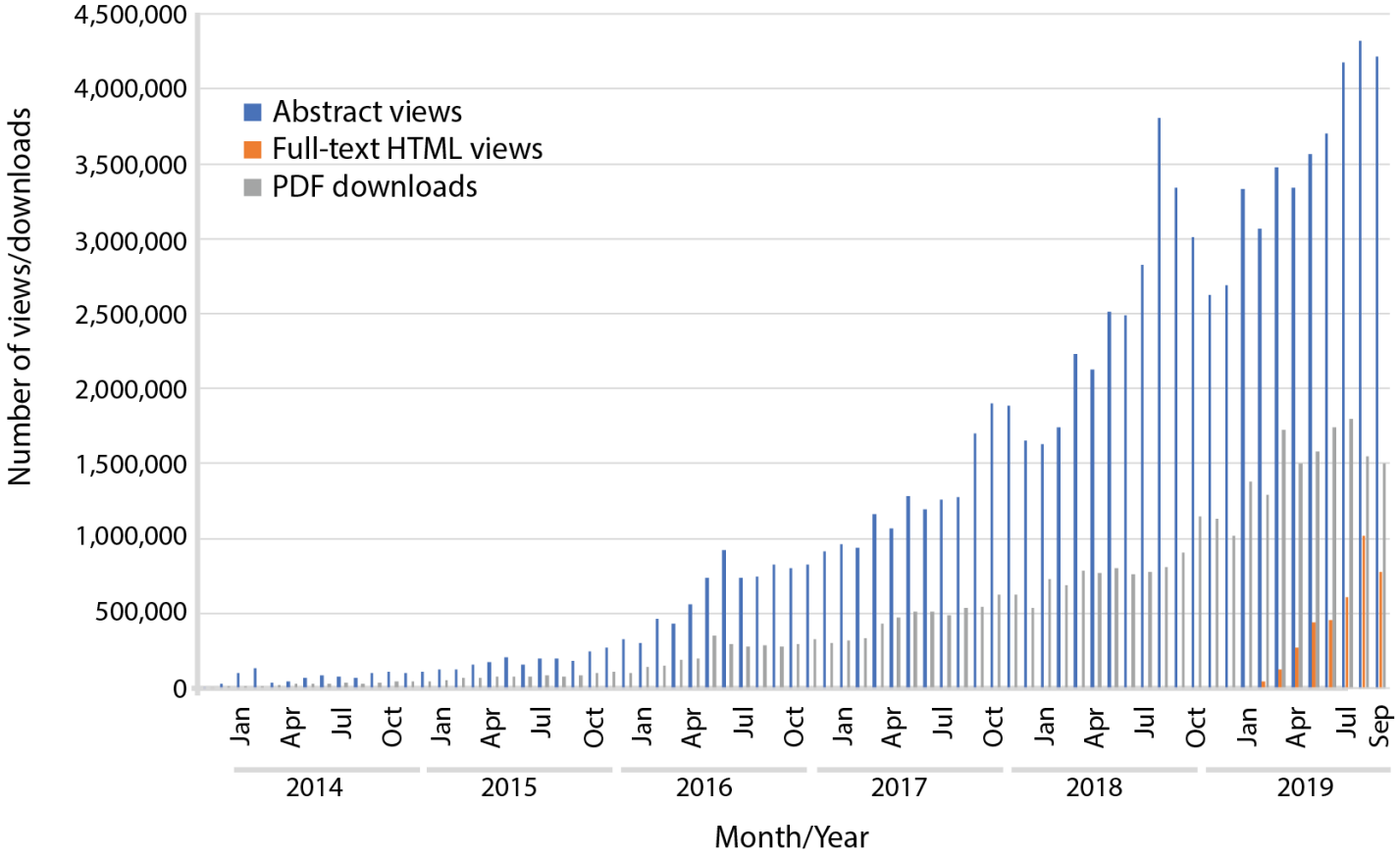

Fig. 5. bioRxiv usage. Growth in abstract views (shown in blue), PDF downloads (shown in grey) and full-text HTML views (shown in red) since the launch of bioRxiv. Note that full-text HTML was only introduced in 2019.

variables such as time to first submission, the number of serial submissions before acceptance, and the extent of revisions required by peer review. For all manuscripts on bioRxiv, the interval between availability on bioRxiv and journal publication currently averages 199 days (median 169 days).

One aspiration for preprints has been that they provide a mechanism for the community to provide feedback on papers to authors. bioRxiv therefore includes an on-site commenting mechanism (powered by Disqus). It also aggregates discussions elsewhere on the Web and in social media. Approximately 5% of papers currently display onsite comments, while just over 1% are covered by discussions on third-party sites such as F1000Prime, PreLights and PubPeer. The latter figure is likely an underestimate as not all independent blogs will be identified. The rate of on-site commenting may appear low; however, these figures are comparable to those for journals. Note also that there are extensive discussions of articles on Twitter (currently more than 30,000 tweets per month) and authors receive private feedback via email (see below); so this may simply reflect the fact that on-site commenting is not yet the preferred medium for feedback. Alternatively, additional cultural change may be required for public commenting to become the norm.

