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: 0000-0001-9565-5832

RB: 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 work to preprint servers at an unprecedented and increasing rate, sharing papers online before (or instead of) publication in peer-reviewed journals. Though the increasing acceptance of preprints is driving policy changes for journals and funders, there is little information about their usage. Here, we collected and analyzed data on all 37,648 preprints uploaded to bioRxiv.org, the largest biology-focused preprint server, in its first five years. We find preprints are being downloaded more than ever before (1.1 million tallied in October 2018 alone) and that the rate of preprints being posted has increased to a recent high of 2,100 per month. We also find that two-thirds of preprints posted before 2017 were later published in peer-reviewed journals, and find a relationship between journal impact factor and preprint downloads. Lastly, we developed Rxivist.org, a web application providing multiple ways of interacting with preprint metadata.

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 than all four—combined—by the end of September 3 (**Figure 1—figure supplement 1**).

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). Researchers have recently 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).

Preprints can short-circuit some of these delays—a practice long familiar to physicists, who began submitting preprints to arXiv, one of the earliest preprint servers, in 1991 (Verma 2017). In the life sciences, however, preprints were approached with greater reluctance (Cobb 2017; Desjardins-Proulx et al. 2013). An early NIH plan for PubMed Central included the hosting of preprints (Varmus 1999; Smaglik 1999) but was scuttled in 1999 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 2013 (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 preprint servers (Anaya 2018), and sporadic updates from bioRxiv leaders show steadily increasing submission numbers (Sever 2018). 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 use of bioRxiv as a research repository and to help quantify trends in biology preprints. 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 becoming an increasingly popular venue for authors to share their work, at a rate that increases almost monthly: 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

108 for five years (**Figure 1b**). The plurality of site-wide growth can be attributed to the
109 neuroscience collection, which has had more submissions than any bioRxiv category in
110 every month since September 2016 (**Figure 1b**). In October 2018, it became the first of
111 bioRxiv's collections to contain 6,000 preprints (**Figure 1a**). The second-largest category
112 is bioinformatics (4,249 preprints), followed by evolutionary biology (2,934). October 2018
113 was also the first month in which bioRxiv posted more than 2,000 preprints, increasing its
114 total preprint count by 6.3 percent (2,119) in 31 days.

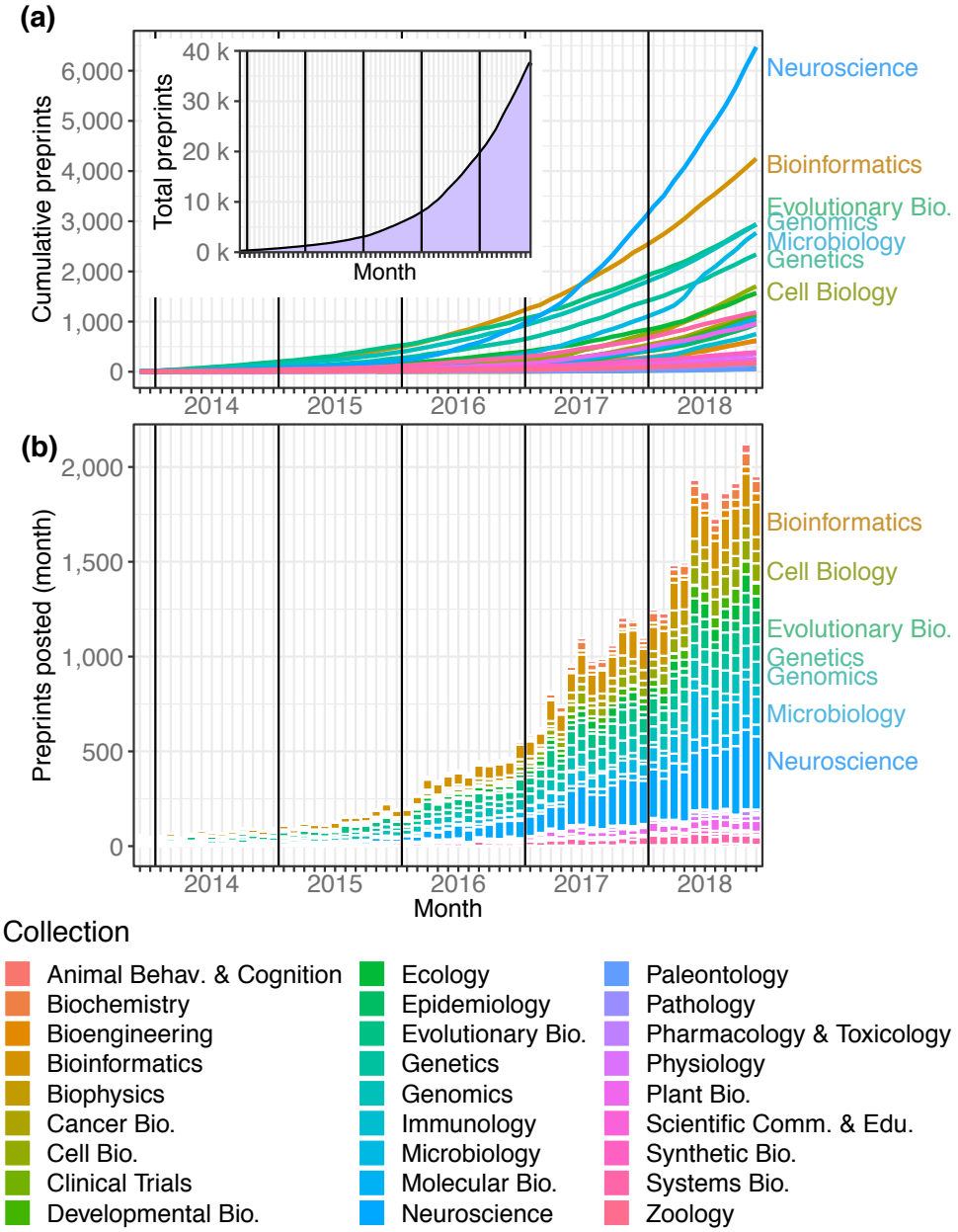

Figure 1. Total preprints posted to bioRxiv over a 61-month period from November 2013 through November 2018. **(a)** The number of preprints (y-axis) at each month (x-axis), with each category depicted as a line in a different color. **(a, inset)** The overall number of preprints on bioRxiv in each month.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

by category. The category color key is provided below the figure.

Figure 1—figure supplement 1: The number of full-length articles published by an arbitrary selection of well-known journals in September 2018.

Figure 1—source data 1: The number of submissions per month to each bioRxiv category, plus running totals. *submissions_per_month.csv*

Figure 1—source data 2: An Excel workbook demonstrating the formulas used to calculate the running totals in Figure 1—source data 1. *submissions_per_month_cumulative.xlsx*

Figure 1—source data 3: The number of submissions per month overall, plus running totals. *submissions_per_month_overall.csv*

Preprint downloads

Using preprint downloads as a metric for readership, we find that bioRxiv's usage among 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**; Mann–Whitney *U* test p = 0.0003).

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 2—figure supplement 2**).

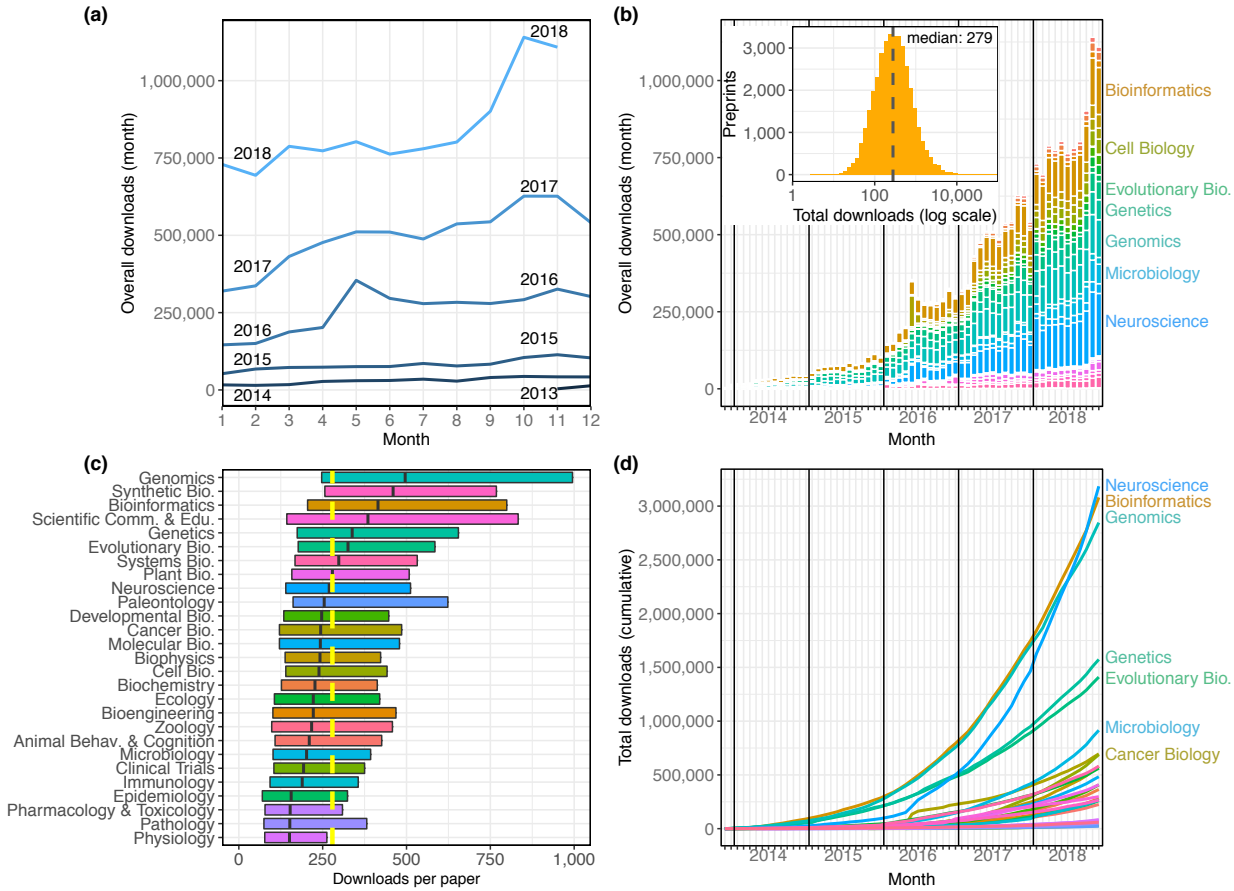

Figure 2. The distribution of all recorded downloads of bioRxiv preprints. **(a)** The downloads recorded in each month, with each line representing a different year. The lines reflect the same totals as the height of the bars in Figure 2b. **(b)** A stacked bar plot of the downloads in each month: The 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.

