# Tracking the popularity and outcomes of all bioRxiv preprints

Richard J. Abdill[1] and Ran Blekhman[1,2]

1 – Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN

2 – Department of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, MN

**ORCID iDs**

RJA: https://orcid.org/0000-0001-9565-5832

RB: https://orcid.org/0000-0003-3218-613X

**Correspondence**

Ran Blekhman, PhD

University of Minnesota

MCB 6-126

420 Washington Avenue SE

Minneapolis, MN 55455

Email: blekhman@umn.edu

## Abstract

Researchers in the life sciences are posting their work to preprint servers at an unprecedented and increasing rate, sharing papers online before (or instead of) publication in peer-reviewed journals. Though the popularity and practical benefits of preprints are driving policy changes at journals and funding organizations, there is little bibliometric data available to measure trends in their usage. Here, we collected and analyzed data on all 37,648 preprints that were uploaded to bioRxiv.org, the largest biology-focused preprint server, in its first five years. We find that preprints on bioRxiv are being read more than ever before (1.1 million downloads in October 2018 alone) and that the rate of preprints being posted has increased to a recent high of more than 2,100 per month. We also find that two-thirds of bioRxiv preprints posted in 2016 or earlier were later published in peer-reviewed journals, and that the majority of published preprints appeared in a journal less than six months after being posted. We evaluate which journals have published the most preprints, and find that preprints with more downloads are likely to be published in journals with a higher impact factor. Lastly, we developed Rxivist.org, a website for downloading and interacting programmatically with indexed metadata on bioRxiv preprints.


## Introduction

In the 30 days of September 2018, *The Journal of Biochemistry* published eight full-length research articles. *PLOS Biology* published 19. *Genetics* published 23. *Cell* published 35. BioRxiv had posted more articles than all four—combined—by the end of September 3 (**Table S1**).

BioRxiv (pronounced "Bio Archive") is a preprint server, a repository to which researchers can post their papers directly to bypass the months-long turnaround time of the publishing process and share their findings with the community more quickly (Berg et al. 2016). Though the idea of preprints is far from new (Cobb 2017), researchers have become vocally frustrated about the lengthy process of distributing research through the conventional pipelines (Powell 2016), and numerous public laments have been published decrying increasingly impractical demands of journals and reviewers (e.g. Raff et al. 2008; Snyder 2013). One analysis found that review times at journals published by the Public Library of Science (PLOS) have doubled over the last decade (Hartgerink 2015); another found a two- to four-fold increase in the amount of data required for publication in top journals between 1984 and 2014 (Vale 2015). Other studies have found more complicated dynamics at play from both authors and publishers that can affect time to press (Powell 2016; Royle 2014).

Against this backdrop, preprints have become a steady source of the most recent research in biology, providing a valuable way to learn about exciting, relevant and high-impact findings—for free—months or years before that research will appear anywhere else, if at all (Kaiser 2017). It's a practice long familiar to physicists, who began submitting preprints to arXiv, one of the earliest preprint servers, in 1991 (Verma 2017). Researchers in fields supported by that server "have developed a habit of checking arXiv every morning to learn about the latest work in their field" (Vale and Hyman 2016), and one survey of published mathematicians found that 81 percent had posted at least one preprint to the site (Fowler 2011). In the life sciences, however, researchers approached preprints with reluctance (O'Roak 2018), even when major publishers made it clear they were not opposed to the practice ("Nature respects preprint servers" 2005; Desjardins-Proulx et al. 2013). An early NIH plan for PubMed Central called "E-Biomed" included the hosting of preprints (Varmus 1999; Smaglik 1999) but was scuttled by the National Academy of Sciences, which successfully negotiated the exclusion of work that had not been peer-reviewed (Marshall 1999; Kling et al. 2003).

Further attempts to circulate biology preprints, such as NetPrints (Delamothe et al. 1999), Nature Precedings (Kaiser 2017), and The Lancet Electronic Research Archive (McConnell and Horton 1999), popped up (and then folded) over time ("ERA Home" 2019). The one that would catch on, bioRxiv, wasn't founded until 14 years after the fall of E-Biomed (Callaway 2013). Now, biology publishers are actively trawling preprint servers for submissions (Barsh et al. 2016; Vence 2017), and more than 100 journals accept submissions directly from the bioRxiv website ("Submission Guide" 2018). The National Institutes of Health announced the explicit acceptance of preprint citations in grant proposals ("Reporting Preprints and Other Interim Research Products" 2017), and multiple funding opportunities from the multi-billion-dollar Chan Zuckerberg Initiative (Abutaleb 2015) require all publications to first be posted to a preprint server ("Funding Opportunities" 2018; Champieux 2018). The conventions of the biology publishing game are changing, in ways that reflect a strong influence from the expanding popularity of preprints. However, details about that ecosystem are hard to come by. We know bioRxiv is the largest of the biology-focused preprint servers: Of the eight websites indexed by PrePubMed (http://www.prepubmed.org), bioRxiv now consistently posts more than three times as many articles per month as the other seven combined (Anaya 2018). Sporadic updates from bioRxiv leaders show a chain of record-breaking months for submission numbers (Sever 2018), and analyses have examined metrics such as total downloads (Serghiou and Ioannidis 2018) and publication rate (Schloss 2017). But long-term questions remain open: Which fields have posted the most preprints, and which collections are growing most quickly? How many times have preprints been downloaded, and which categories are most popular with readers? How many preprints are eventually published elsewhere, and in what journals? Is there a relationship between a preprint's popularity and the journal in which it later appears? Do these conclusions change over time?

Here, we aim to answer these questions by collecting metadata about all 37,648 preprints posted to bioRxiv through November 2018. We use these data to measure the growing popularity of bioRxiv as a research repository and to help quantify trends in biology preprints that have until now been out of reach. In addition, we developed Rxivist (pronounced "Archivist"), a website, API and database (available at https://rxivist.org and gopher://origin.rxivist.org) that provide a fully featured system for interacting programmatically with the periodically indexed metadata of all preprints posted to bioRxiv.

# Results

We developed a Python-based web crawler to visit every content page on the bioRxiv website and download basic data about each preprint across the site's 27 subject-specific categories: title, authors, download statistics, submission date, category, DOI, and abstract. The bioRxiv website also provides the email address and institutional affiliation of each author, plus, if the preprint has been published, its new

DOI and the journal in which it appeared. For those preprints, we also used information from Crossref to determine the date of publication. We have stored these data in a PostgreSQL database; snapshots of the database are available for download, and users can access data for individual preprints and authors on the Rxivist website and API. Additionally, a repository is available online at https://doi.org/10.5281/zenodo.2465689 that includes the database snapshot used for this manuscript, plus the data files used to create all figures. Code to regenerate all figures in this paper is included there and on GitHub (https://github.com/blekhmanlab/rxivist/blob/master/paper/figures.md). See Methods and Supplementary Information for a complete description.

## Preprint submissions

The most apparent trend that can be pulled from the bioRxiv data is that the website is extraordinarily popular with authors, and becoming more so every day: There were 37,648 preprints available on bioRxiv at the end of November 2018, and more preprints were posted in the first 11 months of 2018 (18,825) than in all four previous years combined (**Figure 1a**). The number of bioRxiv preprints doubled in less than a year, and new submissions have been trending upward for five years (**Figure 1b**). The plurality of site-wide growth can be attributed to the neuroscience collection, which has had more submissions than any bioRxiv category in every month since September 2016 (**Figure 1b**). In October 2018, it became the first of bioRxiv's collections to contain 6,000 preprints (**Figure 1a**). The second-largest category is bioinformatics (4,249 preprints), followed by evolutionary biology (2,934). October 2018 was also the first month in which bioRxiv posted more than 2,000 preprints, increasing its total preprint count by 6.3 percent (2,119) in 31 days.