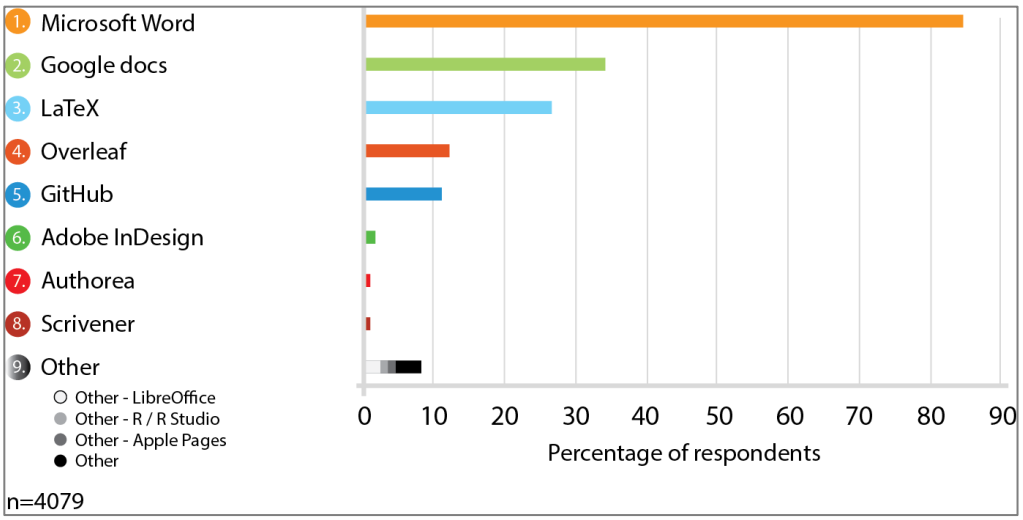

Fig. 6. Authoring software used by authors. In a survey of bioRxiv users, scientists were asked what software they use to prepare manuscripts (see main text and Supplementary Data).

The bioRxiv survey

We recently conducted a survey of more than four thousand bioRxiv users in an effort to understand further how preprints are used among life scientists. There is inevitably some self-selection bias in survey respondents, and the skewed gender (70% male) and geographic representation mean one should be cautious about generalizing from the results (see Supplementary Data). We nevertheless feel the results are informative and highlight some of the key findings below.

bioRxiv uses a submission system in which authors can submit Microsoft Word documents and individual figure files and/or PDF files. This was based on the assumption that most authors in life sciences use Word to compose documents and contrasts with the submission process at arXiv, which focuses on LaTeX users. Figure 6 shows that 85% of bioRxiv survey respondents indeed use Word (Fig. 6). A significant minority use LaTeX (27%) and it is important to emphasize that LaTeX users can submit to bioRxiv; they need simply create a PDF version of their paper as well. Since there is also increasing interest in electronic lab notebooks (ELNs) and potential connections with authoring tools, we also surveyed users on their use of ELNs and related software. The majority (67%) do not use ELNs currently, but this may change and is an area that needs to be monitored as more and more researchers reconsider their experimental workflows. The survey revealed that a variety of reference managers are used, including EndNote (41%), Mendeley (28%), Zotero (14%), Papers (9%) and various others (see Supplementary Table 2), as expected.

There is much discussion among scientists, publishers and IT professionals about authoring technologies and a desire is often voiced for the development and adoption of new tools. The survey findings remind us that it is important to cater for the tools that people are already using, as well as new approaches, particularly when trying to incentivize adoption of new cultural practices such as preprint posting.

We also surveyed authors on their motivation for posting preprints and the consequences of posting. The survey revealed a variety of motivations for posting (Fig. 7), including increasing awareness of research (80%), controlling when research is available (55%), staking a priority claim (54%), a desire to get feedback (53%), and a wish to cite work in a grant application (42%). Most respondents (69%) also felt that immediate sharing of new results benefits the scientific enterprise.

Given the contrast between the anticipated desire for feedback and the relatively low volume of on-site commenting on bioRxiv, we were keen to learn more about the feedback authors received via different channels (Fig. 8A). Importantly, 37% of authors said they had received feedback on preprints by email and 34% through in-person conversations, neither of which bioRxiv can quantify directly. A further 44% had received feedback via Twitter and 14% had received feedback via bioRxiv's online commenting section, figures that indicate some sampling bias among survey respondents given the 5% figure for commenting noted above. Nevertheless, since 55% of surveyed authors express a strong desire for feedback via online comments, that desire is only partly being satisfied (Fig. 8B). Perhaps this is because the technological solutions available are not ideal, but a more likely cause is the absence of meaningful rewards for commenting and providing online feedback within a community already pressed for time. Indeed, 49% of survey respondents had never provided feedback on a preprint. Encouragingly, 54% of surveyed users are discussing preprints at journal clubs (60% of ECRs), which could provide a valuable source of feedback for authors and the community as a whole. However, since the overwhelming

