

1 **Tracking the popularity and outcomes of all bioRxiv preprints**

2 Richard J. Abdill¹ and Ran Blekhman^{1,2}

3 1 – Department of Genetics, Cell Biology, and Development, University of Minnesota,
4 Minneapolis, MN

5 2 – Department of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul,
6 MN

7

8 **ORCID iDs**

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

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

11

12 **Correspondence**

13 Ran Blekhman, PhD

14 University of Minnesota

15 MCB 6-126

16 420 Washington Avenue SE

17 Minneapolis, MN 55455

18 Email: blekhman@umn.edu

19 Abstract

20 Researchers in the life sciences are posting their work to preprint servers at an
21 unprecedented and increasing rate, sharing papers online before (or instead of)
22 publication in peer-reviewed journals. Though the popularity and practical benefits of
23 preprints are driving policy changes at journals and funding organizations, there is little
24 bibliometric data available to measure trends in their usage. Here, we collected and
25 analyzed data on all 37,648 preprints that were uploaded to bioRxiv.org, the largest
26 biology-focused preprint server, in its first five years. We find that preprints on bioRxiv
27 are being read more than ever before (1.1 million downloads in October 2018 alone)
28 and that the rate of preprints being posted has increased to a recent high of more than
29 2,100 per month. We also find that two-thirds of bioRxiv preprints posted in 2016 or
30 earlier were later published in peer-reviewed journals, and that the majority of published
31 preprints appeared in a journal less than six months after being posted. We evaluate
32 which journals have published the most preprints, and find that preprints with more
33 downloads are likely to be published in journals with a higher impact factor. Lastly, we
34 developed Rxivist.org, a website for downloading and interacting programmatically with
35 indexed metadata on bioRxiv preprints.

36 Introduction

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

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

54 Against this backdrop, preprints have become a steady source of the most recent
55 research in biology, providing a valuable way to learn about exciting, relevant and high-
56 impact findings—for free—months or years before that research will appear anywhere
57 else, if at all (Kaiser 2017). It's a practice long familiar to physicists, who began
58 submitting preprints to arXiv, one of the earliest preprint servers, in 1991 (Verma 2017).
59 Researchers in fields supported by that server "have developed a habit of checking
60 arXiv every morning to learn about the latest work in their field" (Vale and Hyman 2016),
61 and one survey of published mathematicians found that 81 percent had posted at least
62 one preprint to the site (Fowler 2011). In the life sciences, however, researchers
63 approached preprints with reluctance (O'Roak 2018), even when major publishers made

64 it clear they were not opposed to the practice ("Nature respects preprint servers" 2005;
65 Desjardins-Proulx et al. 2013). An early NIH plan for PubMed Central called "E-Biomed"
66 included the hosting of preprints (Varmus 1999; Smaglik 1999) but was scuttled by the
67 National Academy of Sciences, which successfully negotiated the exclusion of work that
68 had not been peer-reviewed (Marshall 1999; Kling et al. 2003).

69 Further attempts to circulate biology preprints, such as NetPrints (Delamothe et
70 al. 1999), Nature Precedings (Kaiser 2017), and The Lancet Electronic Research
71 Archive (McConnell and Horton 1999), popped up (and then folded) over time ("ERA
72 Home" 2019). The one that would catch on, bioRxiv, wasn't founded until 14 years after
73 the fall of E-Biomed (Callaway 2013). Now, biology publishers are actively trawling
74 preprint servers for submissions (Barsh et al. 2016; Vence 2017), and more than 100
75 journals accept submissions directly from the bioRxiv website ("Submission Guide"
76 2018). The National Institutes of Health announced the explicit acceptance of preprint
77 citations in grant proposals ("Reporting Preprints and Other Interim Research Products"
78 2017), and multiple funding opportunities from the multi-billion-dollar Chan Zuckerberg
79 Initiative (Abutaleb 2015) require all publications to first be posted to a preprint server
80 ("Funding Opportunities" 2018; Champieux 2018). The conventions of the biology
81 publishing game are changing, in ways that reflect a strong influence from the
82 expanding popularity of preprints. However, details about that ecosystem are hard to
83 come by. We know bioRxiv is the largest of the biology-focused preprint servers: Of the
84 eight websites indexed by PrePubMed (<http://www.prepubmed.org>), bioRxiv now
85 consistently posts more than three times as many articles per month as the other seven
86 combined (Anaya 2018). Sporadic updates from bioRxiv leaders show a chain of