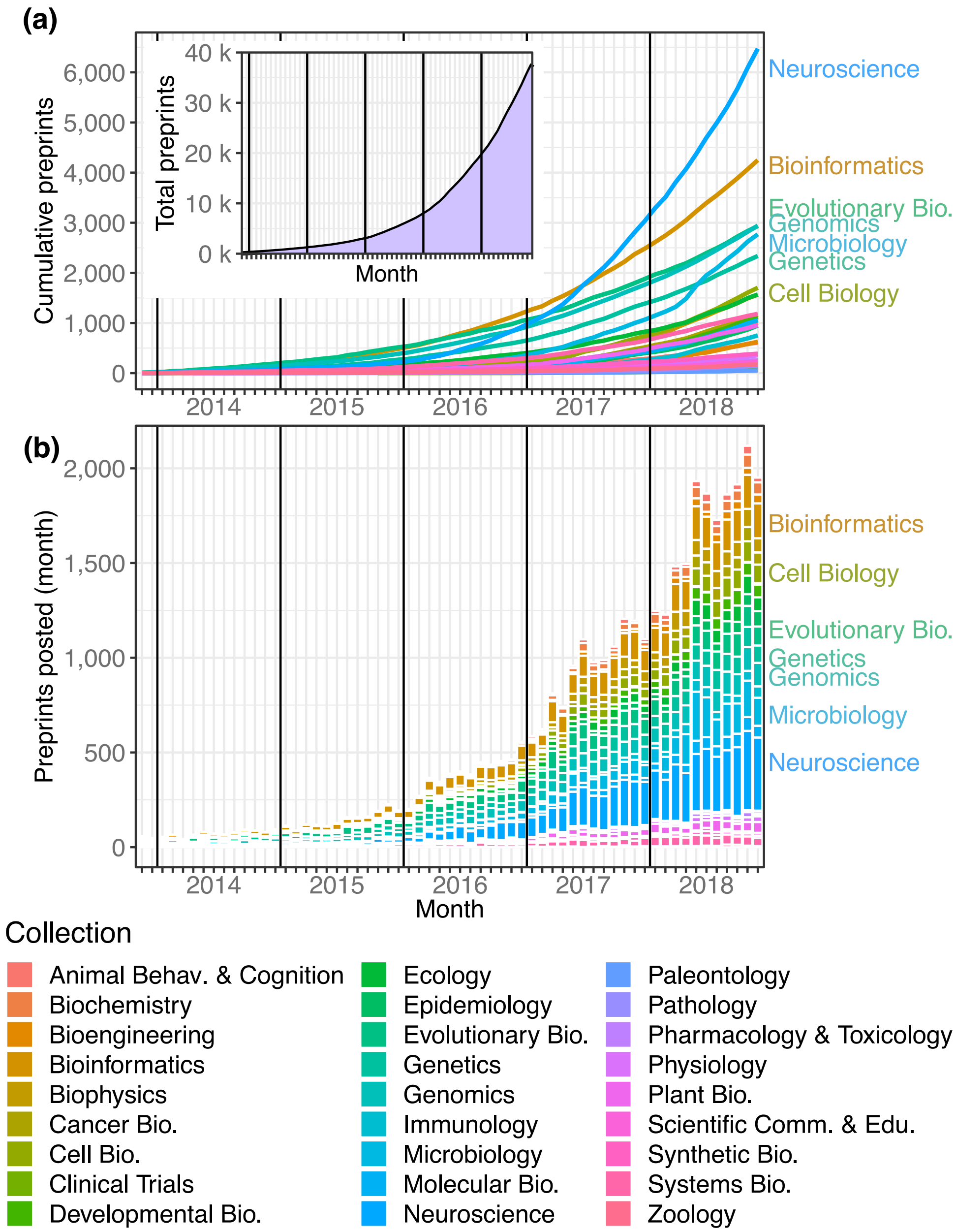



**Figure 1.** Total preprints posted to bioRxiv over a 61-month period from November 2013 through November 2018. **(a)** The number of preprints (y-axis) at each month (x-axis), with each category depicted as a line in a

different color. **(a, inset)** The overall number of preprints on bioRxiv in each month. **(b)** The number of preprints posted (y-axis) in each month (x-axis) by category. The category color key is provided below the figure.

**Supplementary files:** *submissions_per_month.csv*, *submissions_per_month_overall.csv*

## Preprint downloads

Considering the number of downloads for each preprint, we find that bioRxiv's popularity with readers is also increasing rapidly (**Figure 2**): The total download count in October 2018 (1,140,296) was an 82 percent increase over October 2017, which itself was a 115 percent increase over October 2016 (**Figure 2a**). bioRxiv preprints were downloaded almost 9.3 million times in the first 11 months of 2018, and in October and November 2018, bioRxiv recorded more downloads (2,248,652) than in the website's first two and a half years (**Figure 2b**). The overall median downloads per paper is 279 (**Figure 2b, inset**), and the genomics category has the highest median downloads per paper, with 496 (**Figure 2c**). The neuroscience category has the most downloads overall—it overtook bioinformatics in that metric in October 2018, after bioinformatics spent nearly 4 and a half years as the most downloaded category (**Figure 2d**). In total, bioRxiv preprints were downloaded 19,699,115 times from November 2013 through November 2018, and the neuroscience category's 3,184,456 total downloads accounts for 16.2 percent of these (**Figure 2d**). However, this is driven mostly by that category's high volume of preprints: The median downloads per paper in the neuroscience category is 269.5, while the median of preprints in all other categories is 281 (**Figure 2c**).

We also examined traffic numbers for individual preprints relative to the date that they were posted to bioRxiv, which helped create a picture of the change in a preprint's downloads by month after it is posted (**Figure S1**): We can see that preprints typically have the most downloads in their first month, and the download count per month decays most quickly over a preprint's first year on the site. The most downloads recorded in a preprint's first month is 96,047, but the median number of downloads a preprint receives in its debut month on bioRxiv is 73. The median downloads in a preprint's second month falls to 46, and the third month median falls again, to 27. Even so, the average preprint at the end of its first year online is still being downloaded about 12 times per month, and some papers don't have a "big" month until relatively late, receiving the majority of their downloads in their sixth month or later (**Figure S2**).

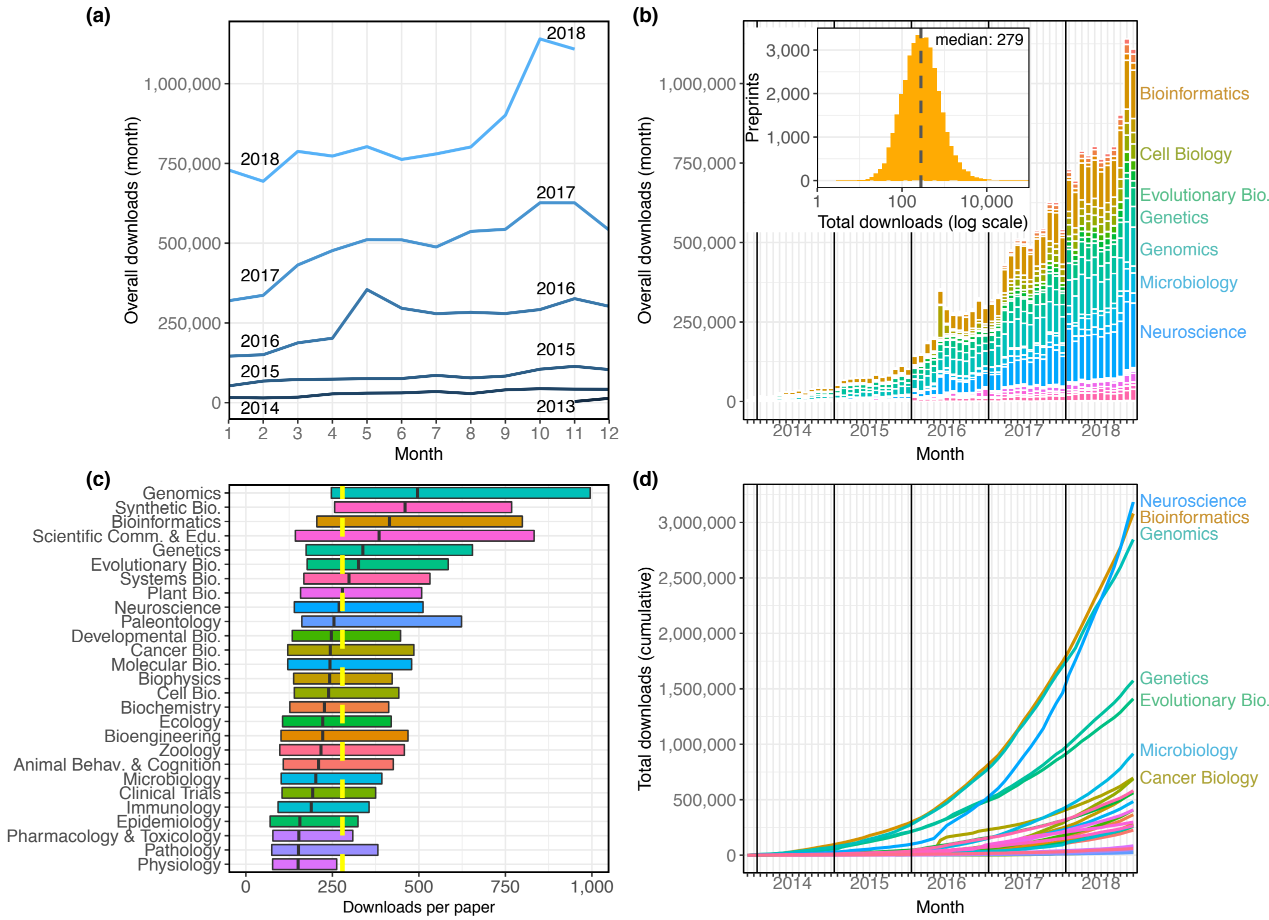



**Figure 2.** The distribution of all recorded downloads of bioRxiv preprints. **(a)** The downloads recorded in each month, with each line representing a different year. The lines reflect the same totals as the height of the bars in Figure 2b. **(b)** A stacked bar plot of the downloads in each month: The height of each bar indicates the total downloads in that month. Each stacked bar shows the number of downloads in that month attributable to each category; the colors of the bars are described in the legend in Figure 1. **(b, inset)** A histogram showing the site-wide distribution of downloads per preprint, as of the end of November 2018. The median download count for a single preprint is 279, marked by a dashed line. **(c)** The

distribution of downloads per preprint, broken down by category. Each box illustrates that category's first quartile, median, and third quartile (similar to a boxplot, but whiskers are omitted due to a long right tail in the distribution). The vertical dashed yellow line indicates the overall median downloads for all preprints. **(d)** Cumulative downloads over time of all preprints in each category. The top seven categories at the end of the plot (November 2018) are labeled using the same category color-coding as above.

**Supplementary files:** *downloads_per_category.csv*, *downloads_per_month_cumulative.csv*, *downloads_per_month_per_year.csv*

## Preprint authors

While data about the authors of individual preprints is easy to organize, associating authors between preprints is difficult due to a lack of consistent unique identifiers (see Methods). We chose to define an author as a unique name in the author list, including middle initials but disregarding letter case and punctuation. Keeping this in mind, we find that there are 170,287 individual authors with content on bioRxiv. Of these, 106,231 (62.4%) posted a preprint in 2018, including 84,339 who posted a preprint for the first time (**Table 1**), indicating that total authors increased by more than 98 percent in 2018.

| Year | New authors | Total authors |
|---|---|---|
| 2013 | 608 | 608 |
| 2014 | 3,873 | 4,012 |
| 2015 | 7,584 | 8,411 |
| 2016 | 21,832 | 24,699 |
| 2017 | 52,051 | 61,239 |
| 2018 | 84,339 | 106,231 |

**Table 1.** Unique authors posting preprints in each year. "New authors" counts authors posting preprints in that year that had never posted before; "Total authors" includes researchers who may have already been counted in a previous year, but are also listed as an author on a preprint posted in that year. Data for table pulled directly from database. An SQL query to generate these numbers is provided in the Methods section.