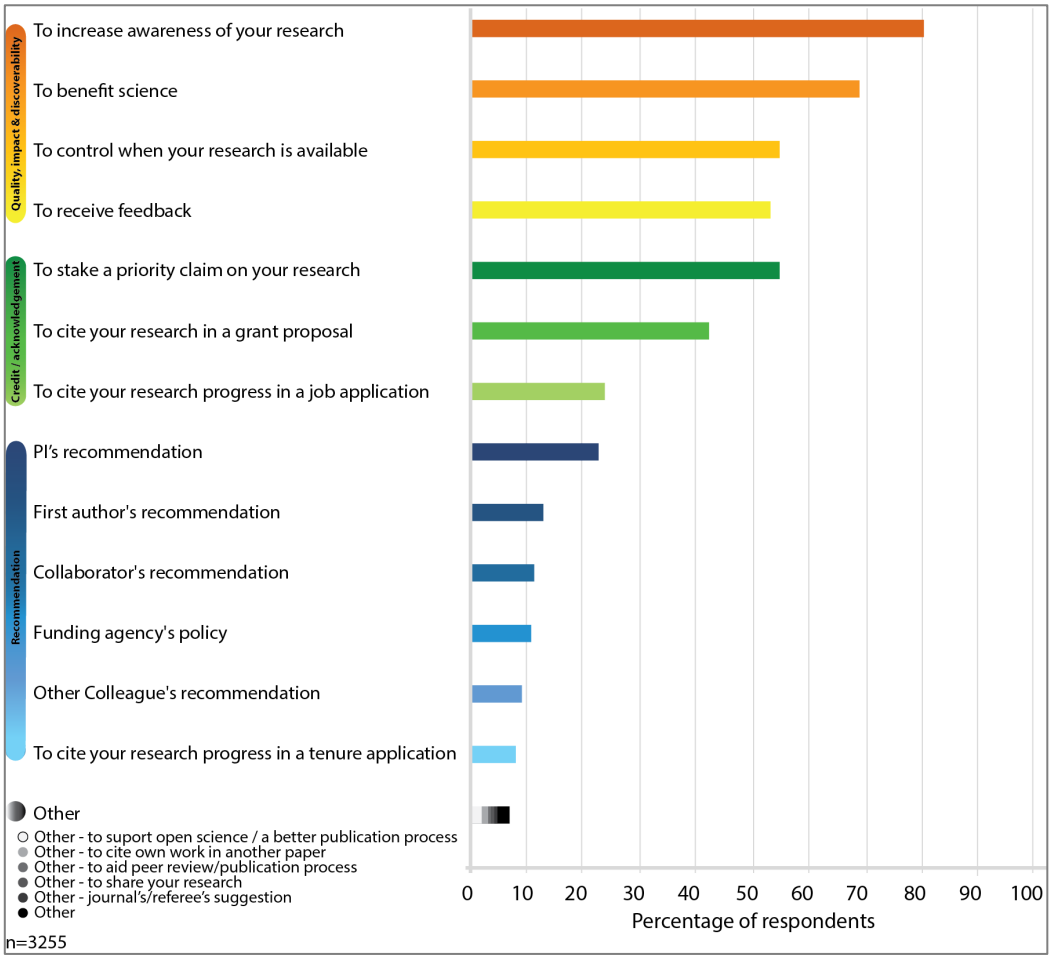

Fig. 7. Motivations for posting work on bioRxiv. In a survey of bioRxiv users, scientists were asked why they post manuscripts on the server (see main text and Supplementary Data).

majority of survey respondents indicated a wish to receive feedback via email (Fig. 8B), there may be a balance to be struck between private and public channels.

Since motivations for preprint posting include both the desire to get work out early and the hope of receiving feedback, we asked when authors post preprints in the course of preparing submissions to journals. 42% of authors said they post before they submit to their first-choice journal; 37% of authors said they post a preprint at the time they submit to their first-choice journal (see Supplementary Table 3). This may indicate there are two main cohorts with slightly different motivations. The ratio between them may change as preprint posting becomes more widely adopted.

Survey respondents reported that posting a preprint had helped in a variety of ways (Fig. 9). 74% said that it had increased awareness of their research. Others found that it had helped them meet new people in their field (19%) and/or make progress in a new field (15%). A smaller number said that it had helped them get a job, grant or seminar invitation (7%, 5%, and 8%, respectively). 28% believe it helped them stake a priority claim, a major motivation for posting in the physical sciences (Ginsparg, 2011). The vast majority of authors (90%) had experienced no negative consequences of preprint posting (Supplementary Table 4). Only 0.7% believed that it had prevented them publishing in a specific journal by giving a competing group an advantage. 6% felt it had limited their choice of journal, however, presumably because a small number of journals will not consider manuscripts previously posted on a preprint server. Given the significant shift in policies among journals over the past few years (see below), we expect this number to fall further in the future.

Discussion

bioRxiv has grown hugely in popularity since its launch in 2013, reflecting an increasing desire within the life science community for rapid and open dissemination of results. There is a positive feedback loop operating, with greater usage and increased familiarity with bioRxiv driving further adoption of the practice of preprinting and its spread to new subdisciplines. The growth of bioRxiv has also helped prompt the launch of numerous similar servers in other fields (e.g. chemRxiv, SocArXiv, PsyArXiv and EarthArXiv) and inspired the creation of medRxiv.

A number of other factors have contributed to preprint adoption. These include changes in many publishers' policies allowing their journals to consider papers previously posted to preprint servers.[1] Furthermore, journals such as the Public Library of Science (PLOS) titles and *eLife*

[1] Compare the current Wikipedia page listing academic journal policies (https://en.wikipedia.org/wiki/List_of_academic_journals_by_preprint_policy) with earlier versions of this page (e.g. https://web.archive.org/web/20130604021231/https://en.wikipedia.org/wiki/List_of_academic_journals_by_preprint_policy)