87 record-breaking months for submission numbers (Sever 2018), and analyses have
88 examined metrics such as total downloads (Serghiou and Ioannidis 2018) and
89 publication rate (Schloss 2017). But long-term questions remain open: Which fields
90 have posted the most preprints, and which collections are growing most quickly? How
91 many times have preprints been downloaded, and which categories are most popular
92 with readers? How many preprints are eventually published elsewhere, and in what
93 journals? Is there a relationship between a preprint's popularity and the journal in which
94 it later appears? Do these conclusions change over time?

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

103 Results

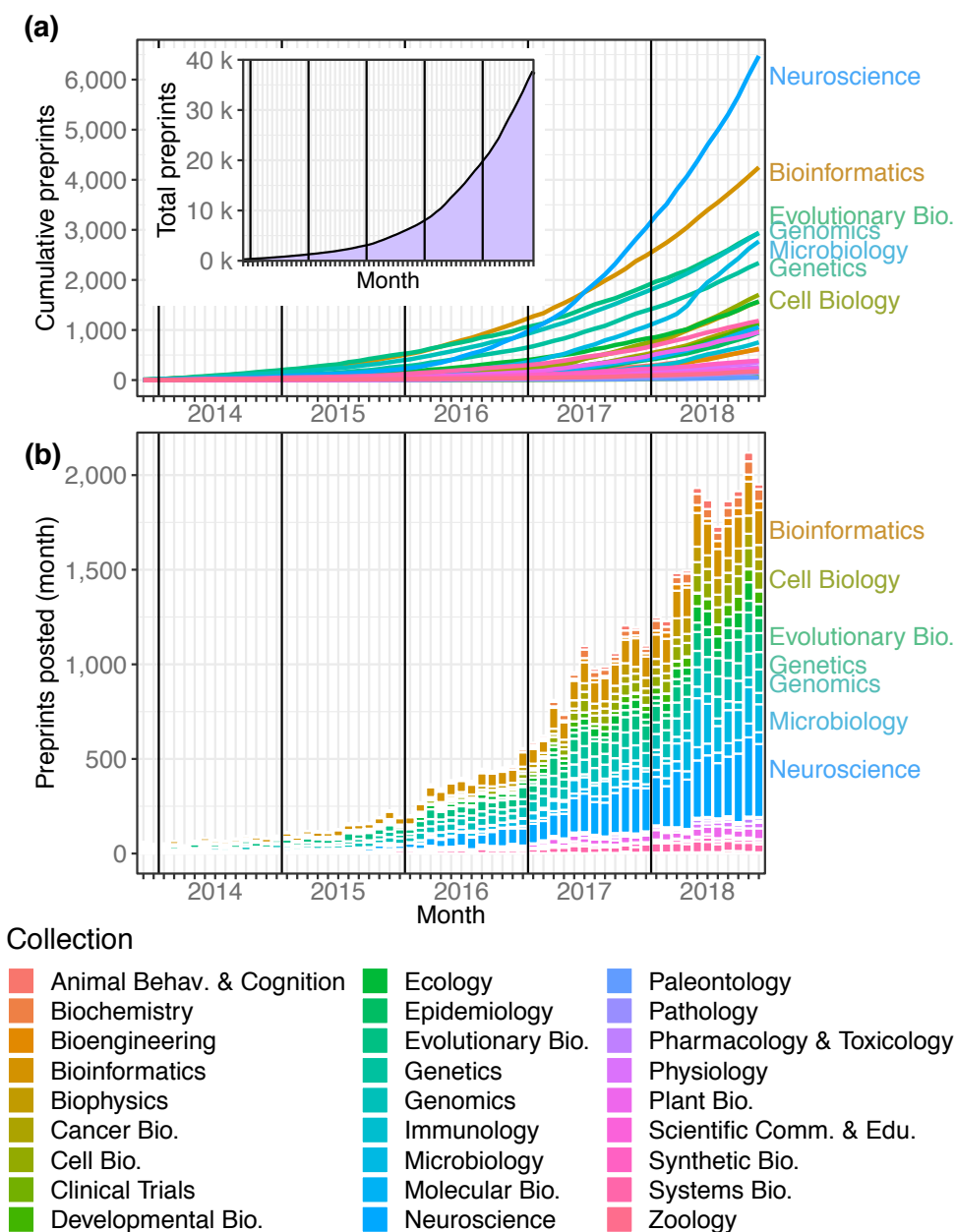
104 We developed a Python-based web crawler to visit every content page on the
105 bioRxiv website and download basic data about each preprint across the site's 27
106 subject-specific categories: title, authors, download statistics, submission date,
107 category, DOI, and abstract. The bioRxiv website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