Even though 129,419 authors (76.0%) are associated with only a single preprint, the mean preprints per author is 1.52 because of a skewed rate of contributions also found in conventional publishing (Rørstad and Aksnes 2015): 10 percent of authors account for 72.8 percent of all preprints, and the most prolific researcher on bioRxiv, George Davey Smith, is listed on 97 preprints across seven categories (**Table S2**). 1,473 authors list their most recent affiliation as Stanford University, the most represented institution on bioRxiv (**Table S3**). Though the majority of the top 100 universities (by author count) are based in the United States, five of the top 11 are from

Great Britain. These results rely on data provided by authors, however, and is confounded by varying levels of specificity: While 530 authors report their affiliation as "Harvard University," for example, there are 528 different institutions that include the phrase "Harvard," and the four preprints from the "Wyss Institute for Biologically Inspired Engineering at Harvard University" don't count toward the "Harvard University" total.

## Publication outcomes

In addition to monthly download statistics, bioRxiv also records whether a preprint has been published elsewhere, and in what journal. In total, 15,797 bioRxiv preprints have been published, or 42.0 percent of all preprints on the site (**Figure 3a**). Proportionally, evolutionary biology preprints have the highest publication rate of the bioRxiv categories: 51.5 percent of all bioRxiv evolutionary biology preprints have been published in a journal (**Figure 3b**). Examining the raw number of publications per category, neuroscience again comes out on top, with 2,608 preprints in that category published elsewhere (**Figure 3c**). When comparing the publication rates of preprints posted in each month we see that more recent preprints are published at a rate close to zero, followed by an increase in the rate of publication every month for about 12–18 months (**Figure 3a**). A similar dynamic was observed in a study of preprints posted to arXiv: After recording lower rates in the most recent time periods, Larivière et al. (2014) found publication rates of arXiv preprints leveled out at about 73 percent. Of bioRxiv preprints posted between 2013 and the end of 2016, 67.0 percent have been published;

if 2017 papers are included, that number falls to 64.0 percent. Of preprints posted in 2018, only 20.0 percent have been printed elsewhere (**Figure 3a**).

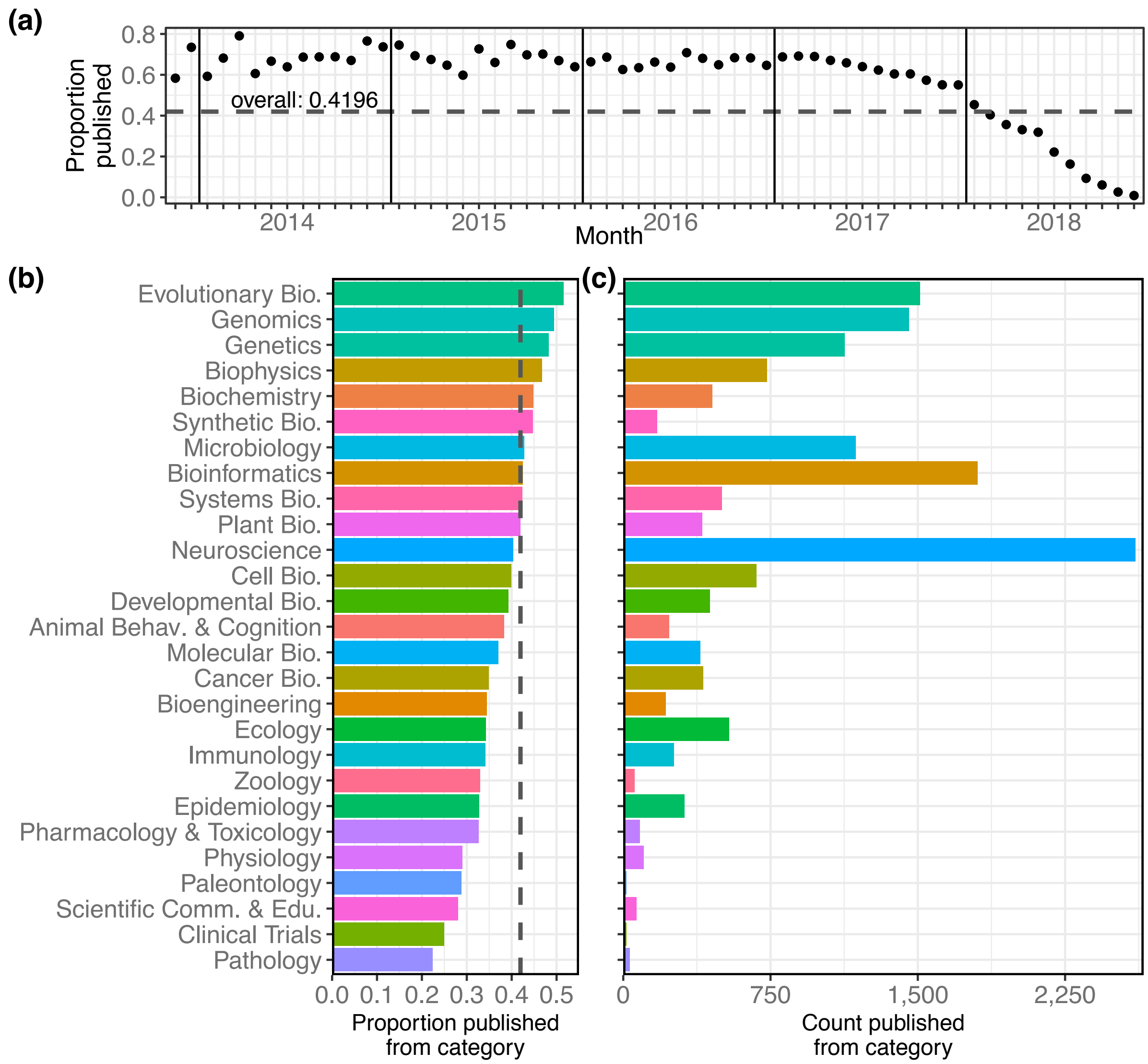



**Figure 3.** Characteristics of the bioRxiv preprints published in journals, across the 27 subject collections. **(a)** The proportion of preprints that have been published (y-axis), broken down by the month in which the preprint was first posted (x-axis). **(b)** The proportion of preprints in each category that have been published elsewhere. The dashed line marks the overall proportion of bioRxiv preprints that have been published and is at the

same position as the dashed line in panel 3a. **(c)** The number of preprints in each category that have been published in a journal.

**Supplementary files:** *publication_rate_month.csv*, *publications_per_category.csv*

Overall, 15,797 bioRxiv preprints have appeared in 1,531 different journals (**Figure 4**). *Scientific Reports* has published the most, with 828 papers, followed by *eLife* and *PLOS ONE* with 750 and 741 papers, respectively. Some journals have accepted a broad range of preprints, though none have hit all 27 of bioRxiv's categories—*PLOS ONE* has published the most diverse category list, with 26. (It has yet to publish a preprint from the clinical trials collection, bioRxiv's second-smallest.) Other journals are much more specialized, though in expected ways: Of the 172 bioRxiv preprints published by *The Journal of Neuroscience*, 169 were in neuroscience, and 3 were from animal behavior and cognition. Similarly, *NeuroImage* has published 211 neuroscience papers, 2 in bioinformatics, and 1 in bioengineering.

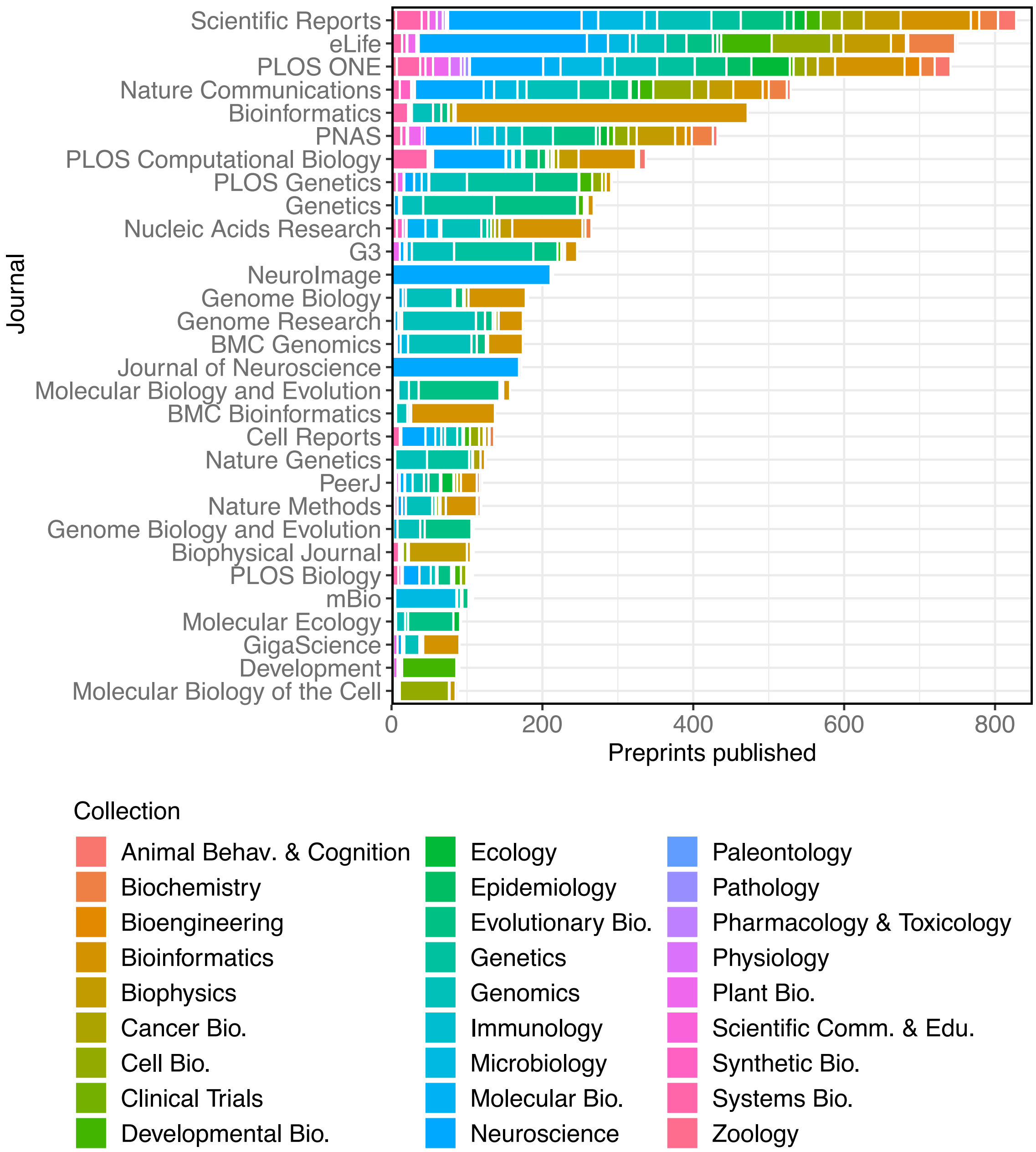



**Figure 4.** A stacked bar graph showing the 30 journals that have published the most preprints. The bars indicate the number of preprints published by each journal, broken down by the bioRxiv categories to which the preprints were originally posted.