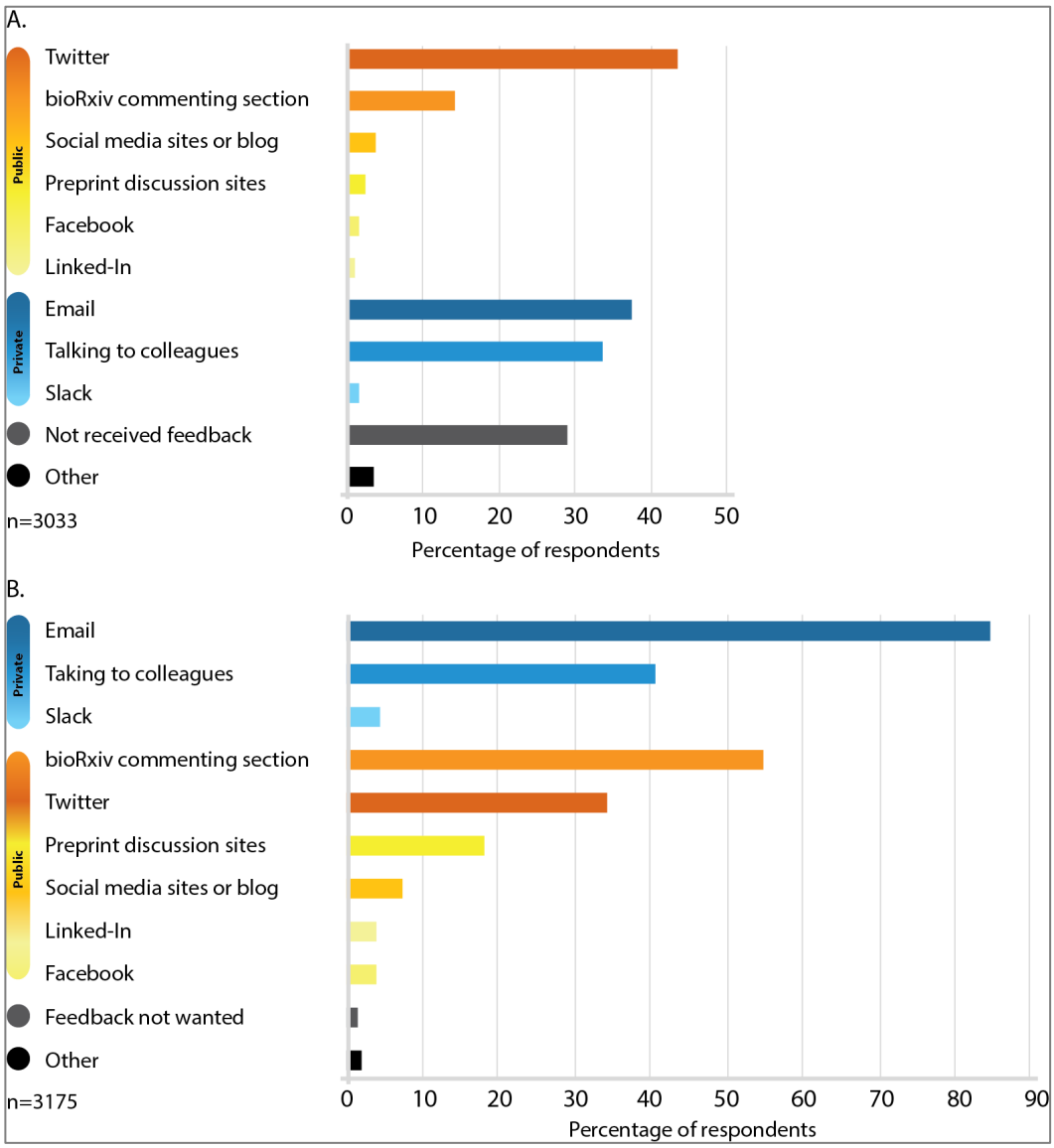

Fig. 8. Feedback mechanisms. A. In a survey of bioRxiv users, scientists were asked the mechanisms by which they have received feedback on papers posted on bioRxiv. B. Authors were asked how they would like to receive feedback on papers (see main text and Supplementary Data).

now actively encourage preprint posting by authors (PLOS, 2018). Similarly, many funders now allow or encourage inclusion of preprints in grant applications and even mandate it in some cases (CZI Science, 2017; Pells, 2018). The NIH recognizes preprints as interim research products (NIH, 2017; NIH, 2019), and some institutions actively recommend that job candidates mention preprints in their applications (ASAPbio, 2019).

It is important to stress, however, the extent to which the research community itself has been the driver of preprint adoption. Genetics and genomics researchers were particularly early adopters and vocal advocates of bioRxiv, and awareness of bioRxiv spread fast among the bioinformatics and evolutionary biology communities. More formal initiatives such as ASAPbio followed and helped spread the word within other subdisciplines, in particular cell and developmental biology. The establishment of preprint discussion sites such as preLights (Brown and Pourquié, 2018) and others has also contributed. bioRxiv has benefited enormously from the enthusiasm with which individual scientists in research communities worldwide have embraced preprints and become active advocates for this approach to dissemination. Twitter has played a very important part in spreading preprint awareness among scientists, alerting readers to individual articles, and providing a conduit for automated article feeds.

It will be interesting to see how greater adoption of preprints further stimulates the evolution of scientific communication and peer review in particular. Anecdotal evidence indicated from the beginning that journal editors were soliciting papers from authors who post preprints on bioRxiv and journals such as *PLOS Genetics* and *Proceedings of the Royal Society B* have editors specifically tasked with such recruitment. B2J is also making the process of journal submission easier for bioRxiv authors. The additional scrutiny of papers prior to journal submission/publication has the potential to improve the quality of papers and optimize peer review. As dissemination and evaluation become decoupled, the pressure to evaluate quickly may be relieved, reducing errors and allowing more thorough and potentially tailored peer review. The very existence of preprints is promoting experimentation with the peer review process at journals (Brainard, 2019) and elsewhere. This is particularly timely given ongoing discussions about the potential for more open and/or transparent peer-review processes (ASAPbio, 2018), additional trust signals (Hall Jamieson et al., 2019), and portable peer review (EMBO and ASAPbio, 2019). Going forward bioRxiv will seek to facilitate such new initiatives,