Figure 2—figure supplement 1: The distribution of downloads that preprints accrue in their first months on bioRxiv.

Figure 2—figure supplement 2. The proportion of downloads that preprints accrue in their first months on bioRxiv.

Figure 2—figure supplement 3. Multiple perspectives on per-preprint download statistics.

Figure 2—figure supplement 4. Total downloads per preprint, segmented by the year in which each preprint was posted.

Figure 2—source data 1: A list of every preprint, its bioRxiv category, and its total downloads. *downloads_per_category.csv*

Figure 2—source data 2: The number of downloads per month in each bioRxiv category, plus running totals. *downloads_per_month_cumulative.csv*

Figure 2—source data 3: An Excel workbook demonstrating the formulas used to calculate the running totals in Figure 2—source data 2. *downloads_per_month.xlsx*

Figure 2—source data 4: The number of downloads per month overall,

plus running totals. *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.

Table 1—table supplement 1: The top 15 authors with the most preprints on bioRxiv.

Table 1—table supplement 2. Top 25 institutions with the most authors listing them as their affiliation, and how many papers have been published by those authors.

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 1—table supplement 1**). 1,473 authors list their most recent affiliation as Stanford University, the most represented

institution on bioRxiv (**Table 1—table supplement 2**). 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**), according to bioRxiv's records linking preprints to their external publications. 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**).

These publication statistics are based on data produced by bioRxiv's internal system that links publications to their preprint versions, a difficult challenge that appears to rely heavily on title-based matching. To better understand the reliability of the linking between preprints and their published versions, we selected a sample of 120 preprints that were not indicated as being published, and manually validated their publication status using Google and Google Scholar (see Methods). Overall, 37.5% of these "unpublished" preprints had actually appeared in a journal. We found earlier years to have a much higher false-negative rate: 53 percent of the evaluated "unpublished" preprints from 2015 had actually been published, though that number dropped to less than 17 percent in 2017 (Figure 3—figure supplement 1). While a more robust study would be required to draw more detailed conclusions about the "true" publication rate, this preliminary examination suggests the data from bioRxiv may be an underestimation of the number of preprints that have actually been published.

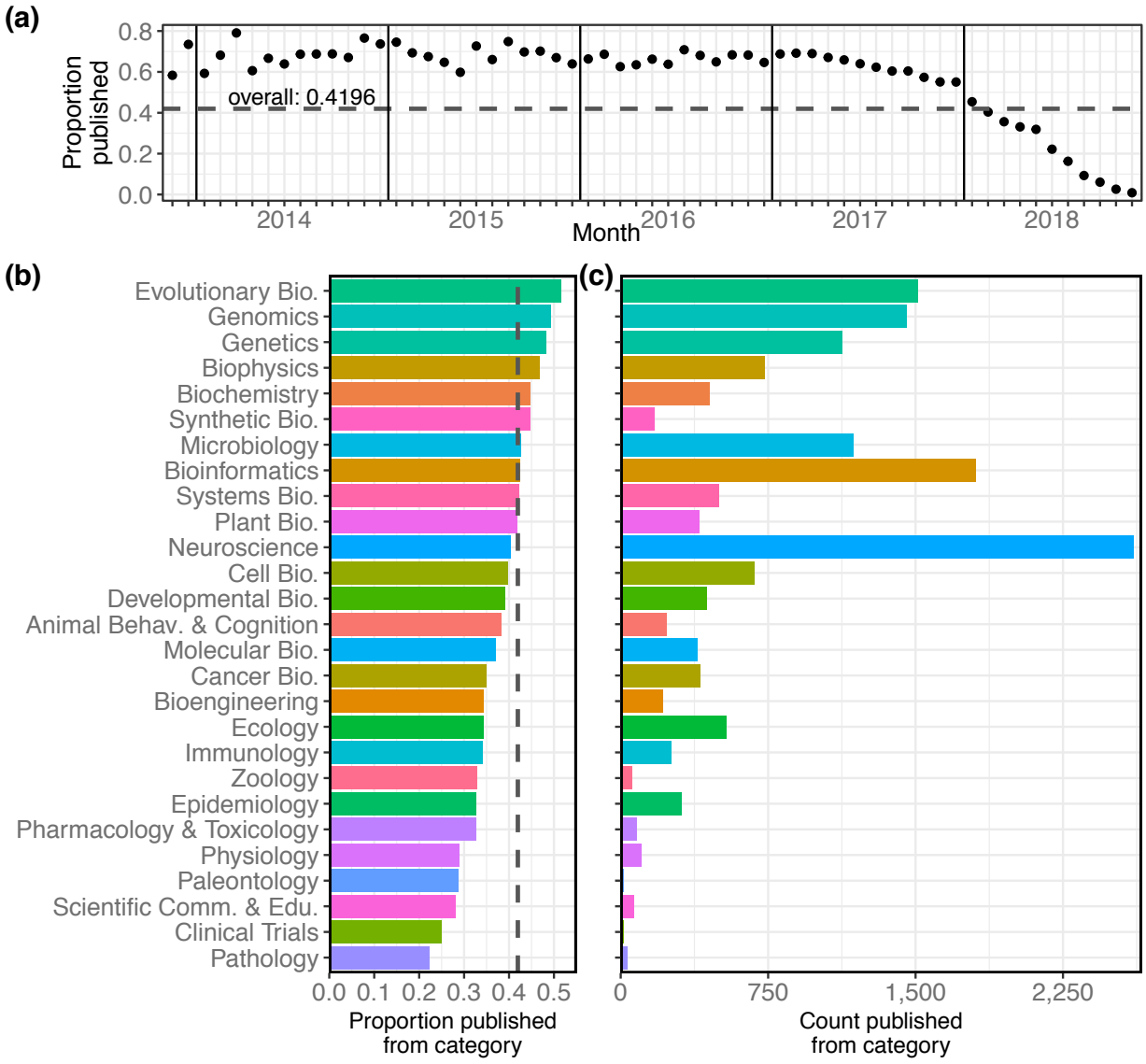

Figure 3. Characteristics of the bioRxiv preprints published in journals, across the 27 subject collections. **(a)** The proportion of preprints that have been published (y-axis), broken down by the month in which the preprint was first posted (x-axis). **(b)** The proportion of preprints in each category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

Figure 3—figure supplement 1: Observed annual publication rates and estimated range for actual publication rates.

Figure 3—source data 1: The number of preprints posted in each month, plus the count and proportion of those later published. *publication_rate_month.csv*

Figure 3—source data 2: The number of preprints posted in each category, plus the count and proportion of those published. *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. However, considering the proportion of preprints of the total papers published in each journal can lead to a different interpretation. For example, *Scientific Reports* published 398 bioRxiv preprints in 2018, but this represents 2.36% of the 16,899 articles it published in that year, as indexed by Web of Science (**Figure 4—figure supplement 1**). In contrast, *eLife* published almost as many bioRxiv preprints (394), which means more than a third of their 1,172 articles from 2018 first appeared on bioRxiv. *GigaScience* had the highest proportion of articles from preprints in 2018 (49.4% of 89 articles), followed by *Genome Biology* (39.9% of 183 articles) and *Genome Research* (36.7% of 169 articles). Incorporating all years in which bioRxiv preprints have been published (2014–2018), these are also the three top journals.

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. It should be noted that these counts are based on the publications detected by bioRxiv and linked to their preprint, so some journals—for example, those that more frequently rewrite the titles of articles—may be underrepresented here.