**Supplementary file:** *publications_per_journal_categorical.csv*

When evaluating the downloads of preprints published in individual journals (**Figure 5**), there is a significant positive correlation (Kendall's tau=0.5862, p=1.364e-06) between the median downloads per paper and journal impact factor: In general, journals with higher impact scores ("Journal Citation Reports Science Edition" 2018) publish preprints that have more downloads. For example, *Nature Methods* (2017 impact score 26.919) has published 119 bioRxiv preprints; the median download count of these preprints is 2,266. By comparison, *PLOS ONE* (2017 impact score 2.766) has published 719 preprints with a median download count of 279 (**Figure 5**). However, we did not evaluate *when* these downloads occurred, relative to a preprint's publication: While it looks like accruing more downloads makes it more likely that a preprint will appear in a higher impact journal, it is also possible that appearance in particular journals drives bioRxiv downloads after publication.

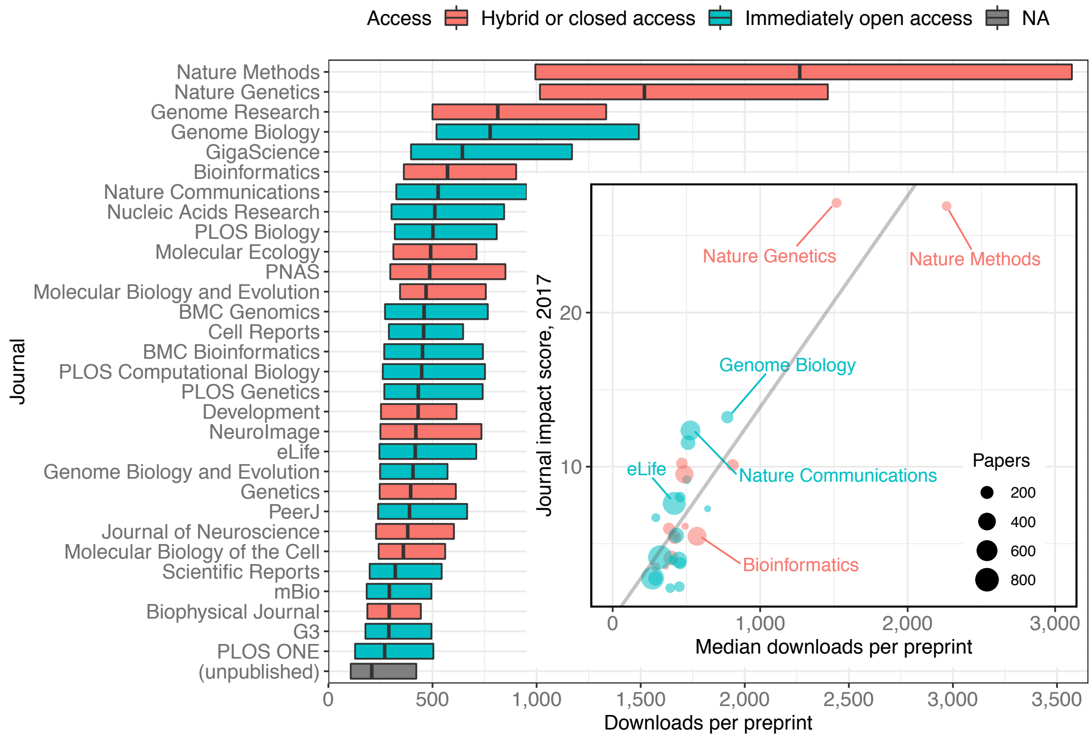



**Figure 5.** A modified box plot (without whiskers) illustrating the median downloads of all bioRxiv preprints published in a journal. Each box illustrates the journal's first quartile, median, and third quartile, as in Figure 2c. Colors correspond to journal access policy as described in the legend. **(inset)** A scatterplot in which each point represents an academic journal, showing the relationship between median downloads of the bioRxiv preprints published in the journal (x-axis) against its most recent impact score (y-axis). The size of each point is scaled to reflect the total number of bioRxiv preprints published by that journal. The regression line in this plot was calculated using the "lm" function in the R "stats" package, but all reported statistics use the Kendall rank correlation coefficient, which does not make as many assumptions about normality or homoscedasticity.

**Supplementary files:** *downloads_journal.csv*, *impact_scores.csv*

If journals are driving post-publication downloads on bioRxiv, however, their efforts are curiously consistent: Preprints that have been published elsewhere have almost twice as many downloads as preprints that have not (**Table 2;** Mann–Whitney *U* test, $p < 2.2e-16$). Site-wide, the median number of downloads per preprint is 208, among papers that have not been published. For preprints that *have* been published, the median download count is 394 (Mann–Whitney *U* test, $p < 2.2e-16$). When preprints published in 2018 are excluded from this calculation, the difference between published and unpublished preprints shrinks, but is still significant (**Table 2;** Mann–Whitney *U* test, $p < 2.2e-16$). Though preprints posted in 2018 received more downloads *in* 2018 than preprints posted in previous years did (**Figure S3**), it appears they have not yet had time to accumulate as many downloads as papers from previous years (**Figure S4**).

| **Posted** | **Published** | **Unpublished** |
|---|---|---|
| 2017 and earlier | 465 | 414 |
| Through 2018 | 394 | 208 |

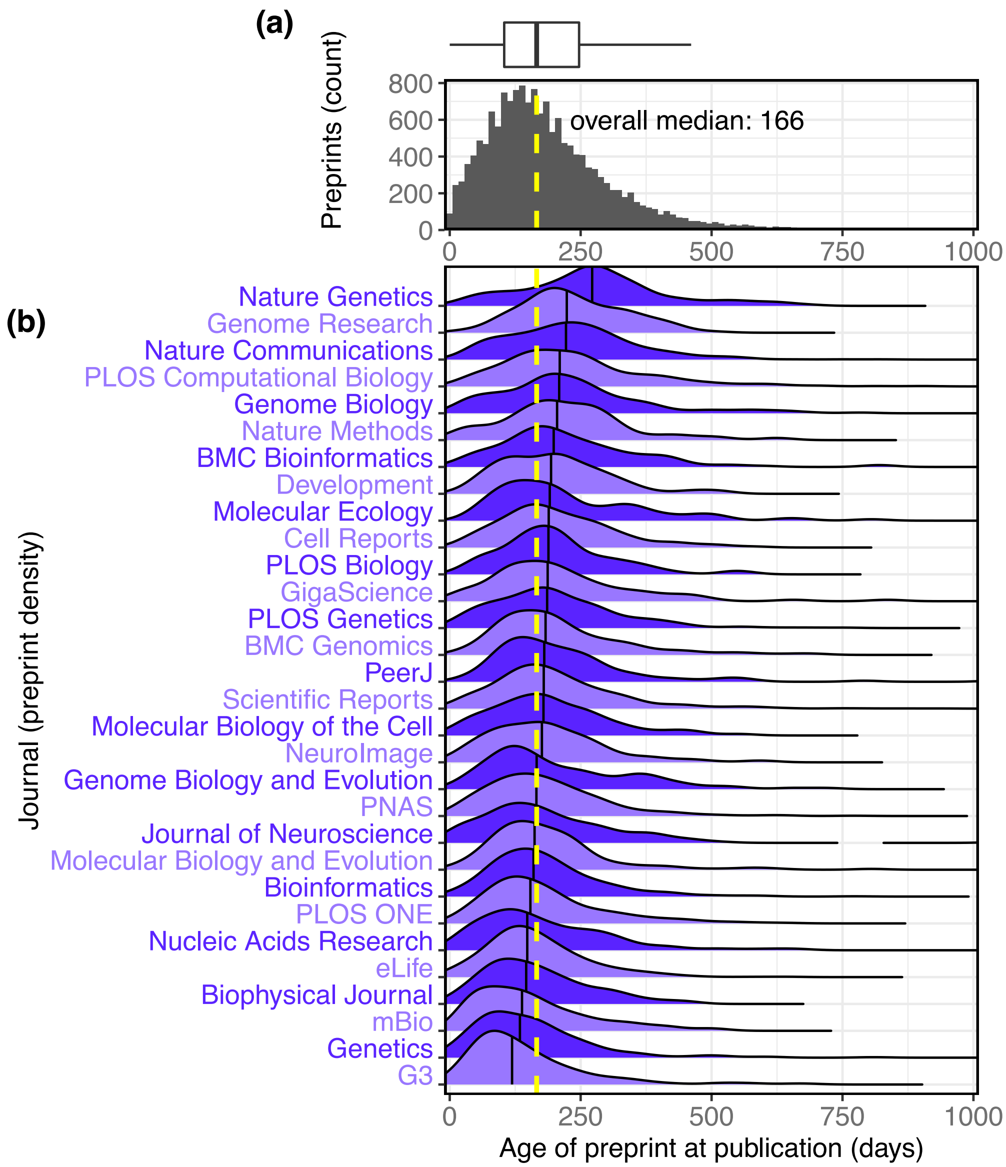
**Table 2.** A comparison of the median downloads per preprint for bioRxiv preprints that have been published elsewhere to those that have not. See Methods section for description of tests used.