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

119 Preprint submissions

120 The most apparent trend that can be pulled from the bioRxiv data is that the
121 website is extraordinarily popular with authors, and becoming more so every day: There
122 were 37,648 preprints available on bioRxiv at the end of November 2018, and more
123 preprints were posted in the first 11 months of 2018 (18,825) than in all four previous
124 years combined (**Figure 1a**). The number of bioRxiv preprints doubled in less than a
125 year, and new submissions have been trending upward for five years (**Figure 1b**). The
126 plurality of site-wide growth can be attributed to the neuroscience collection, which has
127 had more submissions than any bioRxiv category in every month since September 2016
128 (**Figure 1b**). In October 2018, it became the first of bioRxiv's collections to contain
129 6,000 preprints (**Figure 1a**). The second-largest category is bioinformatics (4,249
130 preprints), followed by evolutionary biology (2,934). October 2018 was also the first

131 month in which bioRxiv posted more than 2,000 preprints, increasing its total preprint
 132 count by 6.3 percent (2,119) in 31 days.



133
 134 **Figure 1.** Total preprints posted to bioRxiv over a 61-month period from
 135 November 2013 through November 2018. **(a)** The number of preprints (y-
 136 axis) at each month (x-axis), with each category depicted as a line in a

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

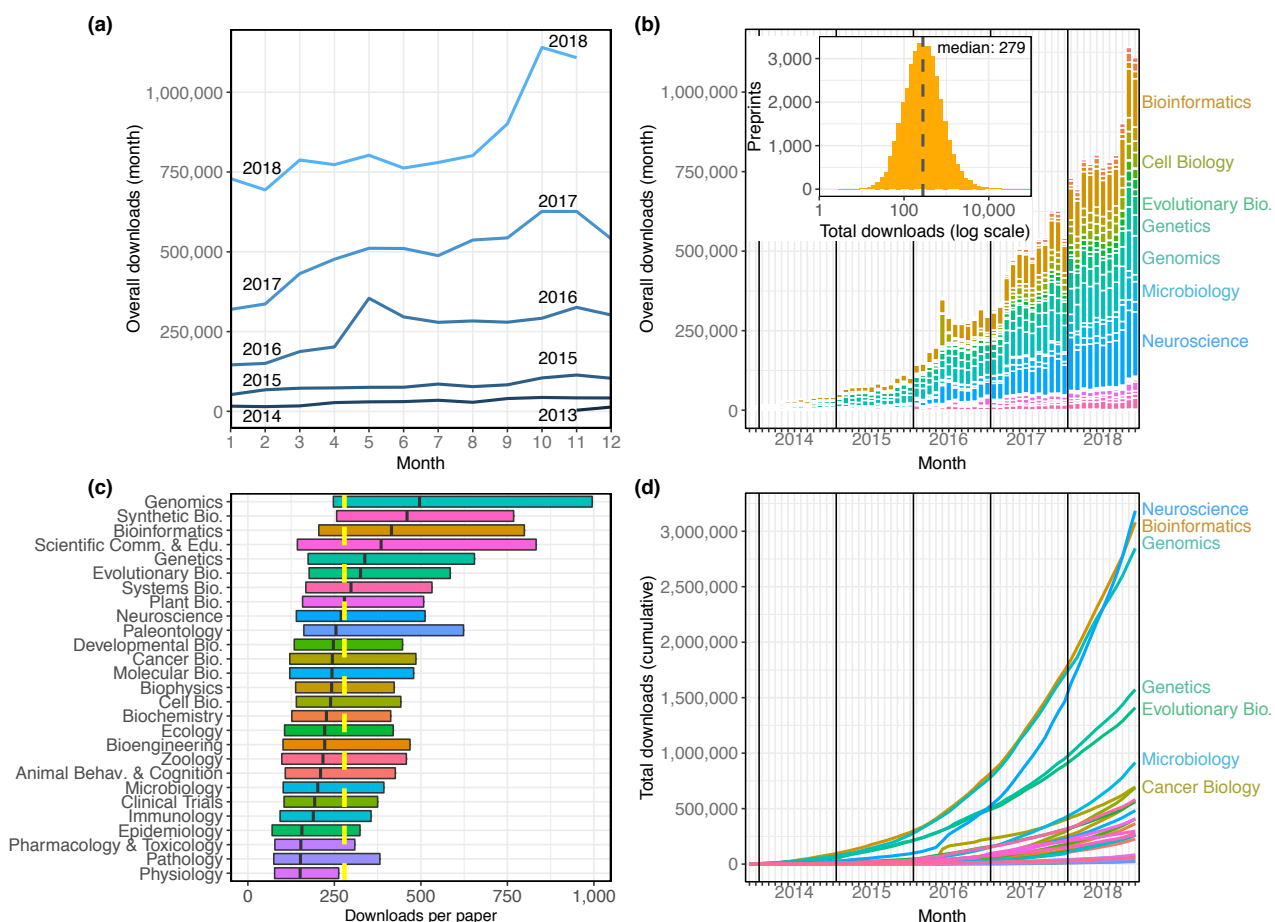
140 **Supplementary files:** *submissions_per_month.csv*,
141 *submissions_per_month_overall.csv*