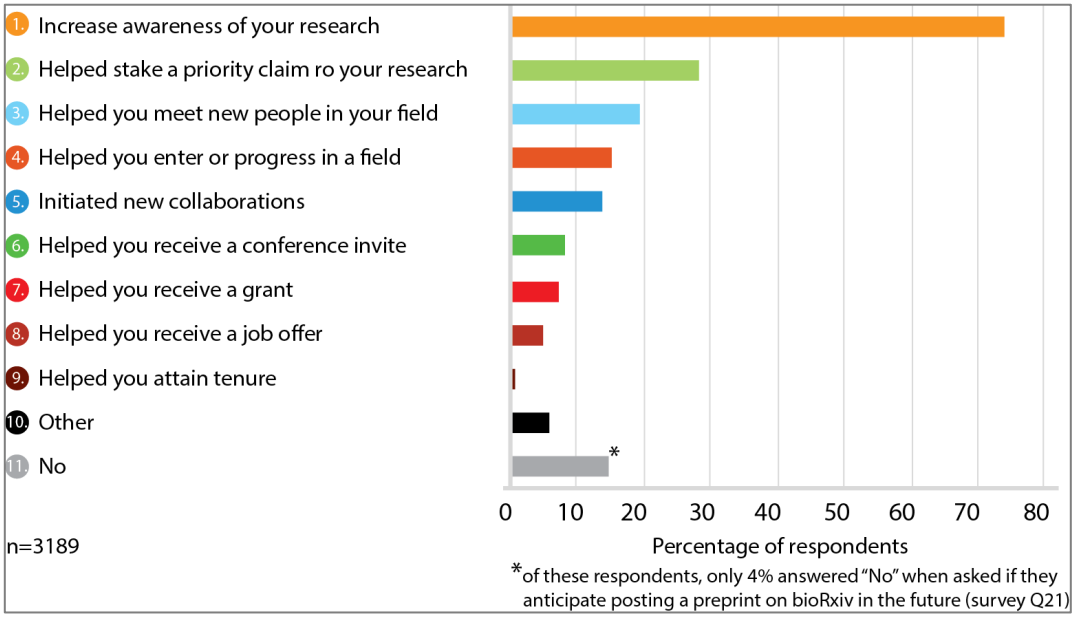

Fig. 9. How preprints help authors. In a survey of bioRxiv users, scientists were asked how posting an article on bioRxiv has helped them in their careers.

as it has journal transfers and linking, community discussion, and reproducibility efforts.

bioRxiv also intends to take advantage of advances in technology and changes in tools used by life scientists. While plagiarism checks are already largely automated, scientific screening is currently performed by individuals. It is unlikely that human judgment could be entirely replaced, but AI approaches offer the hope of automated processes that augment and facilitate human screening. The submission process includes automated aspects of file processing such as PDF generation and verification but still requires manual data entry for other aspects. Improvements in automated text extraction and tagging could make this more efficient, as could a new generation of authoring tools that allow easier generation of XML/HTML. The format of scientific articles has changed little over the years — in many respects it remains tied to a layout dictated by the requirements of print journals. However, the variety of file types employed for different data types, use of tools such as Jupyter notebooks, and broader recognition of code as an integral part of scientific methods and results mean that the content encompassed by the term "research paper" will change, and so too will the outputs with the increasingly anachronistic description 'preprint'.

Concluding remarks

Physicists, computational scientists and mathematicians have been sharing research papers prior to peer review and formal publication for almost three decades. bioRxiv has made this practice widespread in the life sciences and inspired preprint servers in many other disciplines. The decoupling of dissemination and evaluation combined with rapid online posting accelerates awareness of new work and so can increase the pace of research itself. Preprints provide a route to the long-desired goal of making research information freely and immediately available to anyone (Sever et al., 2019). They also create opportunities for evolution of the publishing ecosystem. Broad adoption of preprints, together with technological advances, has the potential to create a more open, equitable and efficient system for the distribution, assessment and archiving of scholarly information.

Acknowledgments

We would like to thank all of those who have worked on and advocated on behalf of bioRxiv, in particular CSHL colleagues Inez Sialiano, Mary Mulligan, Tara Kulesa, Justin Kinney, Bruce Stillman, Terri Grodzicker, Hillary Sussman, Laura DeMare, Dorothy Oddo, Kathy Bubbeo, Denise Weiss, Robert Redmond, Katherine Kelly, and Carol Brown; bioRxiv screeners Andy Tay, Judy Cuddihy, Kaaren Janssen, Heather Cerne, Anqi Zhang, Michael Zierler and Martin Winer; and all the biorxiv Affiliates and Advisory Board members (see biorxiv.org/about-biorxiv). Thanks also to Robert Lourie, Jeremy Freeman, Cori Bargmann, Dario Tarborelli, Paul Ginsparg, Oya Rieger, Gail Steinhart, Jim Entwood, John Sack, Anurag Acharya, Fiona Watt, Leslie Vosshall, Graham Coop, Daniel MacArthur, Leonid Kruglyak, Jessica Polka, Joseph Pickerel, Yaniv Erlich, Steve Shea, Jessica Tollkuhn, Richard Murray, Chris Gunter, Casey Greene, Michael Hoffman, Jim Woodgett, Michael Eisen, Veronique Kiermer, Allison Mudditt, Louise Page, Thomas Lemberger, Bernd Pulverer, Tracey DePellegrin, Eric Topol and Katherine Brown for advice and support. We also wish to acknowledge the support of all the journals that participate in the B2J and J2B programs (see Supplementary Table 5).