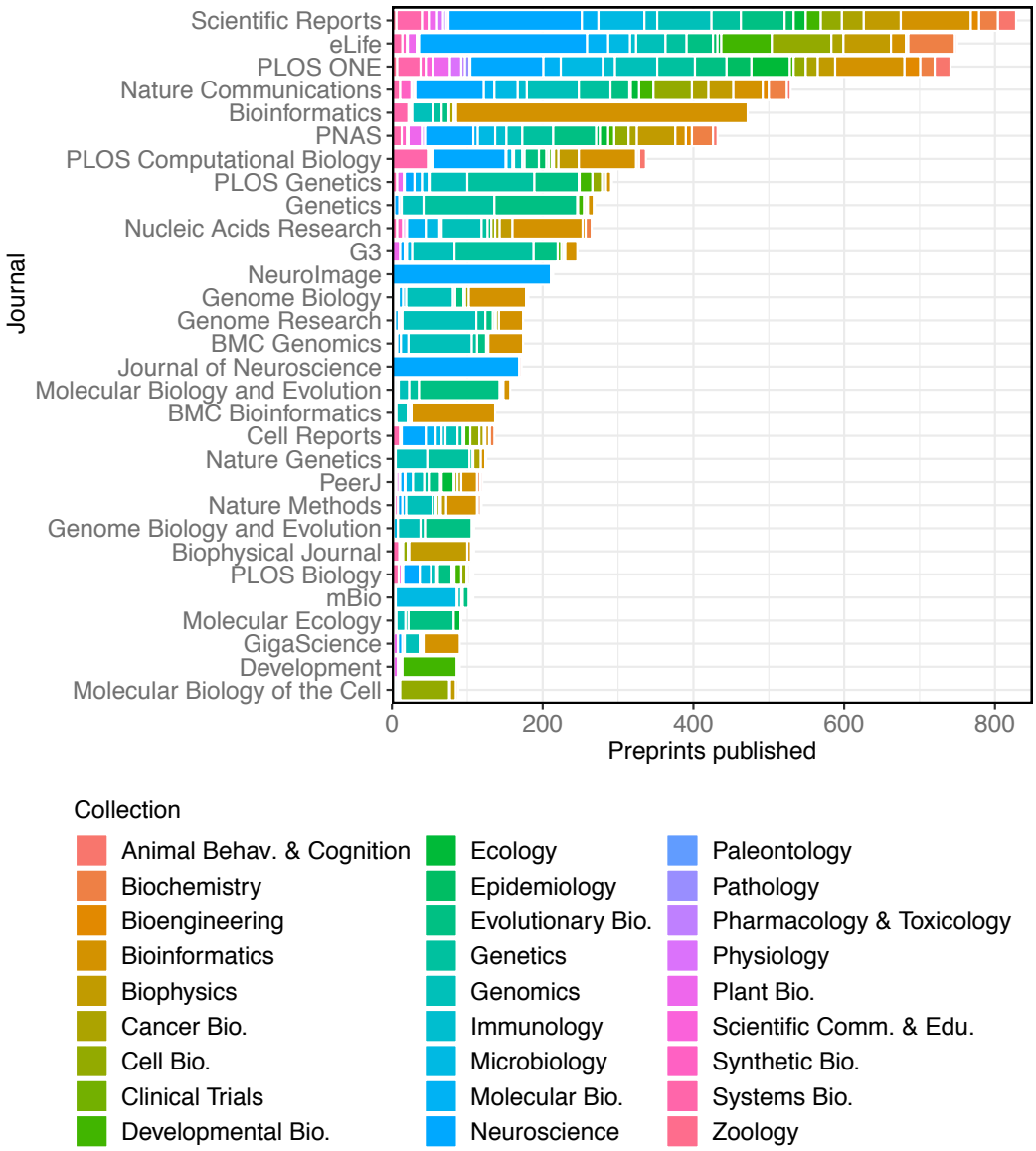

Figure 4. A stacked bar graph showing the 30 journals that have published the most preprints. The bars indicate the number of preprints published by each journal, broken down by the bioRxiv categories to which the preprints were originally posted.

Figure 4—figure supplement 1: The total number of articles published by the top 30 journals that have published the most bioRxiv preprints, compared to how many of those articles appeared on bioRxiv.

Figure 4—source data 1: The number of preprints published in each category by the 30 most prolific publishers of preprints. *publications_per_journal_categorical.csv*

When evaluating the downloads of preprints published in individual journals (**Figure 5**), there is a significant positive correlation between the median downloads per paper and journal impact factor (JIF): In general, journals with higher impact factors ("Journal Citation Reports Science Edition" 2018) publish preprints that have more downloads. For example, *Nature Methods* (2017 journal impact factor 26.919) has published 119 bioRxiv preprints; the median download count of these preprints is 2,266. By comparison, *PLOS ONE* (2017 JIF 2.766) has published 719 preprints with a median download count of 279 (**Figure 5**). In this analysis, each data point in the regression represented a journal, indicating its JIF and the median downloads per paper for the preprints it had published. We found a significant positive correlation between these two measurements (Kendall's τ_b=0.5862, p=1.364e-06). We also found a similar, albeit weaker, correlation when we performed another analysis in which each data point represented a single preprint (n=7,445; Kendall's τ_b=0.2053, p=9.311e-152; see Methods).

It is important to note that 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. The Rxivist dataset has already been used to begin evaluating questions like this (Kramer 2019), and further study may be able to unravel the links, if any, between downloads and journals.

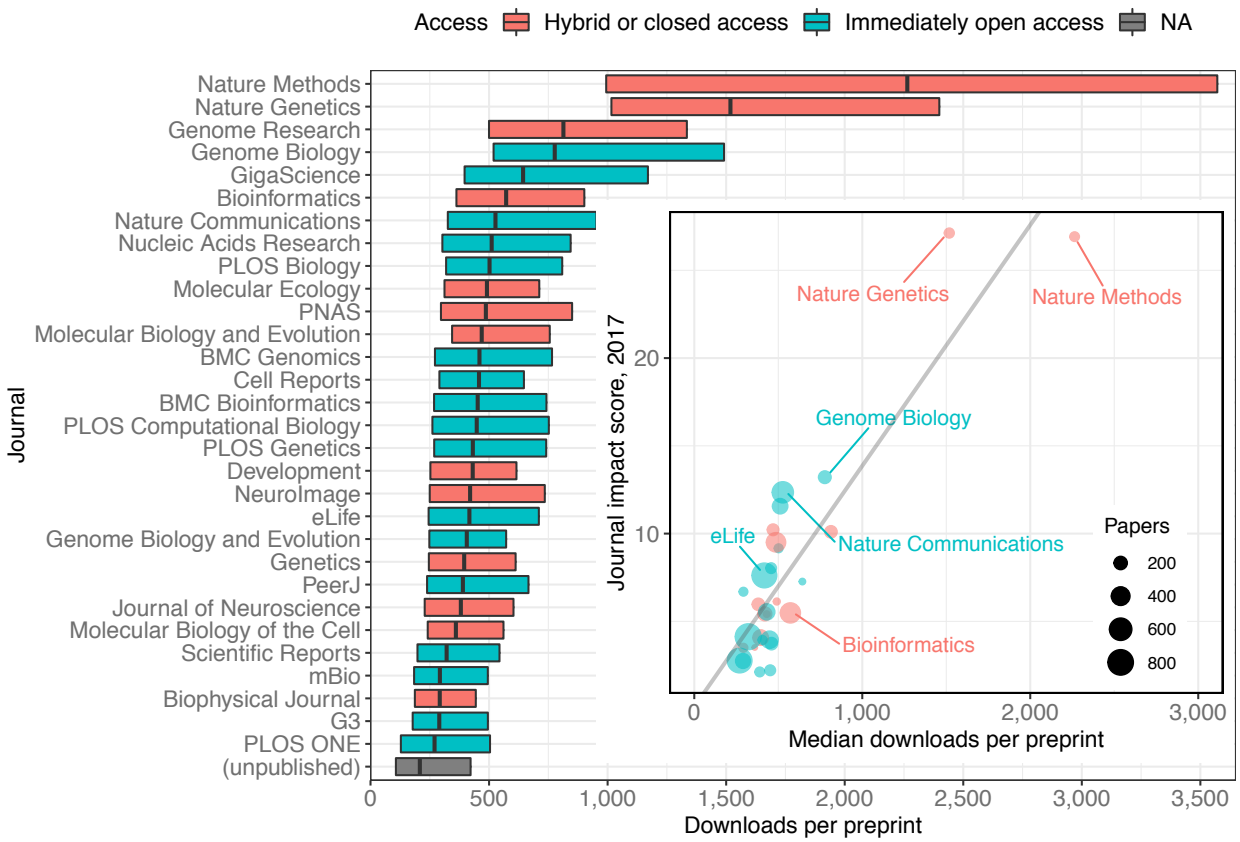

Figure 5. A modified box plot (without whiskers) illustrating the median downloads of all bioRxiv preprints published in a journal. Each box illustrates the journal's first quartile, median, and third quartile, as in Figure 2c. Colors correspond to journal access policy as described in the legend. **(inset)** 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 2017 journal impact factor (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.

Figure 5—source data 1: A list of every preprint with its total download count and the journal in which it was published, if any. *downloads_journal.csv*

Figure 5—source data 2: Journal impact factor and access status of the 30 journals that have published the most preprints. *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, $p < 2.2e\text{-}16$). 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, $p < 2.2e\text{-}16$). 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, $p < 2.2e\text{-}16$). Though preprints posted in 2018 received more downloads *in* 2018 than preprints posted in previous years did (**Figure 2—figure supplement 3**), it appears they have not yet had time to accumulate as many downloads as papers from previous years (**Figure 2—figure supplement 4**).

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.

Table 2—source data 1: A list of every preprint with its total download count, the year in which it was first posted, and whether it has been published. *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, $p < 2.2e\text{-}16$; 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**).