142 Preprint downloads

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

158 category is 269.5, while the median of preprints in all other categories is 281 (**Figure**
159 **2c**).

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



171

172

173 **Figure 2.** The distribution of all recorded downloads of bioRxiv preprints.

174 **(a)** The downloads recorded in each month, with each line representing a

175 different year. The lines reflect the same totals as the height of the bars in

176 Figure 2b. **(b)** A stacked bar plot of the downloads in each month: The

177 height of each bar indicates the total downloads in that month. Each

178 stacked bar shows the number of downloads in that month attributable to

179 each category; the colors of the bars are described in the legend in Figure

180 1. **(b, inset)** A histogram showing the site-wide distribution of downloads

181 per preprint, as of the end of November 2018. The median download

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

190 **Supplementary files:** *downloads_per_category.csv*,
191 *downloads_per_month_cumulative.csv*,
192 *downloads_per_month_per_year.csv*

193 Preprint authors

194 While data about the authors of individual preprints is easy to organize,
195 associating authors between preprints is difficult due to a lack of consistent unique
196 identifiers (see Methods). We chose to define an author as a unique name in the author
197 list, including middle initials but disregarding letter case and punctuation. Keeping this in
198 mind, we find that there are 170,287 individual authors with content on bioRxiv. Of
199 these, 106,231 (62.4%) posted a preprint in 2018, including 84,339 who posted a
200 preprint for the first time (**Table 1**), indicating that total authors increased by more than
201 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