This document was created using an adapted Word preprint template developed by the Finkelstein lab (Finkelstein, 2018).

Data availability

Data underlying the results presented here are available at api.biorxiv.org. The full data set (minus any identifying information) for the bioRxiv author survey is provided in the Supplementary Data.

Funding

bioRxiv is a non-profit initiative. It was initially supported by funding from CSHL and generous donations from Robert Lourie. Since 2017, it has been sustained by grant funding from the Chan Zuckerberg Initiative (CZI) and continued support from CSHL.

Ethics statement

The research in this study was reviewed and approved by the Cold Spring Harbor Laboratory IRB (1218750-1) and deemed exempt under 45 CFR 46.101 (b) 2.

Competing interests

RS is Co-Founder of bioRxiv and medRxiv, and employed as Assistant Director of CSHL Press by CSHL. SH is Content Lead for bioRxiv at CSHL, Co-Founder

of PREreview, an ASAPbio Ambassador, and Associate for the eLife ECR Ambassadors program. TR is Lead Developer for bioRxiv and medRxiv and employed by CSHL. KJB is Product Lead for bioRxiv at CSHL. LS is Director of Publication Services at CSHL Press and employed by CSHL. JA is Screening Lead for bioRxiv and employed by CSHL. WM is Director of Product Development and Marketing at CSHL Press and employed by CSHL. JI is Co-Founder of bioRxiv and medRxiv, employed by CSHL as Executive Director of CSHL Press, and an MIT Press advisory board member. JI and RS are members of the Board of Managers of Science Alliance LLC, a limited liability non-stock corporation jointly owned by Cold Spring Harbor Laboratory, EMBO Press Innovations gGmbH, and The Rockefeller University.

References

Abdil, R. J. and Blekhman (2019) Rxivist.org: Sorting biology preprints using social media and readership metrics *PLOS Biol.* 1: e3000269.

ASAPbio (2018) https://asapbio.org/letter

ASAPbio (2019) https://asapbio.org/university-policies

Bloom, T. (2019) New preprint server for medical research *BMJ* 365, l2301

Brainard, J. (2019) In bid to boost transparency, bioRxiv begins posting peer reviews next to preprints *Science* https://www.sciencemag.org/news/2019/10/bid-boost-transparency-biorxiv-begins-posting-peer-reviews-next-preprints

Brown, K. and Pourquié, O. (2018) Introducing preLights: preprint highlights, selected by the biological community *Development* 145: dev164186

Cobb, M. (2017) The prehistory of biology preprints: A forgotten experiment from the 1960s *PLoS Biol* 15, e2003995.

CZI Science Initiative Privacy Principles (2017) https://chanzuckerberg.com/privacy/science-privacy-principles/

EMBO and ASAPbio (2019) Review Commons https://www.reviewcommons.org/

Finkelstein, I. et al. (2018) BioRxiv-template https://github.com/finkelsteinlab/BioRxiv-Template

Ginsparg, P. (2011) It was twenty years ago today arXiv:1108.2700

Hall Jamieson, K., McNutt, M., Kiermer, V., and Sever, R. (2019) Signaling the trustworthiness of science *Proc Natl Acad Sci USA* 116, 19231-19236

Kaiser, J. (2013) New Preprint Server Aims to Be Biologists' Answer to Physicists' arXiv *Science* https://www.sciencemag.org/news/2013/11/new-preprint-server-aims-be-biologists-answer-physicists-arxiv

Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., and Thelwall, M. (2013) arXiv e-prints and the journal of record: An analysis of roles and relationships arXiv:1306.3261

Marshall, E. (1999) Varmus Circulates Proposal for NIH-Backed Online Venture *Science* 284, 718

Nature Publishing Group (2012) http://precedings.nature.com/

NIH (2017) Reporting Preprints and Other Interim Research Products https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-050.html

NIH (2019) Frequently Asked Questions, Biosketches https://grants.nih.gov/grants/policy/faq_biosketches.htm#4579

Pells, R. (2018) Wellcome mandates publication before peer review in health crises *Time Higher Education* https://www.timeshighereducation.com/news/wellcome-mandates-publication-peer-review-health-crises

PLOS (2018) Power to the Preprint [blog post] https://blogs.plos.org/plos/2018/05/power-to-the-preprint/

Quake, S. (2019) Stanford Big Data Medicine Precision Health https://youtu.be/zt9hlbet2Lk

Rawlinson, C. (2019) New preprint server for medical research *BMJ* 2019, l2301

Rieger, O. Y., Steinhart, G. and Cooper, D. (2018) arXiv@25: Key findings of a user survey arXiv:1607.08212

Royle, S. (2015) Waiting To Happen II: Publication Lag Times [blog post] https://quantixed.org/2015/03/16/waiting-to-happen-ii-publication-lag-times/

Sack, J. (2019) Manuscript Exchange — What MECA Can Do for the Academic Publishing World — And What it Can't [blog post] https://scholarlykitchen.sspnet.org/2018/07/25/guest-post-manuscript-exchange-meca-can-academic-publishing-world-cant/

Sarabipour, S., Debat, H. J., Emmott, E., Burgess, S. J., Schwessinger, B., Hensel, Z. (2019) On the value of preprints: An early career researcher perspective. *PLoS Biol* 17, e3000151

Sever, R. (2019) Twitter thread https://twitter.com/cshperspectives/status/1154394654525272064

Sever R., Eisen M. B., and Inglis J. R. (2019) Plan U: Universal access to scientific and medical research via funder preprint mandates. *PLoS Biol* 17, e3000273.