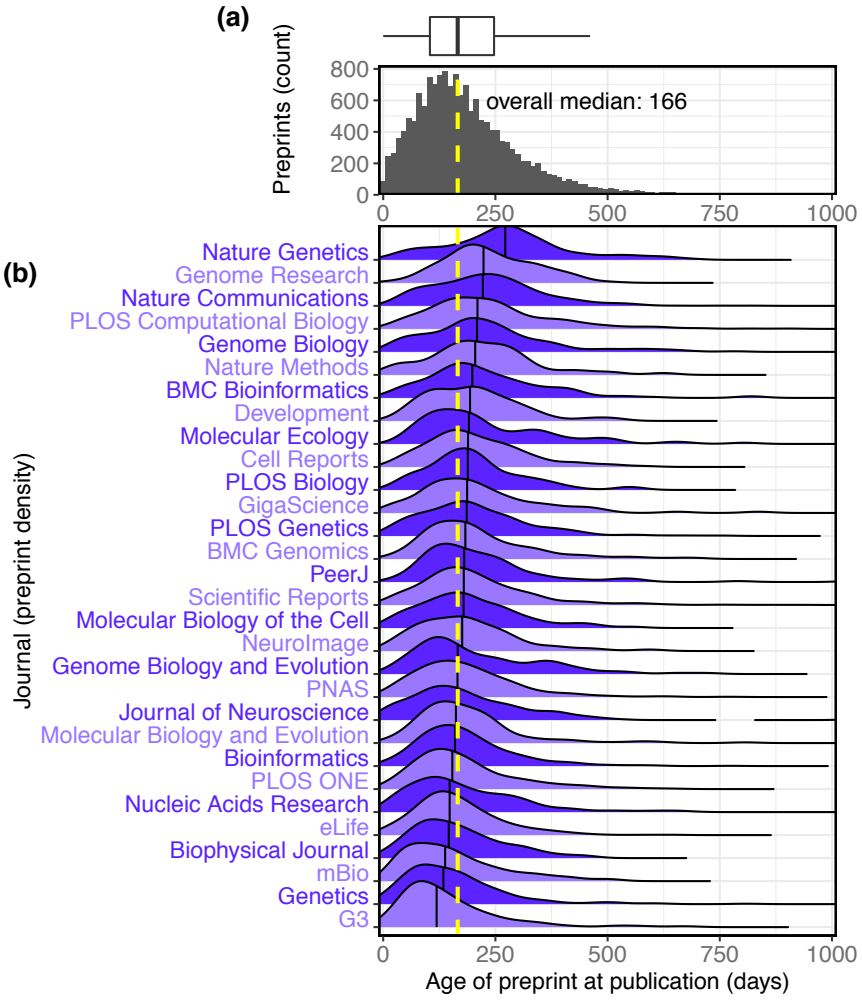

Figure 6. The interval between the date a preprint is posted to bioRxiv and the date it is first published elsewhere. **(a)** A histogram showing the

distribution of publication intervals—the x axis indicates the time between preprint posting and journal publication; the y axis indicates how many preprints fall within the limits of each bin. The yellow line indicates the median; the same data is also visualized using a boxplot above the 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.

Figure 6—source data 1: A list of every published preprint, the year it was first posted, the date it was published, and the interval between posting and publication, in days. *publication_time_by_year.csv*

Figure 6—source data 2: A list of every preprint published in the 30 journals displayed in the figure, the journal in which it was published, and the interval between posting and publication, in days. *publication_interval_journals.csv*

Figure 6—source data 3: The results of Dunn's test, a pairwise comparison of the median publication interval of each journal in the figure. *journal_interval_dunnstest.txt*

413 # Discussion

414 Biology preprints have a growing presence in scientific communication, and we
415 now have ongoing, detailed data to quantify this process. The ability to better characterize
416 the preprint ecosystem can inform decision-making at multiple levels: For authors,
417 particularly those looking for feedback from the community, our results show bioRxiv
418 preprints are being downloaded more than 1 million times per month, and that an average
419 paper can receive hundreds of downloads in its first few months online (**Figure 2—figure**
420 **supplement 1**). Serghiou and Ioannidis (2018) evaluated download metrics for bioRxiv
421 preprints through 2016 and found an almost identical median for downloads in a preprint's
422 first month; we have expanded this to include more detailed longitudinal traffic metrics for
423 the entire bioRxiv collection (**Figure 2b**).

424 For readers, we show that thousands of new preprints are being posted every
425 month. This tracks closely with a widely referenced summary of submissions to preprint
426 servers ("Monthly Statistics for October 2018" 2018) generated monthly by PrePubMed
427 (http://www.prepubmed.org) and expands on submission data collected by researchers
428 using custom web scrapers of their own (Stuart 2016, 2017; Holdgraf 2016). There is also
429 enough data to provide some evidence against the perception that research in preprint is
430 less rigorous than papers appearing in journals ("Methods, preprints and papers" 2017;
431 Vale 2015). In short, the majority of bioRxiv preprints *do* appear in journals eventually, and
432 potentially with very few differences: An analysis of published preprints that had first been
433 posted to arXiv.org found that "the vast majority of final published papers are largely
434 indistinguishable from their pre-print versions" (Klein et al. 2016). A 2016 project measured
435 which journals had published the most bioRxiv preprints (Schmid 2016); despite a six-fold

436 increase in the number of published preprints since then, 23 of the top 30 journals found
437 in their results are also in the top 30 journals we found (**Figure 5**).

438 For authors, we also have a clearer picture of the fate of preprints after they are
439 shared online. Among preprints that are eventually published, we found that 75 percent
440 have appeared in a journal by the time they had spent 247 days (about 8 months) on
441 bioRxiv, an interval similar to results from Larivière et al. (2014) showing preprints on arXiv
442 were most frequently published within a year of being posted there, and to a later study
443 examining bioRxiv preprints that found "the probability of publication in the peer-reviewed
444 literature was 48% within 12 months" (Serghiou and Ioannidis 2018). Another study
445 published in spring 2017 found that 33.6 percent of preprints from 2015 and earlier had
446 been published (Schloss 2017); our data through November 2018 show that 68.2 percent
447 of preprints from 2015 and earlier have been published. Multiple studies have examined
448 the interval between submission and publication at individual journals (e.g. Himmelstein
449 2016a; Royle 2015; Powell 2016), but the incorporation of information about preprints is
450 not as common.

451 We also found a positive correlation between the impact factor of journals and the
452 number of downloads received by the preprints they have published. This finding in
453 particular should be interpreted with caution. Journal impact factor is broadly intended to
454 be a measurement of how citable a journal's "average" paper is (Garfield 2006), though
455 it morphed long ago into an unfounded proxy for scientific quality in individual papers
456 ("The Impact Factor Game" 2006). It is referenced here only as an observation about a
457 journal-level metric correlated with preprint downloads: There is no indication that either
458 factor is influencing the other, nor that download numbers play a direct role in publication

459 decisions.

460 More broadly, our granular data provide a new level of detail for researchers
461 looking to evaluate many remaining questions: What factors may impact the interval
462 between when a preprint is posted to bioRxiv and when it is published elsewhere? Does
463 a paper's presence on bioRxiv have any relationship to its eventual citation count once it
464 is published in a journal, as has been found with arXiv (e.g. Feldman et al. 2018; Wang et
465 al. 2018; Schwarz and Kennicutt 2004)? What can we learn from "altmetrics" as they relate
466 to preprints, and is there value in measuring a preprint's impact using methods rooted in
467 online interactions rather than citation count (Haustein 2018)? One study, published
468 before bioRxiv launched, found a significant association between Twitter mentions of
469 published papers and their citation counts (Thelwall et al. 2013)—have preprints changed
470 this dynamic?

471 Researchers have used existing resources and custom scripts to answer questions
472 like these. Himmelstein (2016b) found that only 17.8 percent of bioRxiv papers had an
473 "open license," for example, and another study examined the relationship between
474 Facebook "likes" of preprints and "traditional impact indicators" such as citation count, but
475 found no correlation for papers on bioRxiv (Ringelhan et al. 2015). Since most bioRxiv
476 data is not programmatically accessible, many of these studies had to begin by scraping
477 data from the bioRxiv website itself. The Rxivist API allows users to request the details of
478 any preprint or author on bioRxiv, and the database snapshots enable bulk querying of
479 preprints using SQL, C, and several other languages ("Procedural Languages" 2019) at a
480 level of complexity currently unavailable using the standard bioRxiv web interface. Using
481 these resources, researchers can now perform detailed and robust bibliometric analysis

482 of the website with the largest collection of preprints in biology, the one that, beginning in
483 September 2018, held more biology preprints than all other major preprint servers
484 combined (Anaya 2018).

485 In addition to our analysis here focused on big-picture trends related to bioRxiv,
486 the Rxivist website provides many additional features that may interest preprint readers
487 and authors. Its primary feature is sorting and filtering preprints based by download count
488 or mentions on Twitter, to help users find preprints in particular categories that are being
489 discussed either in the short term (Twitter) or over the span of months (downloads).
490 Tracking these metrics could also help authors gauge public reaction to their work: While
491 bioRxiv has compensated for a low rate of comments posted on the site itself (Inglis and
492 Sever 2016) by highlighting external sources such as tweets and blogs, Rxivist provides
493 additional context for how a preprint compares to others on similar topics. Several other
494 sites have attempted to use social interaction data to "rank" preprints, though none
495 incorporate bioRxiv download metrics. The "Assert" web application (https://assert.pub)
496 ranks preprints from multiple repositories based on data from Twitter and GitHub. The
497 "PromisingPreprints" Twitter bot (https://twitter.com/PromPreprint) accomplishes a similar
498 goal, posting links to bioRxiv preprints that receive an exceptionally high social media
499 attention score ("How Is the Altmetric Attention Score Calculated?" 2018) from Altmetric
500 (https://www.altmetric.com) in their first week on bioRxiv (De Coster 2017). Arxiv Sanity
501 Preserver (http://www.arxiv-sanity.com) provides rankings of arXiv.org preprints based on
502 Twitter activity, though its implementation of this scoring (Karpathy 2018) is more
503 opinionated than that of Rxivist. Other websites perform similar curation, but based on
504 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 visualization and description of the schema.) These

547 convert the entire Rxivist database into binary files that can be loaded by the free and
548 open-source PostgreSQL database management system to provide a local copy of all
549 collected data on every article and author on bioRxiv.org. (2) We also provide an API
550 (application programming interface) from which users can request information in JSON
551 format about individual preprints and authors, or search for preprints based on similar
552 criteria available on the Rxivist website. Complete documentation is available at
553 https://www.rxivist.org/docs .