**Supplementary file:** *downloads_publication_status.csv*

We also retrieved the publication date for all published preprints using the Crossref "Metadata Delivery" API (Crossref 2018). This, combined with the bioRxiv data, gives us a comprehensive picture of the interval between the date a preprint is first posted to bioRxiv and the date it is published by a journal: These data show the median interval is 166 days, or about 5.5 months. 75 percent of preprints are published within 247 days of appearing on bioRxiv, and 90 percent are published within 346 days (**Figure 6a**). The median interval we found at the end of November 2018 (166 days) is a 23.9 percent increase over the 134-day median interval reported by bioRxiv in mid-2016 (Inglis and Sever 2016).

We also used these data to further examine patterns in the properties of preprints appearing in individual journals: The journal publishing preprints with the highest median age is *Nature Genetics*, whose median interval between bioRxiv posting and publication is 272 days (**Figure 6b**), a significant difference from every journal except *Genome Research* (Kruskal–Wallis rank sum test, $p < 2.2e-16$; Dunn's test $q < 0.05$ comparing *Nature Genetics* to all other journals except *Genome Research*, after Benjamini–Hochberg correction). Among the 30 journals publishing the most bioRxiv preprints, the journal with the most rapid transition from bioRxiv to publication is *G3*, whose median, 119 days, is significantly different from all journals except *Genetics, mBio*, and *The Biophysical Journal* (**Figure 5**).

It is important to note that this metric does not directly evaluate the production processes at individual journals. Authors submit preprints to bioRxiv at different points in the publication process and may work with multiple journals before publication, so individual data points capture a variety of experiences: For example, 122 preprints were published within a week of being posted to bioRxiv, and the longest period between preprint and publication is 3 years, 7 months and 2 days, for a preprint that was posted in March 2015 and not published until October 2018 (**Figure 6a**).




333

334 **Figure 6.** The interval between the date a preprint is posted to bioRxiv
335 and the date it is first published elsewhere. **(a)** A histogram showing the
336 distribution of publication intervals—the x axis indicates the time between
337 preprint posting and journal publication; the y axis indicates how many
338 preprints fall within the limits of each bin. The yellow line indicates the
339 median; the same data is also visualized using a boxplot above the
340 histogram. **(b)** The publication intervals of preprints, broken down by the

journal in which each appeared. The journals in this list are the 30 journals that have published the most total bioRxiv preprints; the plot for each journal indicates the density distribution of the preprints published by that journal, excluding any papers that were posted to bioRxiv after publication. Portions of the distributions beyond 1,000 days are not displayed.

**Supplementary files:** *publication_time_by_year.csv*, *publication_interval_journals.csv*, *journal_interval_dunnstest.csv*

# Discussion

Biology preprints have a large and growing presence in scientific communication, and now we have detailed data to measure and quantify this process. The ability to better characterize the preprint ecosystem can inform decision-making at multiple levels: For authors, particularly those looking for feedback from the community, our results show bioRxiv preprints are being downloaded more than 1 million times per month, and that an average paper can receive hundreds of downloads in its first few months online (**Figure S1**), particularly in genomics, synthetic biology, and bioinformatics (**Figure 2a**). Serghiou and Ioannidis (2018) evaluated download metrics for bioRxiv preprints through 2016 and found an almost identical median for downloads in a preprint's first month; we have expanded this to include more detailed longitudinal traffic metrics for the entire bioRxiv collection (**Figure 2b**). We also quantify which journals have most enthusiastically embraced the publication of biology preprints (**Figure 5**) and begin to evaluate the characteristics of preprints published by individual journals (**Figure 6**). A 2016 project measured which journals had published the most bioRxiv preprints (Schmid 2016); despite a six-fold increase in the number of published preprints since then, 23 of the top 30 journals found in their results are also in the top 30 journals we found.

For readers, we show that more than 2,000 new papers are being posted every month, making bioRxiv an increasingly vital source of information for those seeking to stay on top of the most recent research in their fields. This tracks closely with a widely referenced summary of submissions to preprint servers ("Monthly Statistics for October 2018" 2018) generated monthly by PrePubMed (http://www.prepubmed.org), and expands on submission data collected by researchers using custom web scrapers of their own (Stuart 2016, 2017; Holdgraf 2016). Preprint usage in neuroscience is expanding exceptionally quickly (**Figure 1a**), and collections including bioinformatics, evolutionary biology, and microbiology are growing at a rapid pace (**Figure 1d**). There is also enough data to provide some evidence against the perception that research in preprint is less rigorous than papers appearing in journals ("Methods, preprints and papers" 2017; Vale 2015). In short, the majority of bioRxiv preprints *do* appear in journals eventually, and potentially with very few differences: A 2016 analysis of published preprints that had first been posted to arXiv.org found that "the vast majority of final published papers are largely indistinguishable from their pre-print versions" (Klein et al. 2016).

For preprints that are eventually published, we found the median lag time between posting to bioRxiv and publication in a journal is 166 days (**Figure 6a**), and that 75 percent of preprints are published after 247 days on bioRxiv—more than 8 months. While this number may seem surprisingly short to researchers, it also provides a lengthy head start to readers looking for the most up-to-date research. The distribution of time to publication is similar to the results from Larivière et al. (2014) showing preprints on arXiv were most frequently published within a year of being posted there, and to a later study examining bioRxiv preprints that found "the probability of publication in the peer-reviewed literature was 48% within 12 months" (Serghiou and Ioannidis 2018). Another study published in spring 2017 found that 33.6 percent of preprints from 2015 and earlier had been published (Schloss 2017); our data through November 2018 show that 68.2 percent of preprints from 2015 and earlier have been published. Multiple studies have examined the interval between submission and publication at individual journals (e.g. Himmelstein 2016a; Royle 2015; Powell 2016), but the incorporation of information about preprints is not as common. We believe this is the first time granular publication rates and timeline statistics have been reported for bioRxiv.

More broadly, our data provide a new level of detail. BioRxiv has been the chief facilitator in a paradigmatic shift in biology publishing, and there are still many questions to be answered: What factors may impact the interval between when a preprint is posted to bioRxiv and when it is published elsewhere? Does a paper's presence on bioRxiv have any relationship to its eventual citation count once it is published in a journal, as has been found with arXiv (e.g. Feldman et al. 2018; Wang et al. 2018; Schwarz and Kennicutt 2004)? What can we learn from "altmetrics" as they relate to preprints, and is there value in measuring a preprint's impact using methods rooted in online interactions rather than citation count (Haustein 2018)? One study, published before bioRxiv launched, found a significant association between Twitter mentions of published papers and their citation count (Thelwall et al. 2013)—have preprints changed this dynamic?

Researchers have used existing resources and custom scripts to answer questions like these. Himmelstein (2016b) found that only 17.8 percent of bioRxiv papers had an "open license," for example, and another study examined the relationship between Facebook "likes" of preprints and "traditional impact indicators" such as citation count, but found no correlation for papers on bioRxiv (Ringelhan et al. 2015). Since most bioRxiv data is not programmatically accessible, many of these studies had to begin by scraping data from the bioRxiv website itself. There have been stated plans to change transition to a new, open-source system (Callaway 2017), but the database and API developed here (https://rxivist.org) bring bioRxiv data one step closer to parity with the programmatic interface available for arXiv ("arXiv API" 2018). The Rxivist API allows users to request the details of any preprint or author on the bioRxiv website, and the database snapshots enable bulk querying of preprints using SQL, C, and several other languages ("Procedural Languages" 2019) at a level of complexity currently unavailable using the standard bioRxiv web interface. Using these resources, researchers can now perform detailed and robust bibliometric analysis of the website with the largest collection of preprints in biology, the one that, beginning in September 2018, held more biology preprints than all other major preprint servers combined (Anaya 2018).

In addition to our analysis here focused on big picture trends related to bioRxiv, the Rxivist website provides many additional features that may interest preprint readers. Its primary feature is sorting and filtering preprints based by download count or mentions on Twitter, to help users find preprints in particular categories that are being discussed either in the short term (Twitter) or over the span of months (downloads). Several other sites have attempted to use social interaction data to "rank" preprints, though none incorporate bioRxiv download metrics. The "Assert" web application (https://assert.pub) ranks preprints from multiple repositories based on data from Twitter and GitHub. The "PromisingPreprints" Twitter bot (https://twitter.com/PromPreprint) accomplishes a similar goal, posting links to bioRxiv preprints that receive an exceptionally high social media attention score ("How Is the Altmetric Attention Score Calculated?" 2018) from Altmetric (https://www.altmetric.com) in their first week on bioRxiv (De Coster 2017). Arxiv Sanity Preserver (http://www.arxiv-sanity.com) provides rankings of arXiv.org preprints based on Twitter activity, though its implementation of this scoring (Karpathy 2018) is more opinionated than that of Rxivist. Other websites perform similar curation, but based on user interactions within the sites themselves: SciRate (https://scirate.com), Paperkast (https://paperkast.com) and upvote.pub allow users to vote on articles that should receive more attention (van der Silk et al. 2018; Özturan 2018), though upvote.pub is no longer online ("Frontpage" 2018). By comparison, Rxivist doesn't rely on user interaction—by pulling "popularity" metrics from Twitter and bioRxiv, we aim to decouple the quality of our data from the popularity of the website itself.