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

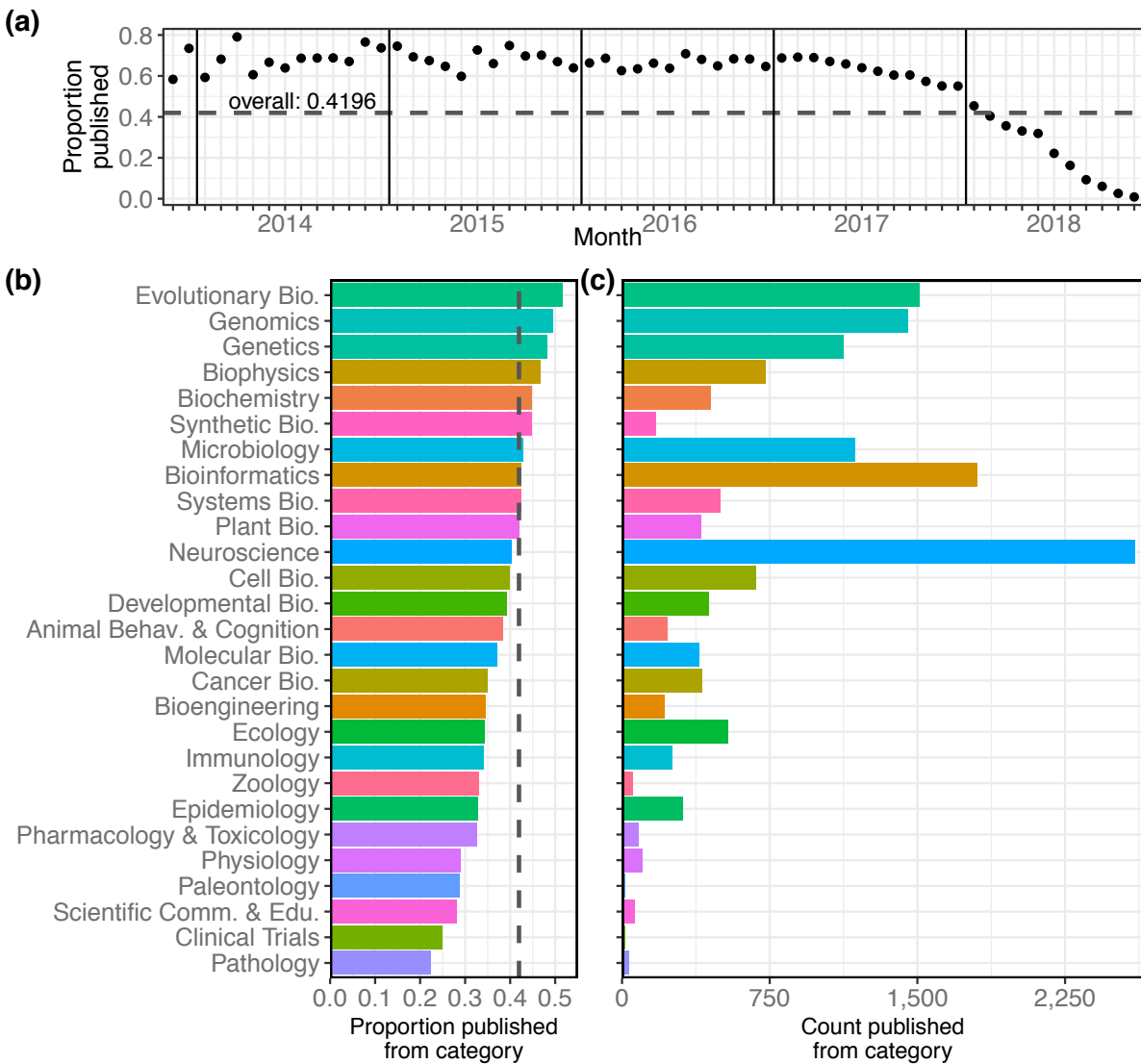
208 Even though 129,419 authors (76.0%) are associated with only a single preprint,
209 the mean preprints per author is 1.52 because of a skewed rate of contributions also
210 found in conventional publishing (Rørstad and Aksnes 2015): 10 percent of authors
211 account for 72.8 percent of all preprints, and the most prolific researcher on bioRxiv,
212 George Davey Smith, is listed on 97 preprints across seven categories (**Table S2**).
213 1,473 authors list their most recent affiliation as Stanford University, the most
214 represented institution on bioRxiv (**Table S3**). Though the majority of the top 100
215 universities (by author count) are based in the United States, five of the top 11 are from

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

222 Publication outcomes

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

237 if 2017 papers are included, that number falls to 64.0 percent. Of preprints posted in
 238 2018, only 20.0 percent have been printed elsewhere (**Figure 3a**).