554 While the analysis presented here deals mostly with overall trends on bioRxiv, the
555 primary entity of the Rxivist API is the individual research preprint, for which we have a
556 straightforward collection of metadata: title, abstract, DOI (digital object identifier), the
557 name of any journal that has also published the preprint (and its new DOI), and which
558 collection the preprint was submitted to. We also collected monthly traffic information for
559 each preprint, as reported by bioRxiv. We use the PDF download statistics to generate
560 rankings for each preprint, both site-wide and for each collection, over multiple timeframes
561 (all-time, year to date, etc.). In the API and its underlying database schema, "authors"
562 exist separately from "preprints" because an author can be associated with multiple
563 preprints. They are recorded with three main pieces of data: name, institutional affiliation
564 and a unique identifier issued by ORCID. Like preprints, authors are ranked based on the
565 cumulative downloads of all their preprints, and separately based on downloads within
566 individual bioRxiv collections. Emails are collected for each researcher, but are not
567 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

591 available there: the abstract, the date the preprint was first posted, and monthly download
592 statistics are pulled from here, as well as information about the preprint's authors—name,
593 email address and institution. These authors are then compared against the list of those
594 already indexed by Rxivist, and any unrecognized authors have profiles created in the
595 database.

596 **Consolidation of author identities.** Authors are most reliably identified across
597 multiple papers using the bioRxiv feature that allows authors to specify an identifier
598 provided by ORCID (https://orcid.org), a nonprofit that provides a voluntary system to
599 create unique identification numbers for individuals. These ORCID ("Open Researcher
600 and Contributor ID") numbers are intended to serve approximately the same role for
601 authors that DOI numbers do for papers (Haak 2012), providing a way to identify
602 individuals whose other information may change over time. 29,559 bioRxiv authors, or
603 17.4 percent, have an associated ORCID. If an individual included in a preprint's list of
604 authors doesn't have an ORCID already recorded in the database, authors are
605 consolidated if they have an identical name to an existing Rxivist author.

606 There are certainly authors who are duplicated within the Rxivist database, an
607 issue arising mostly from the common complaint of unreliable source data. 68.4 percent
608 of indexed authors have at least one email address associated with them, for example,
609 including 7,085 (4.40 percent) authors with more than one. However, of the 118,490 email
610 addresses in the Rxivist database, 6,517 (5.50 percent) are duplicates that are associated
611 with more than one author. Some of these are because real-life authors occasionally
612 appear under multiple names, but other duplicates are caused by uploaders to bioRxiv
613 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 posting date information. Posting dates were pulled from

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

655 **Calculation of download rankings.** The web crawler's "ranking" step orders
656 preprints and authors based on download count in two populations (overall and by bioRxiv
657 category) and over several periods: all-time, year-to-date, and since the beginning of the
658 previous month. The last metric was chosen over a "month-to-date" ranking to avoid
659 ordering papers based on the very limited traffic data available in the first days of each

660 month—in addition to a short lag in the time bioRxiv takes to report downloads, an
661 individual preprint's download metrics may only be updated in the Rxivist database once
662 every two or three weeks, so metrics for a single month will be biased in favor of those
663 that happen to have been crawled most recently. This effect is not eliminated in longer
664 windows, but is diminished. The step recording the rankings takes a more unusual
665 approach to loading the data: Because each article ranking step could require more than
666 37,000 "insert" or "update" statements, and each author ranking requires more than
667 170,000 of the same, these modifications are instead written to a text file on the application
668 server and loaded by running an instance of the Postgres command-line client "psql,"
669 which can use the more efficient "copy" command, a change that reduced the ranking
670 process from several hours to less than one minute.

671 **Reporting of small p-values:** In several locations, p-values are reported as
672 "< 2.2e-16". It is important to note that this is an inequality, and these p-values are not
673 necessarily identical. The upper limit, 2.2×10^{-16}, is not itself a particularly meaningful
674 number and is an artifact of the limitations of the floating-point arithmetic used by R, the
675 software used in the analysis. 2.2×10^{-16} is the "machine epsilon," or the smallest number
676 that can be added to 1.0 that would generate a result measurably different from 1.0.
677 Though smaller numbers can be represented by the system, those smaller than the
678 machine epsilon are not reported by default; we elected to do the same.

679 Data preparation

680 Several steps were taken to organize the data that was used for this paper. First,
681 the production data being used for the Rxivist API was copied to a separate "schema"—

682 a PostgreSQL term for a named set of tables. This was identical to the full database, but
683 had a specifically circumscribed set of preprints. Once this was copied, the table
684 containing the associations between authors and each of their papers ("article_authors")
685 was pruned to remove references to any articles that were posted after 30 Nov 2018, and
686 any articles that were not associated with a bioRxiv collection. For unknown reasons, 10
687 preprints (0.03%) could not be associated with a bioRxiv collection; because the bioRxiv
688 profile page for some papers does not specify which collection it belongs to, these papers
689 were ignored. Once these associations were removed, any articles meeting those criteria
690 were removed from the "articles" table. References to these articles were also removed
691 from the table containing monthly bioRxiv download metrics for each paper
692 ("article_traffic"). We also removed all entries from the "article_traffic" table that recorded
693 downloads after November 2018. Next, the table containing author email addresses
694 ("author_emails") was pruned to remove emails associated with any author that had zero
695 preprints in the new set of papers; those authors were then removed from the "authors"
696 table.

697 Before evaluating data from the table linking published preprints to journals and
698 their post-publication DOI ("article_publications"), journal names were consolidated to
699 avoid under-counting journals with spelling inconsistencies. First, capitalization was
700 stripped from all journal titles, and inconsistent articles ("The Journal of…" vs. "Journal
701 of…"; "and" vs. "&" and so on) were removed. Then, the list of journals was reviewed by
702 hand to remove duplication more difficult to capture automatically: "PNAS" and
703 "Proceedings of the National Academy of Sciences," for example. Misspellings were rare,
704 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 publication data did not include any instances of papers appearing in those journals ("List of Predatory Journals" 2018). It is important to note that the absence of this information does not necessarily indicate that preprints have not appeared in these journals—we performed this search to ensure our analysis of publication rates was not inflated with numbers from illegitimate publications.

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.

Estimation of ranges for true publication rates. To evaluate the sensitivity of the system bioRxiv uses to detect published versions of preprints, we pulled a random sample of 120 preprints that had not been marked as published on bioRxiv.org—30 preprints from each year between 2014 and 2017. We then performed a manual online literature search for each paper to determine whether they had actually been published. The primary search method was searching on Google.com for the preprint's title and the

749 senior author's last name—if this did not return any results that looked like publications,
750 other author names were added to the search to replace the senior author's name. If this
751 did not return any positive results, we also checked Google Scholar
752 (https://scholar.google.com) for papers with similar titles. If any of the preprint's authors,
753 particularly the first and last authors, had Google Scholar profiles, they were reviewed for
754 publications on subject matter similar to the preprint. If a publication looked similar to the
755 preprint, a visual comparison between the preprint and published paper's abstract and
756 introduction was used to determine if they were simply different version of the same paper.
757 The paper was marked as a true negative If none of these returned positive results, or if
758 the suspected published paper described a study that was different enough that the
759 preprint effectively described a different research project.

760 Once all 120 preprints had been evaluated, the results were used to approximate
761 a false-negative rate to each year—the proportion of preprints that had been incorrectly
762 excluded from the list of published papers. The sample size for each year (30) was used
763 to calculate the margin of error using a 95% confidence interval (17.89 percentage points).
764 This margin was then used to generate the minimum and maximum false-negative rates
765 for each year, which were then used to calculate the minimum and maximum number of
766 incorrectly classified preprints from each year. These numbers yielded a range for each
767 year's actual publication rate: For 2015, for example, bioRxiv identified 1,218 preprints (out
768 of 1,774) that had been published. The false-negative rate and MOE suggest between 197
769 and 396 additional preprints have been published but not detected, yielding a final range
770 of 1,415–1,614 preprints published in that year.

To evaluate the specificity of the publication detection system, we pulled 40 samples (10 from each of the years listed above) that bioRxiv had listed as published, and found that all 40 had been accurately classified. Though this helps establish that bioRxiv is not consistently finding all preprint publications, it should be noted that the determination of a more precise estimation for publication rates would require deeper analysis and sampling.

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.