In summary, our approach provides multiple perspectives on trends in biology preprints: (1) the Rxivist.org website, where readers can prioritize preprints and generate reading lists tailored to specific topics, (2) a dataset that can provide a foundation for developers and bibliometric researchers to build new tools, websites, and studies that can further improve the ways we interact with preprints, and (3) an analysis that brings together a comprehensive summary of trends in bioRxiv preprints and an examination of the crossover points between preprints and conventional publishing.

# Methods

## Data availability

There are multiple web links to resources related to this project:

- The Rxivist application is available on the web at https://rxivist.org and via Gopher at gopher://origin.rxivist.org
- The source for the web crawler and API is available at https://github.com/blekhmanlab/rxivist
- The source for the Rxivist website is available at https://github.com/blekhmanlab/rxivist_web
- Data files used to generate the figures in this manuscript are available on Zenodo at https://doi.org/10.5281/zenodo.2465689, as is a snapshot of the database used to create the files.

## The Rxivist website

We attempted to put the Rxivist data to good use in a relatively straightforward web application. Its main offering is a ranked list of all bioRxiv preprints that can be filtered by areas of interest. The rankings are based on two available metrics: either the count of PDF downloads, as reported by bioRxiv, or the number of Twitter messages linking to that preprint, as reported by Crossref (https://crossref.org). Users can also specify a timeframe for the search—for example, one could request the most downloaded preprints in microbiology over the last two months, or view the preprints with the most Twitter activity since yesterday across all categories. Each preprint and each author is given a separate profile page, populated only by Rxivist data available from the API. These include rankings across multiple categories, plus a visualization of where the download totals for each preprint (and author) fall in the overall distribution across all 37,000 preprints and 170,000 authors.

## The Rxivist API and dataset

The full data described in this paper is available through Rxivist.org, a website developed for this purpose. BioRxiv data is available from Rxivist in two formats: (1) SQL "database dumps" are currently pulled and published weekly on zenodo.org. (See Supplementary Information for a description of the schema.) These convert the entire Rxivist database into binary files that can be loaded by the free and open-source PostgreSQL database management system to provide a local copy of all collected data on every article and author on bioRxiv.org. (2) We also provide an API (application programming interface) from which users can request information in JSON format about individual preprints and authors, or search for preprints based on similar criteria available on the Rxivist website. Complete documentation is available at https://www.rxivist.org/docs .

While the analysis presented here deals mostly with overall trends on bioRxiv, the primary entity of the Rxivist API is the individual research preprint, for which we have a straightforward collection of metadata: title, abstract, DOI (digital object identifier), the name of any journal that has also published the preprint (and its new DOI), and which collection the preprint was submitted to. We also collected monthly traffic information for each preprint, as reported by bioRxiv. We use the PDF download statistics to generate rankings for each preprint, both site-wide and for each collection, over multiple timeframes (all-time, year to date, etc.). In the API and its underlying database schema, "authors" exist separately from "preprints" because an author can be associated with multiple preprints. They are recorded with three main pieces of data: name, institutional affiliation and a unique identifier issued by ORCID. Like preprints, authors are ranked based on the cumulative downloads of all their preprints, and separately based on downloads within individual bioRxiv collections. Emails are collected for each researcher, but are not necessarily unique (See below).

## Data acquisition

**Web crawler design.** To collect information on all bioRxiv preprints, we developed an application that pulled preprint data directly from the bioRxiv website. The primary issue with managing this data is keeping it up to date: Rxivist aims to essentially maintain an accurate copy of a subset of bioRxiv's production database, which means routinely running a web crawler against the website to find any new or updated content as it is posted. We have tried to find a balance between timely updates and observing courteous web crawler behavior; currently, each preprint is re-crawled once every two to three weeks to refresh its download metrics and publication status. The web crawler itself uses Python 3 and requires two primary modules for interacting with external services: Requests-HTML (Reitz 2018) is used for fetching individual web pages and pulling out the relevant data, and the psycopg2 module (Di Gregorio et al. 2018) is used to communicate with the PostgreSQL database that stores all of the Rxivist data (PostgreSQL Global Development Group 2017). PostgreSQL was selected over other similar database management systems because of its native support for text search, which, in our implementation, enables users to search for preprints based on the contents of their titles, abstracts and author list. The API, spider and web application are all hosted within separate Docker containers (Docker Inc. 2018), a decision we made to simplify the logistics required for others to deploy the components on their own: Docker is the only dependency, so most workstations and servers should be able to run any of the components.

New preprints are recorded by parsing the section of the bioRxiv website that lists all preprints in reverse-chronological order: At this point, a preprint's title, URL and DOI are recorded. The bioRxiv webpage for each preprint is then crawled to obtain details only available there: the abstract, the date the preprint was first posted, and monthly download statistics are pulled from here, as well as information about the preprint's authors—name, email address and institution. These authors are then compared against the list of those already indexed by Rxivist, and any unrecognized authors have profiles created in the database.

**Consolidation of author identities.** Authors are most reliably identified across multiple papers using the bioRxiv feature that allows authors to specify an identifier provided by ORCID (https://orcid.org), a nonprofit that provides a voluntary system to create unique identification numbers for individuals. These ORCID ("Open Researcher and Contributor ID") numbers are intended to serve approximately the same role for authors that DOI numbers do for papers (Haak 2012), providing a way to identify individuals whose other information may change over time. 29,559 bioRxiv authors, or 17.4 percent, have an associated ORCID. If an individual included in a preprint's list of authors doesn't have an ORCID already recorded in the database, authors are consolidated if they have an identical name to an existing Rxivist author.

There are certainly authors who are duplicated within the Rxivist database, an issue arising mostly from the common complaint of unreliable source data. 68.4 percent of indexed authors have at least one email address associated with them, for example, including 7,085 (4.40 percent) authors with more than one. However, of the 118,490 email addresses in the Rxivist database, 6,517 (5.50 percent) are duplicates that are associated with more than one author. Some of these are because real-life authors occasionally appear under multiple names, but other duplicates are caused by uploaders to bioRxiv using the same email address for multiple authors on the same preprint, making it far more difficult to use email addresses as unique identifiers. There are also cases like one from 2017, in which 16 of the 17 authors of a preprint were listed with the email address "test@test.com."

Inconsistent naming patterns cause another chronic issue that is harder to detect and account for. For example, at one point thousands of duplicate authors were indexed in the Rxivist database with various versions of the same name—including a full middle name, or a middle initial, or a middle initial with a period, and so on—which would all have been recorded as separate people if they did not all share an ORCID, to say nothing of authors who occasionally skip specifying a middle initial altogether. Accommodations could be made to account for inconsistencies such as these (using institutional affiliation or email address as clues, for example), but these methods also have the potential to increase the opposite problem of incorrectly combining different authors with similar names who intentionally introduce slight modifications such as a middle initial to help differentiate themselves. One allowance was made to normalize author names: When the web crawler searches for name matches in the database, periods are now ignored in string matches, so "John Q. Public" would be a match with "John Q Public." The other naming problem we encountered was of the opposite variety: multiple authors with identical names (and no ORCID). For example, the Rxivist profile for author "Wei Wang" is associated with 40 preprints and 21 different email addresses but is certainly the conglomeration of multiple researchers. A study of more than 30,000 Norwegian researchers found that when using full names rather than initials, the rate of name collisions was 1.4 percent (Aksnes 2008).

**Retrieval of publication date information.** Publication dates were pulled from the Crossref Metadata Delivery API (Crossref 2018) using the publication DOI numbers provided by bioRxiv. Dates were found for all but 31 (0.2%) of the 15,797 published bioRxiv preprints. Because journals measure "publication date" in different ways, several metrics were used. If a "published—online" date was available from Crossref with a day, month and year, then that was recorded. If not, "published—print" was used, and the Crossref "created" date was the final option evaluated. Requests for which we received a 404 response were assigned a publication date of 1 Jan 1900, to prevent further attempts to fetch a date for those entries. These results were filtered out of the analysis. There was no practical way to validate the nearly 16,000 values retrieved, but anecdotal evaluation reveals some inconsistencies: For example, the preprint with the longest interval before publication (1,371 days) has a publication date reported by Crossref of 1 Jul 2018, when it appeared in *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 15(4). However, the IEEE website lists a date of 15 Dec 2015, two and a half years earlier, as that paper's "publication date," which they define as "the very first instance of public dissemination of content." Since every publisher is free to make their own unique distinctions, these data are difficult to compare at a granular level.

**Calculation of download rankings.** The web crawler's "ranking" step orders preprints and authors based on download count in two populations (overall and by bioRxiv category) and over several periods: all-time, year-to-date, and since the beginning of the previous month. The last metric was chosen over a "month-to-date" ranking to avoid ordering papers based on the very limited traffic data available in the first days of each month—in addition to a short lag in the time bioRxiv takes to report downloads, an individual preprint's download metrics may only be updated in the Rxivist database once every two or three weeks, so metrics for a single month will be biased in favor of those that happen to have been crawled most recently. This effect is not eliminated in longer windows, but is diminished. The step recording the rankings takes a more unusual approach to loading the data: Because each article ranking step could require more than 37,000 "insert" or "update" statements, and each author ranking requires more than 170,000 of the same, these modifications are instead written to a text file on the application server and loaded by running an instance of the Postgres command-line client "psql," which can use the more efficient "copy" command, a change that reduced the ranking process from several hours to less than one minute.

## Data preparation

Several steps were taken to organize the data that was used for this paper. First, the production data being used for the Rxivist API was copied to a separate "schema"—a PostgreSQL term for a named set of tables. This was identical to the full database, but had a specifically circumscribed set of preprints. Once this was copied, the table containing the associations between authors and each of their papers ("article_authors") was pruned to remove references to any articles that were posted after 30 Nov 2018, and any articles that were not associated with a bioRxiv collection. For unknown reasons, 10 preprints (0.03%) could not be associated with a bioRxiv collection; because the bioRxiv profile page for each paper does not specify which collection it belongs to, these papers were ignored. Once these associations were removed, any articles meeting those criteria were removed from the "articles" table. References to these articles were also removed from the table containing monthly bioRxiv download metrics for each paper ("article_traffic"). We also removed all entries from the "article_traffic" table that recorded downloads after November 2018. Next, the table containing author email addresses ("author_emails") was pruned to remove emails associated with any author that had zero preprints in the new set of papers; those authors were then removed from the "authors" table.