Vale, R. (2015) Accelerating scientific publication in biology *Proc Natl Acad Sci USA* 112, 13439-13446

Wyde, M. et al. (2016) Report of Partial findings from the National Toxicology Program Carcinogenesis Studies of Cell Phone Radiofrequency Radiation in Hsd: Sprague Dawley® SD rats (Whole Body Exposures) bioRxiv doi: https://doi.org/10.1101/055699 version 1 https://www.biorxiv.org/content/10.1101/055699v1

Supplementary Table 1 – Top 20 journal destinations for bioRxiv preprints

Journal	2014 (%)	2015 (%)	2016 (%)	2017 (%)	2018 (%)	2019 (%)
BMC Bioinformatics	0.96	1.22	1.17	0.87	0.80	0.78
Bioinformatics	4.18	4.88	4.15	2.78	2.43	1.87
B*iophysical Journal*	0.96	0.24	0.53	0.64	0.80	1.08
Cell Reports	0	0.85	0.64	0.83	0.97	0.96
eLife	3.86	3.17	3.72	5.26	4.90	4.64
G3: Genes|Genomes|Genetics	0.64	2.44	2.60	1.38	0.85	0.74
Genetics	2.25	4.76	3.46	1.80	0.99	0.74
Genome Biology	3.54	2.8	1.44	1.02	0.85	0.65
mBio	0	0.12	0.58	0.68	0.85	0.70
Molecular Biology and Evolution	2.25	2.32	1.65	1.12	0.57	0.75
Nature Communications	1.61	0.98	2.23	2.52	4.38	3.92
NeuroImage	0	0.85	1.01	1.48	1.36	1.34
Nucleic Acids Research	2.89	2.20	1.70	1.93	1.43	1.21
PLOS Biology	0.32	0.49	0.48	0.57	0.76	0.90
PLOS Computational Biology	2.25	3.29	2.39	2.23	1.84	2.14
PLOS Genetics	3.86	4.02	2.29	1.84	1.48	1.23
PLOS ONE	10.3	6.71	4.73	3.56	4.93	7.93
PNAS	2.25	3.05	2.98	2.78	2.63	2.46
Scientific Reports	0.32	2.44	5.37	6.53	4.54	4.29
The Journal of Neuroscience	0	0	0	1.02	1.16	1.05

Supplementary Table 2 – Reference packages used by survey respondents

Reference manager	% of respondents
EndNote	40.9
Zotero	14.3
Paperpile	7.16
Papers	9.28
F1000Workspace	1.17
Authorea	0.25
Overleaf	3.42
Mendeley	27.7
Other	14.3

N=4009

Supplementary Table 3 – Preprint posting times reported by survey respondents

Posting time	% of respondents
Before submitting to first-choice journal	37.1
When submitting to first-choice journal	41.9
After submitting to first-choice journal, but before receiving reviewers' comments	6.65
After receiving reviewers' comments from first-choice journal	3.08
After manuscript rejected from first-choice journal	4.89
Not sure	4.03
Other	2.37

N=3250

Supplementary Table 4 – Negative consequences of posting reported by survey respondents

Has posting a preprint negatively affected you in any of the following ways	% of respondents
No	89.6
Limited your choice of journal for publication	6.43
Prevented you from publishing in your journal of choice because another lab published before you	0.70
Affected your priority claim to the research	1.25
Other	4.41

N=3127

Supplementary Table 5 – Journal partners

B2J participants
Access Microbiology
Acta Crystallographica Section D
American Journal of Human Genetics
Animal
Antimicrobial Agents and Chemotherapy
Applied and Environmental Microbiology
Archives of Animal Nutrition
Autoimmunity Highlights
Autophagy
Biochemical Journal
BioDesign Research
Biology Open
Biology of Reproduction
Biophysical Journal
Bioscience Reports
BMC Biology
BME Frontiers
BMJ Open Science
Brain Topography
Bulletin of Entomological Research
Bulletin of Mathematical Biology
Cell Cycle
Cell Reports
The Cerebellum
Clinical and Vaccine Immunology
Clinical Science
The CRISPR Journal
Current Biology
Development
Disease Models & Mechanisms
Drug Metabolism and Disposition
eLife
The EMBO Journal
EMBO Molecular Medicine
EMBO Reports
Endocrinology
eNeuro
Epigenetics
Environmental Health Perspectives
Experimental Agriculture
Experimental Brain Research
Experimental Physiology
F1000Research
FASEB BioAdvances
The FASEB Journal
Fly
Frontiers in Aging Neuroscience
Frontiers in Behavioral Neuroscience
Frontiers in Big Data
Frontiers in Bioengineering and Biotechnology
Frontiers in Cell and Developmental Biology
Frontiers in Cellular and Infection Microbiology
Frontiers in Cellular Neuroscience
Frontiers in Computational Neuroscience
Frontiers in Ecology and Evolution
Frontiers in Endocrinology
Frontiers in Genetics
Frontiers in Human Neuroscience
Frontiers in Immunology
Frontiers in Integrative Neuroscience
Frontiers in Marine Science
Frontiers in Microbiology
Frontiers in Molecular Biosciences
Frontiers in Molecular Neuroscience
Frontiers in Neural Circuits
Frontiers in Neuroanatomy
Frontiers in Neuroenergetics
Frontiers in Neuroengineering
Frontiers in Neurorobotics
Frontiers in Neuroinformatics
Frontiers in Neurology
Frontiers in Neuroscience
Frontiers in Pharmacology
Frontiers in Physics
Frontiers in Physiology
Frontiers in Plant Science
Frontiers in Psychiatry
Frontiers in Research Metrics and Analytics
Frontiers in Sustainable Food Systems
Frontiers in Synaptic Neuroscience
Frontiers in Systems Neuroscience
Frontiers in Veterinary Science
G3: Genes, Genomes, Genetics
Genetics
Genetics in Medicine
Genetics Research
Genome Biology
Genome Research
GigaScience
GM Crops and Food
Hormones and Cancer
Infection and Immunity
Invertebrate Neuroscience
iScience
JCI Insight
Journal of Apicultural Research
Journal of Bacteriology
Journal of Biological Chemistry
Journal of Bryology
The Journal of Cell Biology
Journal of Cell Science
The Journal of Clinical Endocrinology & Metabolism
Journal of Clinical Investigation
Journal of Clinical Microbiology
Journal of the Endocrine Society
Journal of Experimental Biology
Journal of Experimental Botany