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

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

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

"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

Garfield, Eugene. 2006. “The History and Meaning of the Journal Impact Factor,” JAMA 295(1) pp. 90–93. doi: 10.1001/jama.295.1.90

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

"The Impact Factor Game," 2006. PLOS Medicine 3(6): e291. doi: 10.1371/journal.pmed.0030291

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

Kramer, Bianca. 2019. "Rxivist Analysis, " Google Docs. https://docs.google.com/spreadsheets/d/18-zIlfgrQaGo6e4SmyfzMTY7AN1dUYiI5l6PyX5pWtg/edit#gid=1455314569 (accessed 15 Mar 2019).

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/

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

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

"Web of Science." 2018. Clarivate Analytics.

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

Supplements

Figure supplements

Source	Articles
Cell vol. 174(6)	17
Cell vol. 175(1)	18
Genetics vol. 210(1)	23
Jour of Biochem vol. 164(3)	8
PLoS Biology vol. 16(9)	19
bioRxiv, 1–3 Sep 2018	100

Figure 1—figure supplement 1: The number of full-length articles published by an arbitrary selection of well-known journals in September 2018. The *Cell* count is limited to the "Articles" and "Resources" categories; *Genetics* is limited to their "Investigations" category, and *PLoS Biology* to "Research Articles," "Methods and Resources" and "Meta-Research Articles." Links to each issue's table of contents are included in supplementary file *figures.md*.

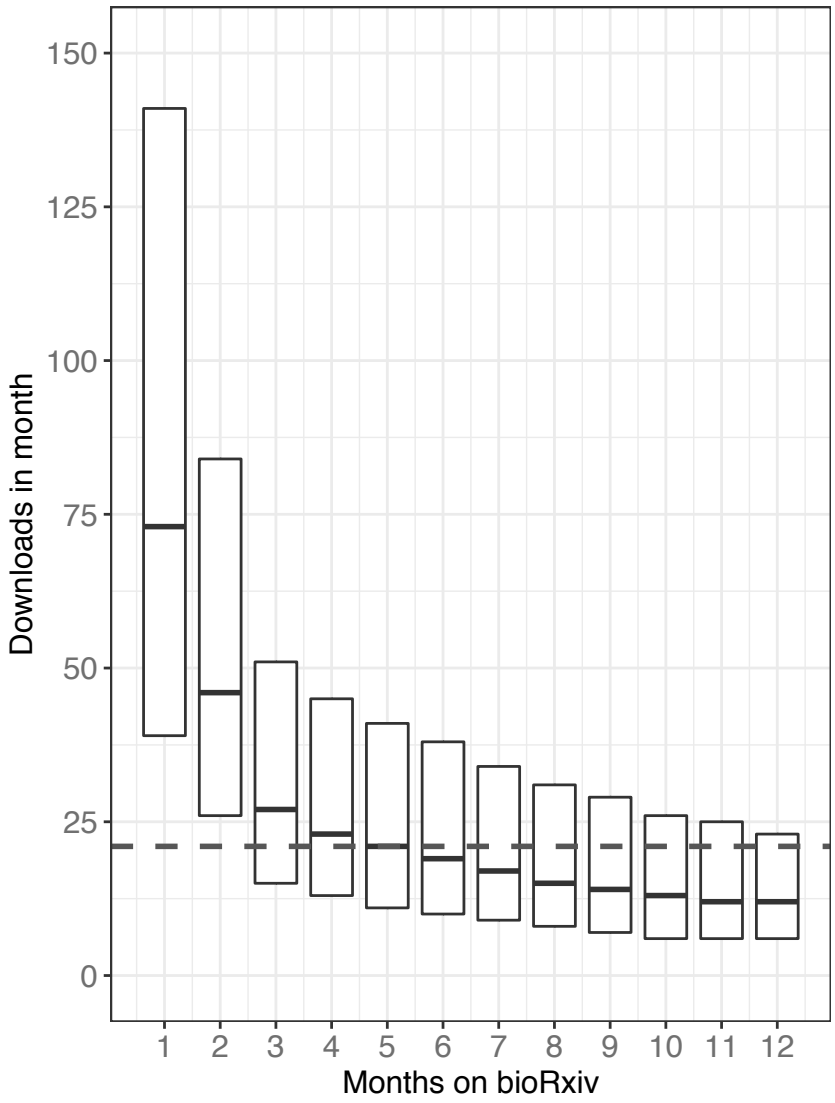

Figure 2—figure supplement 1. The distribution of downloads that preprints accrue in their first months on bioRxiv. For example, the box at "1" on the x axis indicates the downloads that all preprints have received in their first month online. The dashed line represents the median downloads per month over each paper's first 12 months.

Figure 2—figure supplement 1—source data 1: A list of every preprint posted before 2018, the category to which it was posted, and the number of downloads it received in each of its first 12 months.

downloads_by_months.csv

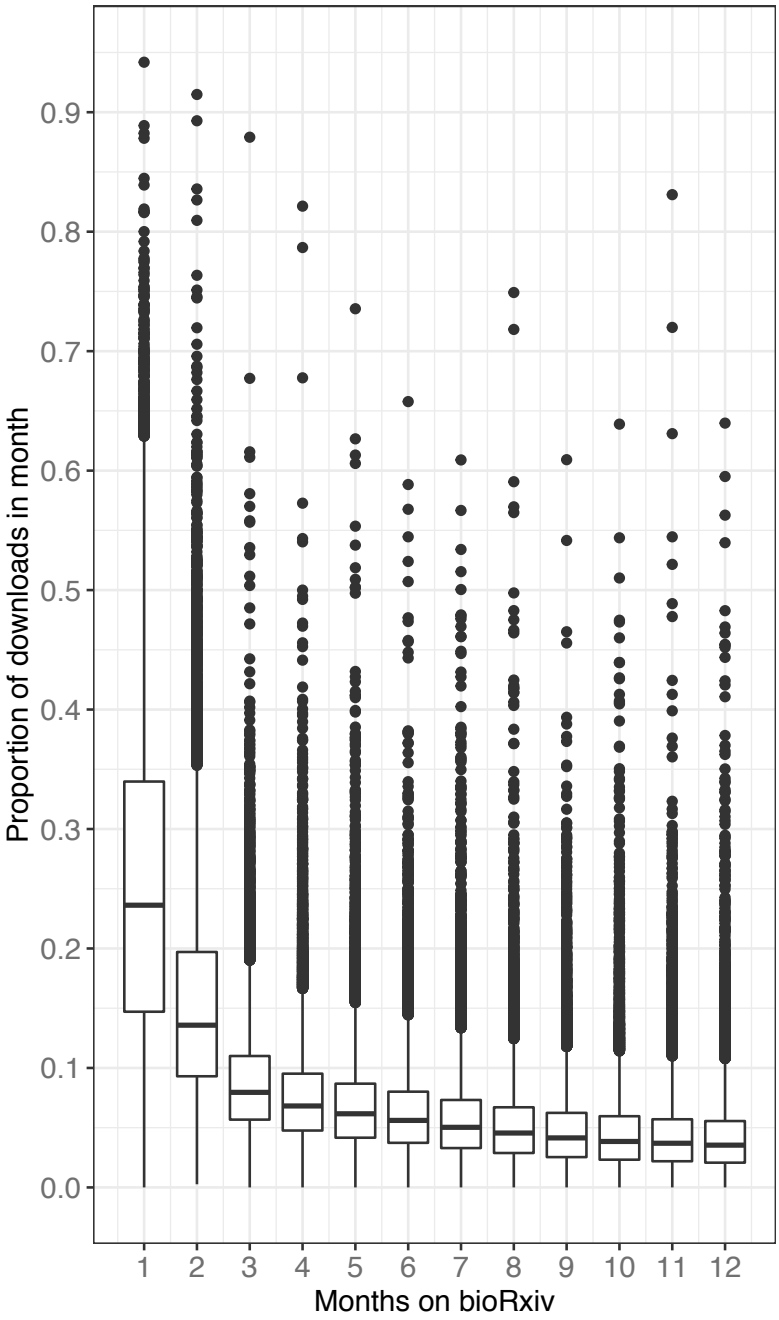

Figure 2—figure supplement 2. The proportion of downloads that preprints accrue in their first months on bioRxiv.

Figure 2—figure supplement 2—source data 1: See Figure 2—figure

supplement 1—source data 1.

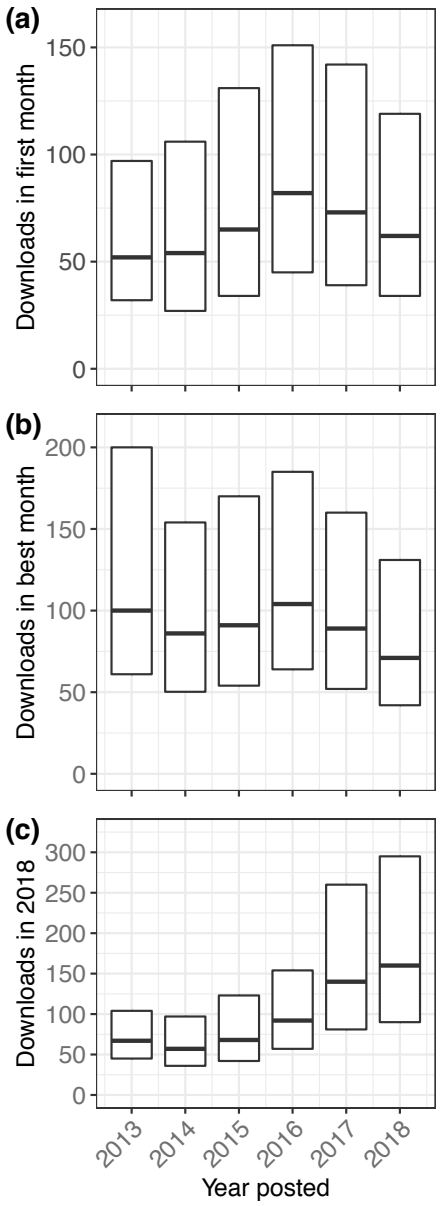

Figure 2—figure supplement 3. Multiple perspectives on per-preprint download statistics. Preprints in all three plots are categorized by the year in which they were first posted. **(a)** The median number of downloads for each preprint's first calendar month on bioRxiv. (There was no correction

done for preprints posted toward the end of their first month.)**(b)** The median number of downloads in each preprint's *best* month, in any year. **(c)** The median downloads per preprint in 2018, for preprints posted in any year.