Before evaluating data from the table linking published preprints to journals and their post-publication DOI ("article_publications"), journal names were consolidated to avoid under-counting journals with spelling inconsistencies. First, capitalization was stripped from all journal titles, and inconsistent articles ("The Journal of…" vs. "Journal of…"; "and" vs. "&" and so on) were removed. Then, the list of journals was reviewed by hand to remove duplication more difficult to capture automatically: "PNAS" and "Proceedings of the National Academy of Sciences," for example. Misspellings were rare, but one publication in "integrrative biology" did appear. See *figures.md* in the project's GitHub repository (https://github.com/blekhmanlab/rxivist/blob/master/paper/figures.md) for a full list of corrections made to journal titles. We also evaluated preprints for publication in "predatory journals," organizations that use irresponsibly low academic standards to bolster income from publication fees (Xia et al. 2015). A search for 1,345 journals based on the list compiled by Stop Predatory Journals (https://predatoryjournals.com) showed that bioRxiv lists zero papers appearing in those publications ("List of Predatory Journals" 2018).

## Data analysis

**Reproduction of figures.** Two files are needed to recreate the figures in this manuscript: a compressed database backup containing a snapshot of the data used in this analysis, and a file called *figures.md* storing the SQL queries and R code necessary to organize the data and draw the figures. The PostgreSQL documentation for restoring database dumps should provide the necessary steps to "inflate" the database snapshot, and each figure and table is listed in *figures.md* with the queries to generate comma-separated values files that provide the data underlying each figure. (Those who wish to skip the database reconstruction step will find CSVs for each figure provided along with these other files.) Once the data for each figure is pulled into files, executing the accompanying R code should create figures containing the exact data as displayed here.

**Tallying institutional authors and preprints.** When reporting the counts of bioRxiv authors associated with individual universities, there are several important caveats: First, these counts only include the most recently observed institution for an author on bioRxiv: If someone submits 15 preprints at Stanford, then moves to the University of Iowa and posts another preprint afterward, that author will be associated with the University of Iowa, which will receive all 16 preprints in the inventory. Second, this count is also confounded by inconsistencies in the way authors report their affiliations: For example, "Northwestern University," which has 396 preprints, is counted separately from "Northwestern University Feinberg School of Medicine," which has 76. Overlaps such as these were not filtered, though commas in institution names were omitted when grouping preprints together.

**Evaluation of publication rates.** Data referenced in this manuscript is limited to preprints posted through the end of November 2018. However, determining which preprints had been published in journals by the end of November required refreshing the entries for all 37,000 preprints *after* the month ended. Consequently, it's possible that papers published after the end of November (but not after the first weeks of December) are included in the publication statistics.

**Calculation of publication intervals.** There are 15,797 distinct preprints with an associated date of publication in a journal, a corpus too large to allow detailed manual validation across hundreds of journal websites. Consequently, these dates are only as accurate as the data collected by Crossref from the publishers. We attempted to use the earliest publication date, but researchers have found that some publishers may be intentionally manipulating dates associated with publication timelines (Royle 2015), particularly the gap between online and print publication, which can inflate journal impact factor (Tort et al. 2012). Intentional or not, these gaps may be inflating the time to press measurements of some preprints and journals in our analysis. In addition, there are 66 preprints (0.42 percent) that have a publication date that falls before the date it was posted to bioRxiv; these were excluded from analyses of publication interval.

**Counting authors with middle initials.** To obtain the comparatively large counts of authors using one or two middle initials, results from a SQL query were used without any curation. For the counts of authors with three or four middle initials, the results of the database call were reviewed by hand to remove "author" names that look like initials, but are actually the name of consortia ("International IBD Genetics Consortium") or authors who provided non-initialized names using all capital letters.

## Acknowledgements

We thank the members of the Blekhman lab, Kevin M. Hemer, and Kevin LaCherra for helpful discussions. We also thank the bioRxiv staff at Cold Spring Harbor Laboratory for building a valuable tool for scientific communication, and also for not blocking our web crawler even when it was trying to read every web page they have. We are grateful to Crossref for maintaining an extensive, freely available database of publication data. The research was supported in part by funds from the University of Minnesota College of Biological Sciences, NIH grant R35-GM128716, and a McKnight Land-Grant Professorship.

## Competing interests

The authors declare no competing interests.

## References


Abutaleb, Yasmeen, "Facebook's CEO and wife to give 99 percent of shares to their new foundation." Reuters, 1 Dec 2015. https://www.reuters.com/article/us-markzuckerberg-baby/facebooks-ceo-and-wife-to-give-99-percent-of-shares-to-their-new-foundation-idUSKBN0TK5UG20151202

Aksnes, Dag W. 2008. "When different persons have an identical author name. How frequent are homonyms?" *Journal of the Association for Information Science and Technology* 59: 838-841. doi: 10.1002/asi.20788

Anaya, Jordan. 2018. PrePubMed: analyses (version 674d5aa). https://github.com/OmnesRes/prepub/tree/master/analyses/preprint_data.txt

"arXiv API," arXiv (accessed 18 Dec 2018). https://arxiv.org/help/api/index

Barsh, Gregory S., Casey M. Bergman, Christopher D. Brown, Nadia D. Singh, and Gregory P. Copenhaver. 2016. "Bringing PLOS Genetics Editors to Preprint Servers," *PLOS Genetics* 12(12): e1006448. doi: 10.1371/journal.pgen.1006448

Berg, Jeremy M., Needhi Bhalla, Philip E. Bourne, Martin Chalfie, David G. Drubin, James S. Fraser, Carol W. Greider, Michael Hendricks, Chonnettia Jones, Robert Kiley, Susan King, Marc W. Kirschner, Harlan M. Krumholz, Ruth Lehmann, Maria Leptin, Bernd Pulverer, Brooke Rosenzweig, John E. Spiro, Michael Stebbins, Carly Strasser, Sowmya Swaminathan, Paul Turner, Ronald D. Vale, K. VijayRaghavan, and Cynthia Wolberger. 2016. "Preprints for the life sciences," *Science* 352(6288), pp. 899–901. doi: 10.1126/science.aaf9133

Callaway, Ewen. 2013. "Preprints come to life," *Nature* 503, p. 180. doi: 10.1038/503180a

———. 2017. "BioRxiv preprint server gets cash boost from Chan Zuckerberg Initiative," *Nature* 545(18). doi: 10.1038/nature.2017.21894

Champieux, Robin. "Gathering Steam: Preprints, Librarian Outreach, and Actions for Change," The Official PLOS Blog, 15 Oct 2018 (accessed 18 Dec 2018). https://blogs.plos.org/plos/2018/10/gathering-steam-preprints-librarian-outreach-and-actions-for-change/

Cobb, Matthew. 2017. "The prehistory of biology preprints: A forgotten experiment from the 1960s," *PLOS Biology* 15(11): e2003995. doi: 10.1371/journal.pbio.2003995

De Coster, Wouter. "A Twitter bot to find the most interesting bioRxiv preprints," Gigabase or gigabyte, 8 Aug 2017 (accessed 11 Dec 2018).

https://gigabaseorgigabyte.wordpress.com/2017/08/08/a-twitter-bot-to-find-the-most-interesting-biorxiv-preprints/

Crossref Metadata Delivery REST API. Web service (accessed 19 Dec 2018). https://www.crossref.org/services/metadata-delivery/rest-api/

Delamothe, Tony, Richard Smith, Michael A Keller, John Sack, and Bill Witscher. 1999. "Netprints: the next phase in the evolution of biomedical publishing," *BMJ* 319(7224): 1515–6. doi: 10.1136/bmj.319.7224.1515

Desjardins-Proulx, Philippe, Ethan P. White, Joel J. Adamson, Karthik Ram, Timothée Poisot, and Dominique Gravel. 2013. "The case for open preprints in biology," *PLOS Biology* 11(5). doi: 10.1371/journal.pbio.1001563

Di Gregorio, Federico, and Daniele Varrazzo. 2018. psycopg2 (version 2.7.5). https://github.com/psycopg/psycopg2

Docker Inc. 2018. Docker (version 18.06.1-ce). https://www.docker.com

Feldman, Sergey, Kyle Lo, and Waleed Ammar. 2018. "Citation Count Analysis for Papers with Preprints," arXiv. https://arxiv.org/abs/1805.05238

Fowler, Kristine K. 2011. "Mathematicians' Views on Current Publishing Issues: A Survey of Researchers," *Issues in Science and Technology Librarianship* 67. doi: 10.5062/F4QN64NM

"Frontpage," upvote.pub. Archive.org snapshot, 30 Apr 2018 (accessed 29 Dec 2018). https://web.archive.org/web/20180430180959/https://upvote.pub/

"Funding Opportunities," Chan Zuckerberg Initiative, accessed 18 Dec 2018. https://chanzuckerberg.com/science/#funding-opportunities

Haak, Laure. "The O in ORCID," ORCiD, 5 Dec 2012 (accessed 30 Nov 2018). https://orcid.org/blog/2012/12/06/o-orcid

Hartgerink, C.H.J. 2015. "Publication cycle: A study of the public Library of Science (PLOS)," accessed 4 Dec 2018. https://www.authorea.com/users/2013/articles/36067-publication-cycle-a-study-of-the-public-library-of-science-plos/_show_article

Haustein, Stefanie. 2018. "Scholarly Twitter Metrics," arXiv. http://arxiv.org/abs/1806.02201

Himmelstein, Daniel, "The history of publishing delays," Satoshi Village, 10 Feb 2016 (accessed 29 Dec 2018). https://blog.dhimmel.com/history-of-delays/

———, "The licensing of bioRxiv preprints," Satoshi Village, 5 Dec 2016 (accessed 29 Dec 2018). https://blog.dhimmel.com/biorxiv-licenses/

Holdgraf, Christopher R. "The bleeding edge of publishing, Scraping publication amounts at biorxiv," Predictably Noisy, 19 Dec 2016 (accessed 30 Nov 2018). https://predictablynoisy.com/scrape-biorxiv

"How Is the Altmetric Attention Score Calculated?" Altmetric Support, 5 Apr 2018 (accessed 30 Nov 2018).

https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated

Inglis, John R., and Richard Sever, "bioRxiv: a progress report." ASAPbio. 12 Feb 2016 (accessed 5 Dec 2018). http://asapbio.org/biorxiv

"ERA Home," The Lancet Electronic Research Archive. Archive.org snapshots, 22 Apr 2005 and 30 Jul 2005 (accessed 3 Jan 2019). https://web.archive.org/web/20050422224839/http://www.thelancet.com/era

"Journal Citation Reports Science Edition." 2018. Clarivate Analytics.