239

240 **Figure 3.** Characteristics of the bioRxiv preprints published in journals,

241 across the 27 subject collections. **(a)** The proportion of preprints that have

242 been published (y-axis), broken down by the month in which the preprint

243 was first posted (x-axis). **(b)** The proportion of preprints in each category

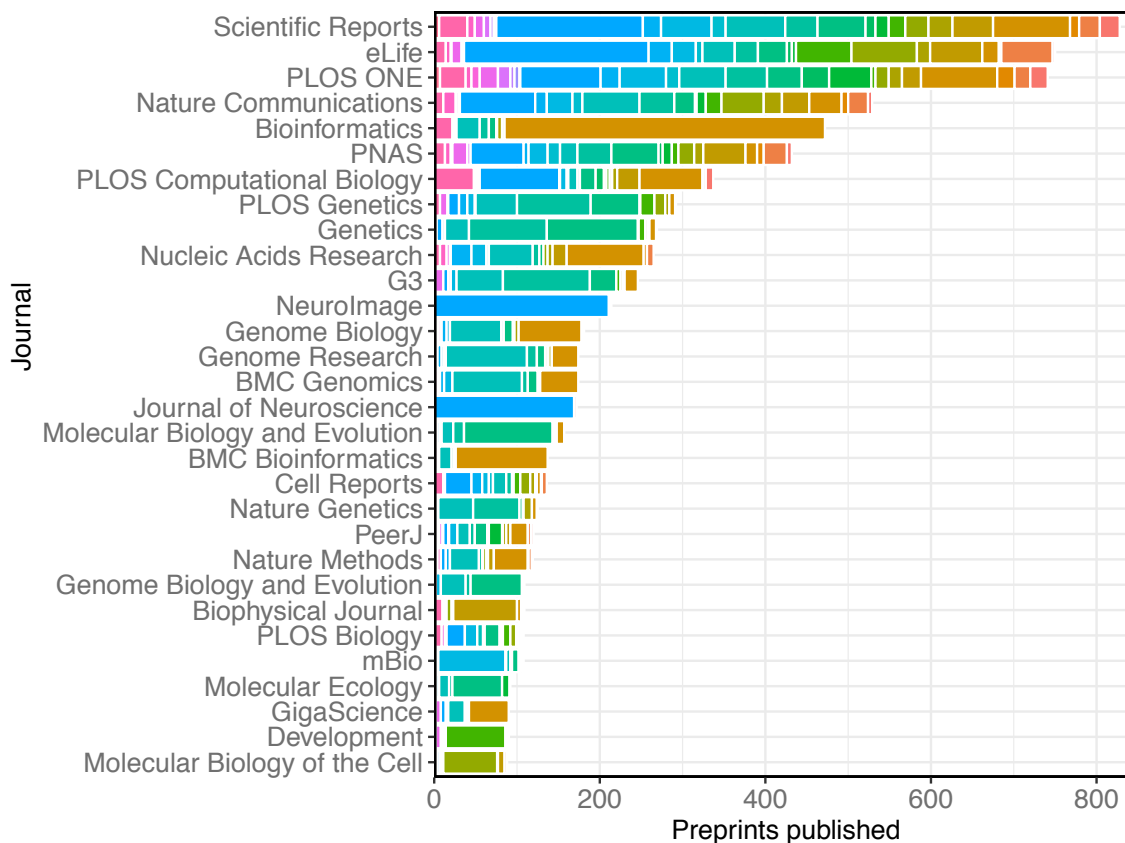
244 that have been published elsewhere. The dashed line marks the overall

245 proportion of bioRxiv preprints that have been published and is at the

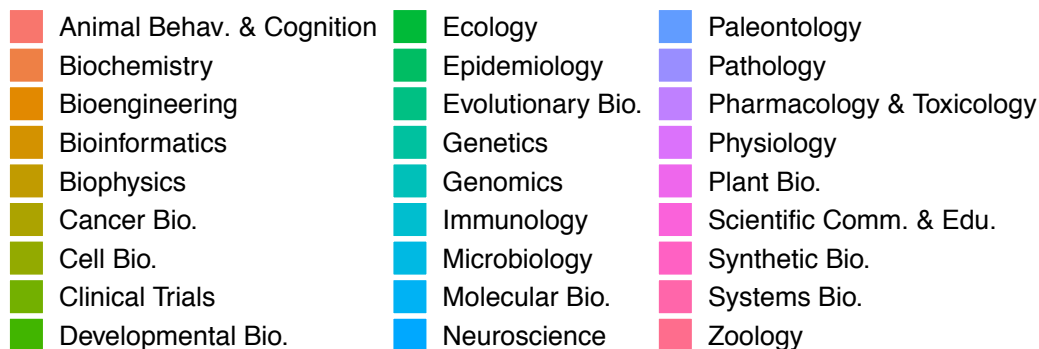
246 same position as the dashed line in panel 3a. **(c)** The number of preprints
247 in each category that have been published in a journal.