Figure 2—figure supplement 3—source data 1: A list of every preprint, the month and year in which it was posted, and the number of downloads it received in its first month. *downloads_by_first_month.csv*

Figure 2—figure supplement 3—source data 2: A list of every preprint, the year in which it was posted, and the maximum number of downloads it received in a single month. *downloads_max_by_year_posted.csv*

Figure 2—figure supplement 3—source data 3: A list of every preprint, the year in which it was posted, and the number of downloads it received in the first 11 months of 2018. *2018_downloads_per_year.csv*

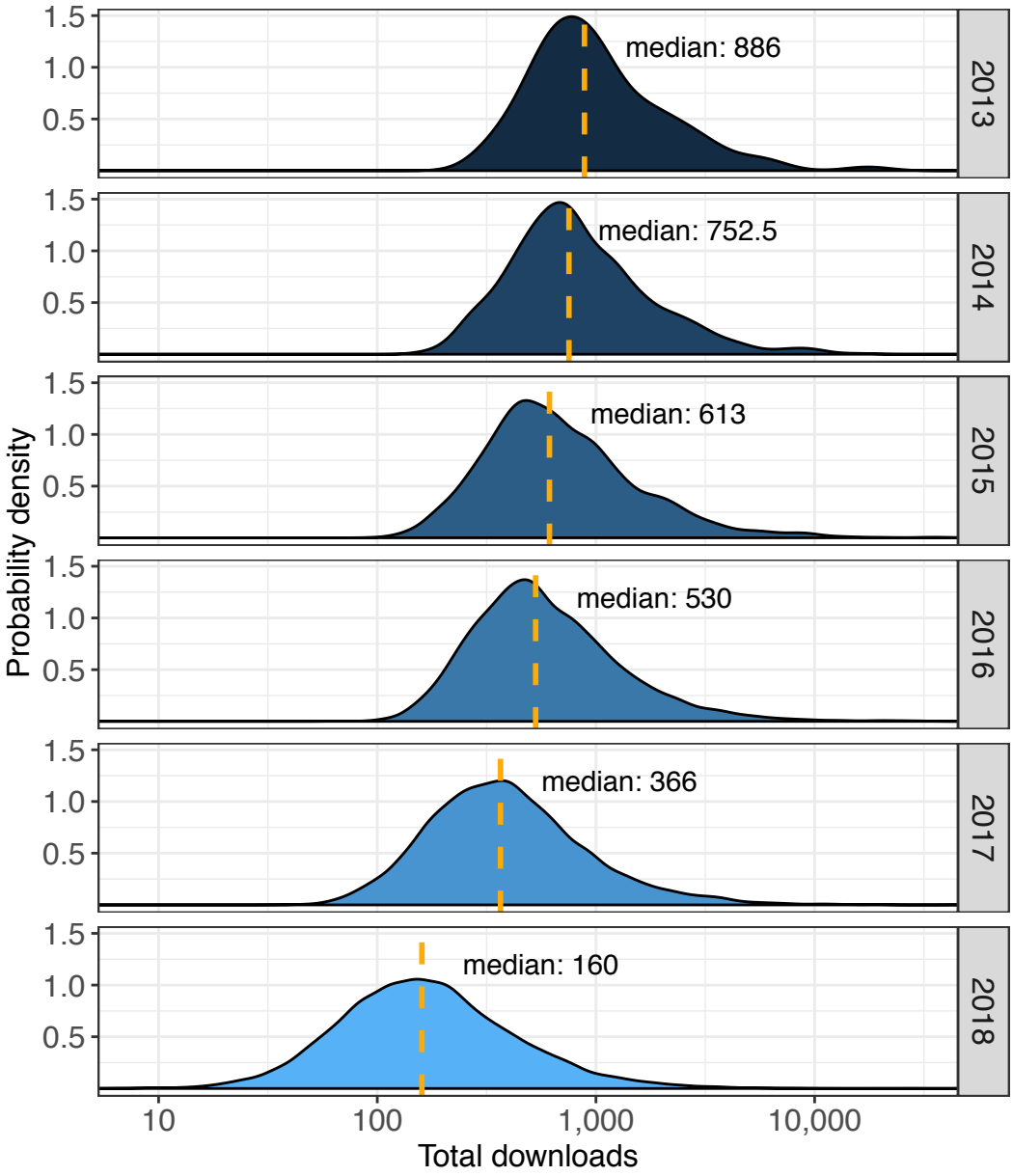

Figure 2—figure supplement 4. Total downloads per preprint, segmented by the year in which each preprint was posted. The Y axis indicates a probability density rather than overall count. The dashed orange lines in each plot represent the median downloads (labeled separately in each plot) for preprints posted in that year.

1046 **Figure 2—figure supplement 4—source data 1:** A list of all preprints, the

1047 year in which it was posted, and its total download count.

1048 *downloads_per_year.csv*

	2018			2014 through 2018		
Journal	**Total**	**Preprints**	**Proportion**	**Total**	**Preprints**	**Proportion**
GigaScience	89	44	49.44%	375	90	24.00%
Genome Biology	183	73	39.89%	1,145	181	15.81%
Genome Research	169	62	36.69%	1,020	174	17.06%
eLife	1172	394	33.62%	6,118	750	12.26%
Nature Methods	137	46	33.58%	847	119	14.05%
PLOS Computational Bio.	490	153	31.22%	3,310	337	10.18%
Nature Genetics	188	55	29.26%	1,169	125	10.69%
G3	344	95	27.62%	1,962	246	12.54%
Genetics	276	71	25.72%	1,899	269	14.17%
Bioinformatics	829	209	25.21%	4,694	472	10.06%
PLOS Genetics	515	117	22.72%	3,885	292	7.52%
Molecular Bio. and Evolution	228	47	20.61%	1,544	157	10.17%
PLOS Biology	333	67	20.12%	1,379	108	7.83%
Genome Bio. and Evolution	199	40	20.10%	1,571	109	6.94%
Molecular Bio. of the Cell*	215	38	17.67%	2,012	88	4.37%
mBio	373	64	17.16%	2,383	108	4.53%
NeuroImage	862	117	13.57%	5,186	214	4.13%
Biophysical Journal*	482	65	13.49%	3,103	109	3.51%

BMC Bioinformatics	456	61	13.38%	3,176	137	4.31%
Journal of Neuroscience	777	98	12.61%	6,569	172	2.62%
Development	360	38	10.56%	2,296	90	3.92%
Molecular Ecology	315	32	10.16%	2,245	92	4.10%
Nucleic Acids Research	1142	115	10.07%	7,896	265	3.36%
BMC Genomics	904	75	8.30%	6,366	174	2.73%
Nature Communications	4979	355	7.13%	22,338	529	2.37%
PNAS	3195	212	6.64%	19,697	432	2.19%
Cell Reports	1230	79	6.42%	6,846	136	1.99%
PeerJ	1877	53	2.82%	7,047	120	1.70%
Scientific Reports	16899	398	2.36%	9,7451	827	0.85%
PLOS ONE	17021	395	2.32%	138,009	741	0.54%

Figure 4—figure supplement 1. The total number of articles published by the top 30 journals that have published the most bioRxiv preprints, compared to how many of those articles appeared on bioRxiv. The list is ordered by the proportion of 2018 publications first appeared on bioRxiv. Only 2018 and the total number are displayed here; annual data from 2014 through November 2018 is available in the source data. Total article counts are from the number of works in the “article” category as indexed by Web of Science (Clarivate Analytics). Journals marked with an asterisk have a large number of published works categorized on Web of Science as “meeting abstracts”; for consistency, those are not included in the counts

here, though it is possible some of the preprints published by these journals fall into that category.

Figure 4—figure supplement 4—source data 1: An annual count of how many articles were published by the journals publishing the most bioRxiv preprints, coupled with the number of preprints published in that year. *preprint_proportions_per_journal.xlsx*

Author name	bioRxiv preprints	Primary field	Email addresses
George Davey Smith	97	Epidemiology	4
Ian J. Deary	61	Genetics	4
Andrew M. McIntosh	57	Genetics	1
Mark J. Daly	47	Genetics	3
Richard M. Murray	45	Synthetic biology	2
George M. Church	43	Synthetic biology	4
Wei Wang	39	Bioinformatics	23
Benjamin M. Neale	39	Genetics	2
Alkes L. Price	36	Genetics	1
Jian Yang	36	Genetics	6
Po-Ru Loh	35	Genetics	2
Caroline Hayward	35	Genetics	1
Aarno Palotie	34	Genetics	4
Jay Shendure	34	Genomics	3
Ole A. Andreassen	34	Genetics	2

Table 1—table supplement 1: The top 15 authors with the most preprints on bioRxiv. Names are listed as they appear on biorxiv.org, after making corrections outlined in the Methods section. An author's "Primary field" is the bioRxiv collection to which they have submitted the most preprints. Preprint count does not account for duplicates: For example, Ian J. Deary and Andrew M. McIntosh are both high on the list, but their counts include multiple preprints that they co-authored together. The "Email addresses" field lists the number of email addresses observed in that author's preprints that is attributed to them, and is used to approximate the risk that the author is actually a conglomeration of multiple researchers with the same name.