Kaiser, Jocelyn. 2017. "The preprint dilemma," *Science* 357(6358):1344–1349. doi: 10.1126/science.357.6358.1344

Karpathy, Andrej. 2018. Arxiv Sanity Preserver, "twitter_daemon.py" (version 8e52b8b). https://github.com/karpathy/arxiv-sanity-preserver/blob/8e52b8ba59bfb5684f19d485d18faf4b7fba64a6/twitter_daemon.py

Klein, Martin, Peter Broadwell, Sharon E. Farb, and Todd Grappone. 2016. "Comparing published scientific journal articles to their pre-print versions," *Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 153–162. doi: 10.1145/2910896.2910909

Kling, Rob, Lisa B. Spector, and Joanna Fortuna. 2003. "The real stakes of virtual publishing: The transformation of E-Biomed into PubMed central," *Journal of the Association for Information Science and Technology* 55(2):127–48. doi: 10.1002/asi.10352

Larivière, Vincent, Cassidy R. Sugimoto, Benoit Macaluso, Staša Milojević, Blaise Cronin, and Mike Thelwall. 2014. "arXiv E-prints and the journal of record: An analysis of roles and relationships," *Journal of the Association for Information Science and Technology* 65(6), pp. 1157–69. doi: 10.1002/asi.23044

"List of Predatory Journals," Stop Predatory Journals (accessed 28 Dec 2018). https://predatoryjournals.com/journals/

Marshall, Eliot. 1999. "PNAS to Join PubMed Central--On Condition," *Science* 286(5440):655–6. doi: 10.1126/science.286.5440.655a

McConnell, John, and Richard Horton. 1999. "Lancet electronic research archive in international health and eprint server," *The Lancet* 354(9172):2–3. doi: 10.1016/S0140-6736(99)00226-3.

"Methods, preprints and papers." 2017. *Nature Biotechnology* 35(12). doi: 10.1038/nbt.4044

"Monthly Statistics for October 2018," PrePubMed, accessed 17 Dec 2018. http://www.prepubmed.org/monthly_stats/

"Nature respects preprint servers," 2005. *Nature* 434, p. 257. doi: 10.1038/434257b

O'Roak, Brian. "How I learned to stop worrying and love preprints," Spectrum, 22 May 2018 (accessed 30 Nov 2018). https://www.spectrumnews.org/opinion/learned-stop-worrying-love-preprints/

Özturan, Doğancan. "Paperkast: Academic article sharing and discussion," 2 Sep 2018 (accessed 8 Jan 2019). https://medium.com/@dogancan/paperkast-academic-article-sharing-and-discussion-e1aebc6fe66d

PostgreSQL Global Development Group. 2017. PostgreSQL (version 9.6.6). https://www.postgresql.org

Powell, Kendall. 2016. "Does it take too long to publish research?" *Nature* 530, pp. 148–151. doi: 10.1038/530148a

"Procedural Languages," PostgreSQL Documentation (version 9.4.20), accessed 1 Jan 2019. https://www.postgresql.org/docs/9.4/xplang.html

Raff, Martin, Alexander Johnson, and Peter Walter. 2008. "Painful Publishing," *Science* 321(5885):36. doi: 10.1126/science.321.5885.36a

Reitz, Kenneth. 2018. Requests-HTML (version 0.9.0). https://github.com/kennethreitz/requests-html

"Reporting Preprints and Other Interim Research Products," notice number NOT-OD-17-050. National Institutes of Health. 24 Mar 2017 (accessed 7 Jan 2019). https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-050.html

Ringelhan, Stefanie, Jutta Wollersheim, and Isabell M. Welpe. 2015. "I Like, I Cite? Do Facebook Likes Predict the Impact of Scientific Work?" *PLOS ONE* 10(8): e0134389. doi: 10.1371/journal.pone.0134389

Rørstad, Kristoffer, and Dag W. Aksnes. 2015. "Publication rate expressed by age, gender and academic position – A large-scale analysis of Norwegian academic staff," *Journal of Informetrics* 9(2). doi: 10.1016/j.joi.2015.02.003

Royle, Stephen, "What The World Is Waiting For," quantixed, 17 Oct 2014 (accessed 29 Dec 2018). https://quantixed.org/2014/10/17/what-the-world-is-waiting-for/

———, "Waiting to happen II: Publication lag times," quantixed, 16 Mar 2015 (accessed 29 Dec 2018). https://quantixed.org/2015/03/16/waiting-to-happen-ii-publication-lag-times/

Schloss, Patrick D. 2017. "Preprinting Microbiology," *mBio* 8:e00438-17. doi: 10.1128/mBio.00438-17

Schmid, Marc W. 2016. crawlBiorxiv (version e2af128). https://github.com/MWSchmid/crawlBiorxiv/blob/master/README.md

Schwarz, Greg J., and Robert C. Kennicutt Jr. 2004. "Demographic and Citation Trends in Astrophysical Journal papers and Preprints," arXiv. https://arxiv.org/abs/astro-ph/0411275

Sever, Richard. Twitter Post. 1 Nov 2018, 9:29 AM. https://twitter.com/cshperspectives/status/1058002994413924352

van der Silk, Noon, Aram Harrow, Jaiden Mispy, Dave Bacon, Steven Flammia, Jonathan Oppenheim, James Payor, Ben Reichardt, Bill Rosgen, Christian Schaffner, and Ben Toner. "About," SciRate, accessed 30 Nov 2018. https://scirate.com/about

Smaglik, Paul. "E-Biomed Becomes Pubmed Central," *The Scientist*, 27 Sep 1999 (accessed 29 Dec 2018). https://www.the-scientist.com/news/e-biomed-becomes-pubmed-central-56359

Snyder, Solomon H. 2013. "Science interminable: Blame Ben?" *PNAS* 110(7):2428–9. doi: 10.1073/pnas.201300924

Stuart, Tim, "bioRxiv," 1 Mar 2016 (accessed 2 Jan 2019). http://timoast.github.io/blog/2016-03-01-biorxiv/

———, "bioRxiv 2017 update," 4 Oct 2017 (accessed 2 Jan 2019). http://timoast.github.io/blog/biorxiv-2017-update/

Serghiou, Stylianos, and John P.A. Ioannidis. 2018. "Altmetric Scores, Citations, and Publication of Studies Posted as Preprints," JAMA 318(4): 402–4. doi: 10.1001/jama.2017.21168

"Submission Guide," bioRxiv, accessed 30 Nov 2018. https://www.biorxiv.org/submit-a-manuscript

Thelwall, Mike, Stefanie Haustein, Vincent Larivière, and Cassidy R. Sugimoto. 2013. "Do Altmetrics Work? Twitter and Ten Other Social Web Services," *PLOS ONE* 8(5): e64841. doi: 10.1371/journal.pone.0064841

Tort, Adriano B.L., Zé H. Targino, and Olavo B. Amaral. 2012. "Rising Publication Delays Inflate Journal Impact Factors," *PLOS ONE* 7(12): e53374. doi: 10.1371/journal.pone.0053374

Vale, Ronald D. 2015. "Accelerating scientific publication in biology," *PNAS* 112(44):13439–46. doi: 10.1073/pnas.1511912112

Vale, Ronald D., and Anthony A. Hyman. 2016. "Priority of discovery in the life sciences," *eLife* 5(e16931). Doi: 10.7554/eLife.16931

Varmus, Harold. "E-BIOMED: A Proposal for Electronic Publications in the Biomedical Sciences," National Institutes of Health, 5 May 1999. Archive.org snapshot, 18 Oct 2015 (accessed 29 Dec 2018). https://web.archive.org/web/20151018182443/https://www.nih.gov/about/director/pubmedcentral/ebiomedarch.htm

Vence, Tracy. "Journals Seek Out Preprints," *The Scientist*, 18 Jan 2017 (accessed 7 Jan 2019). https://www.the-scientist.com/news-opinion/journals-seek-out-preprints-32183

Verma, Inder M. 2017. "Preprint servers facilitate scientific discourse," *PNAS* 114(48). doi: 10.1073/pnas.1716857114

Wang, Zhiqi, Wolfgang Glänzel, and Yue Chen. 2018. "How Self-Archiving Influences the Citation Impact of a Paper: A Bibliometric Analysis of arXiv Papers and Non-arXiv Papers in the Field of Information and Library Science," Leiden, The Netherlands: Proceedings of the 23rd International Conference on Science and Technology Indicators (ISBN: 978-90-9031204-0), pages 323–30. https://openaccess.leidenuniv.nl/handle/1887/65329

Xia, Jingfeng, Jennifer L. Harmon, Kevin G. Connolly, Ryan M. Donnelly, Mary R. Anderson, and Heather A. Howard. 2015. "Who published in ‘predatory’ journals?" *Journal of the Association for Information Science and Technology* 66(7). doi: 10.1002/asi.23265