The Journal of Experimental Medicine
The Journal of General Physiology
Journal of General Virology
Journal of Helminthology
The Journal of Histotechnology
The Journal of Immunology
Journal of Lipid Research
Journal of Mammary Gland Biology and Neoplasia
Journal of Medical Microbiology
Journal of Molecular Evolution
Journal of Neural Transmission
Journal of Neurophysiology
The Journal of Neuroscience
Journal of Pharmacology and Experimental Therapeutics
The Journal of Physiology
Journal of Virology
Learning & Memory
Life Science Alliance
mBio
Medicinal Chemistry Research
Microbial Genomics
Microbiology
Microbiology Resource Announcements
Molecular and Cellular Biology
Molecular & Cellular Proteomics
Molecular Biology of the Cell
Molecular Pharmacology
Molecular Systems Biology
Molecular Vision
mSphere
mSystems
Mycologia
Neuronal Signaling
Oecologia
Open Biology
Pathogens and Immunity
Peerage of Science
PeerJ
Pharmaceutical Research
Photosynthesis Research
The Plant Cell
Plant Cell, Tissue and Organ Culture
Plant Direct
Plant Molecular Biology
Plant Phenomics
Plant Physiology
PLOS Biology
PLOS Computational Biology
PLOS Genetics
PLOS Medicine
PLOS Neglected Tropical Diseases
PLOS One
PLOS Pathogens
PNAS
Research
RNA
RNA Biology
Science
Science Advances
Science Immunology
Science Signaling
Science Translational Medicine
Seed Science Research
Theoretical and Applied Genetics

J2B Participants
Antimicrobial Agents and Chemotherapy
Applied and Environmental Microbiology
Biology Open
Biophysical Journal
BMJ Open Science
Development
Disease Models & Mechanisms
eLife
The EMBO Journal
EMBO Molecular Medicine
EMBO Reports
G3: Genes Genomes Genetics
Genetics
Genome Research
Infection and Immunity
Journal of Bacteriology
Journal of Cell Science
Journal of Clinical Microbiology
Journal of Experimental Biology
Journal of Experimental Botany
Journal of Virology
Life Science Alliance
mBio
Molecular and Cellular Biology
Molecular Systems Biology
mSphere
mSystems
The Plant Cell
Plant Physiology
PLOS Biology
PLOS Computational Biology
PLOS Genetics
PLOS Neglected Tropical Diseases
PLOS One
PLOS Pathogens

Supplementary Material — Survey design, execution and analysis

Prior to the bioRxiv survey, we asked for community input to ensure we asked questions relevant to scientists who use, or are interested in, preprints. We emailed a pre-survey — consisting of five available spaces to input suggested questions —to authors and participants who expressed interest in bioRxiv at various scientific conferences, and we used social media to promote this. We received 517 responses and composed questions covering common themes, supplemented by questions inspired by the arXiv@25 survey (Reiger *et al*., 2016) and additional questions intended to help bioRxiv cater to the needs of users and get input from non-users.

The survey design struck a balance between length/completion time and depth of understanding/utility. The final survey comprised 39 multiple-choice questions and one open-ended question, and used the Survey Monkey tool. Questions were divided across user type; authors, readers, and non-users viewed a maximum of 37, 24, and 16 questions, respectively. The survey took an average of ~8 minutes to complete, with a 91% completion rate.

The same target audience contacted for the pre-survey were also notified of the launch of the final survey. We also alerted bioRxiv readers by adding a banner to all pages at bioRxiv.org. In an attempt to reach non-users, we asked bioRxiv Affiliates to post flyers at their institutions and use institutional email listservs to amplify the message. The questions and answer options were fixed at launch, except for the addition of the "bioRxiv survey flyer/poster on bulletin board" answer to Q1 following our attempt to increase the number of non-user respondents. This option was added after 3209 responses had been received. At the same time, the word "survey" was underlined in Q1 to emphasize that the question referred to the survey, not how respondents heard about bioRxiv, as there appeared to be some confusion from the responses supplied in the "Other (please specify)" free-text field.

For all multiple-choice questions containing a free-text field — for example, "Other (please specify)," — responses were read, and common categories were identified. Categories with ~1% or more of the total responses were included as a sub-category in the "Other" response totals. Each multiple-choice answer was tallied and expressed as a percentage of the total responses for that question. Graphs were generated in Microsoft Excel (Version 16.16.14) and modified using Adobe Illustrator CS6 (Version 16.0.4). To avoid survey responses being used to identify individuals, the answers to free-text questions were removed prior to uploading as Supplementary data.

Supplementary Data — Survey Results

See csv file posted online.