Table 1—table supplement 1—source data 1: A list of every author, the number of preprints for which they are listed as an author, and the number of email addresses they are associated with. *papers_per_author.csv*

Institution	**Authors**	**Preprints**
Stanford University	1,473	1,045
University of Oxford	1,192	902
University of Cambridge	1,109	842
University of Washington	924	609
University College London	801	644
University of Pennsylvania	764	544

University of Michigan	763	484
University of California, San Francisco	750	511
University of California, San Diego	725	495
Imperial College London	703	472
University of Edinburgh	646	487
University of California, Berkeley	620	528
Yale University	555	392
Duke University	554	323
Harvard University	532	557
Harvard Medical School	529	453
Columbia University	520	422
Cornell University	486	365
University of Toronto	462	334
Johns Hopkins University	461	407
University of California, Davis	461	291
Icahn School of Medicine at Mount Sinai	448	281
University of Chicago	444	353

University of British Columbia	431	281
University of Minnesota	430	310

Table 1—table supplement 2. Top 25 institutions with the most authors listing them as their affiliation, and how many papers have been published by those authors. Each institution's count of total preprints is based on the number of papers posted by authors currently listed with those affiliations, but preprints attributed to authors from multiple institutions count toward the total for all institutions mentioned. A paper with multiple authors from the same institution is counted only once for that institution.

Table 1—table supplement 2—source data 1: A list of every indexed institution, the number of authors associated with that institution, and the number of papers authored by those researchers. *authors_per_institution.csv*

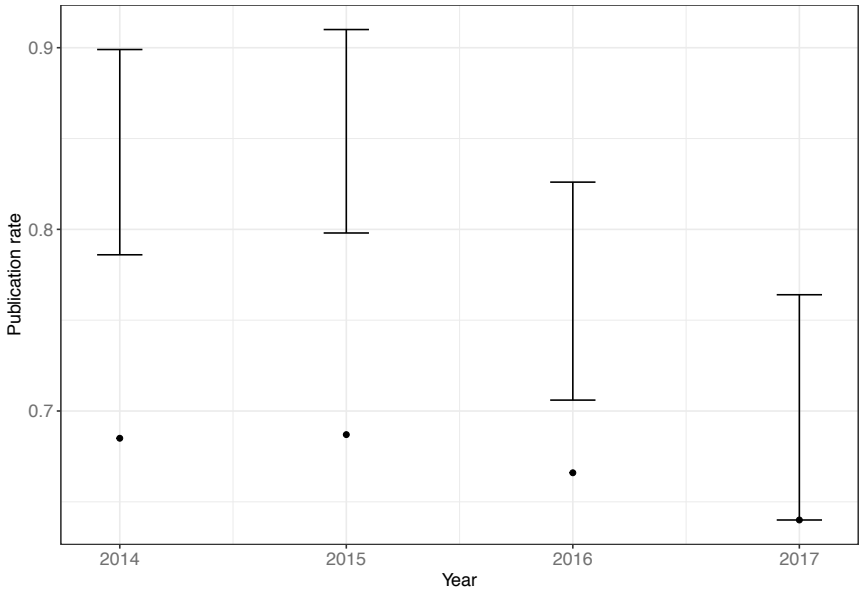

Figure 3—figure supplement 1. The publication rate (y-axis) of preprints posted in four consecutive years (x-axis). The points show the proportion of preprints posted in each year that bioRxiv has associated with a journal publication. The error bars indicate the range for an estimated publication rate based on our manual examination of 120 preprints (30 from each year) that were misclassified by bioRxiv as unpublished. The ranges are calculated by combining the observed rate of publication in the "unpublished" samples with the margin of error at a 95% confidence interval.

Figure 3—figure supplement 1—source data 1: An Excel workbook that includes the results of the manual trials, plus a sheet calculating the revised publication rates for each year. *biorxiv_linking.xlsx*

List of supplements and source data files

- **Supplementary file 1:** Database schema. A description of each field in each table of the database used to store the scraped bioRxiv data.
- **Source code file 1:** Figure generation code. The R and SQL code used to generate all figures in this paper.
- **Table 1**
    - **Table supplement 1:** The top 15 authors with the most preprints on bioRxiv.

1114 ■ **Source data 1:** A list of every author, the number of preprints for
1115 which they are listed as an author, and the number of email
1116 addresses they are associated with. *papers_per_author.csv*
1117 ○ **Table supplement 2.** Top 25 institutions with the most authors listing them
1118 as their affiliation, and how many papers have been published by those
1119 authors.
1120 ■ **Source data 1:** A list of every indexed institution, the number of
1121 authors associated with that institution, and the number of papers
1122 authored by those researchers. *authors_per_institution.csv*
1123 ● **Figure 1**
1124 ○ **Figure supplement 1:** The number of full-length articles published by an
1125 arbitrary selection of well-known journals in September 2018.
1126 ○ **Source data 1:** The number of submissions per month to each bioRxiv
1127 category, plus running totals. *submissions_per_month.csv*
1128 ○ **Source data 2:** An Excel workbook demonstrating the formulas used to
1129 calculate the running totals in Figure 1—source data 1.
1130 *submissions_per_month_cumulative.xlsx*
1131 ○ **Source data 3:** The number of submissions per month overall, plus running
1132 totals. *submissions_per_month_overall.csv*
1133 ● **Figure 2**
1134 ○ **Figure supplement 1:** The distribution of downloads that preprints accrue
1135 in their first months on bioRxiv.
1136 ■ **Source data 1:** A list of every preprint posted before 2018, the

category to which it was posted, and the number of downloads it received in each of its first 12 months. *downloads_by_months.csv*

- **Figure supplement 2.** The proportion of downloads that preprints accrue in their first months on bioRxiv.
- **Figure supplement 3.** Multiple perspectives on per-preprint download statistics.
  - **Source data 1:** A list of every preprint, the month and year in which it was posted, and the number of downloads it received in its first month. *downloads_by_first_month.csv*
  - **Source data 2:** A list of every preprint, the year in which it was posted, and the maximum number of downloads it received in a single month. *downloads_max_by_year_posted.csv*
  - **Source data 3:** A list of every preprint, the year in which it was posted, and the number of downloads it received in the first 11 months of 2018. *2018_downloads_per_year.csv*
- **Figure supplement 4.** Total downloads per preprint, segmented by the year in which each preprint was posted.
  - **Source data 1:** A list of all preprints, the year in which it was posted, and its total download count. *downloads_per_year.csv*
- **Source data 1:** A list of every preprint, its bioRxiv category, and its total downloads. *downloads_per_category.csv*
- **Source data 2:** The number of downloads per month in each bioRxiv category, plus running totals. *downloads_per_month_cumulative.csv*

- **Source data 3:** An Excel workbook demonstrating the formulas used to calculate the running totals in Figure 2—source data 2. *downloads_per_month.xlsx*
- **Source data 4:** The number of downloads per month overall, plus running totals. *downloads_per_month_per_year.csv*

- **Figure 3**
  - **Source data 1:** The number of preprints posted in each month, plus the count and proportion of those later published. *publication_rate_month.csv*
  - **Source data 2:** The number of preprints posted in each category, plus the count and proportion of those published. *publications_per_category.csv*
  - **Figure supplement 1:** Observed annual publication rates and estimated range for actual publication rates.
    - **Source data 1:** An Excel workbook that includes the results of the manual trials, plus a sheet calculating the revised publication rates for each year. *biorxiv_linking.xlsx*
  -

- **Figure 4**
  - **Source data 1:** The number of preprints published in each category by the 30 most prolific publishers of preprints. *publications_per_journal_categorical.csv*
  - **Figure supplement 1:** The total number of articles published by the top 30 journals that have published the most bioRxiv preprints, compared to how many of those articles appeared on bioRxiv.

1183 ■ **Source data 1:** An annual count of how many articles were
1184 published by the journals publishing the most bioRxiv preprints,
1185 coupled with the number of preprints published in that year.
1186 *preprint_proportions_per_journal.xlsx*

1187 - **Figure 5**
1188 ○ **Source data 1:** A list of every preprint with its total download count and the
1189 journal in which it was published, if any. *downloads_journal.csv*
1190 ○ **Source data 2:** Journal impact factor and access status of the 30 journals
1191 that have published the most preprints. *impact_scores.csv*
1192 - **Table 2**
1193 ○ **Source data 1:** A list of every preprint with its total download count, the year
1194 in which it was first posted, and whether it has been published.
1195 *downloads_publication_status.csv*
1196 - **Figure 6**
1197 ○ **Source data 1:** A list of every published preprint, the year it was first posted,
1198 the date it was published, and the interval between posting and publication,
1199 in days. *publication_time_by_year.csv*
1200 ○ **Source data 2:** A list of every preprint published in the 30 journals displayed
1201 in the figure, the journal in which it was published, and the interval between
1202 posting and publication, in days. *publication_interval_journals.csv*
1203 ○ **Source data 3:** The results of Dunn's test, a pairwise comparison of the
1204 median publication interval of each journal in the figure.
1205 *journal_interval_dunnstest.txt*