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

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



Collection

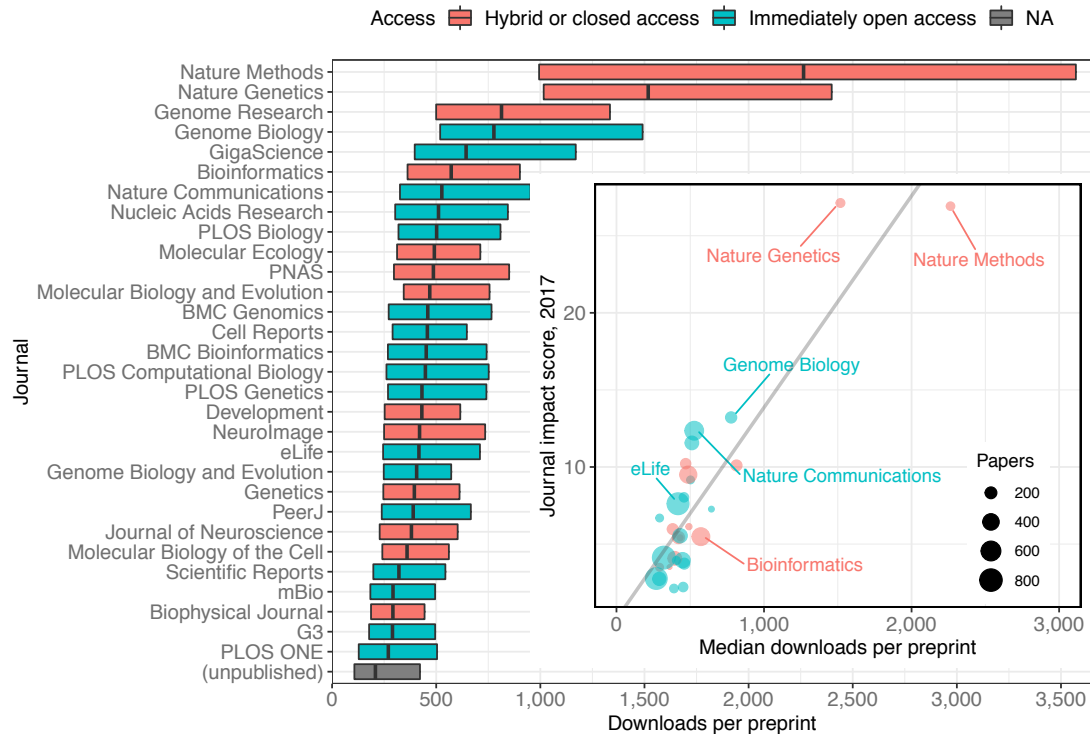


260

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

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

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



278

279

280 **Figure 5.** A modified box plot (without whiskers) illustrating the median

281 downloads of all bioRxiv preprints published in a journal. Each box

282 illustrates the journal's first quartile, median, and third quartile, as in Figure

283 2c. Colors correspond to journal access policy as described in the legend.

284 **(inset)** A scatterplot in which each point represents an academic journal,

285 showing the relationship between median downloads of the bioRxiv

286 preprints published in the journal (x-axis) against its most recent impact

287 score (y-axis). The size of each point is scaled to reflect the total number

288 of bioRxiv preprints published by that journal. The regression line in this

289 plot was calculated using the "lm" function in the R "stats" package, but all

290 reported statistics use the Kendall rank correlation coefficient, which does

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

292 If journals are driving post-publication downloads on bioRxiv, however, their
293 efforts are curiously consistent: Preprints that have been published elsewhere have
294 almost twice as many downloads as preprints that have not (**Table 2**; Mann–Whitney U
295 test, $p < 2.2e-16$). Site-wide, the median number of downloads per preprint is 208,
296 among papers that have not been published. For preprints that *have* been published,
297 the median download count is 394 (Mann–Whitney U test, $p < 2.2e-16$). When preprints
298 published in 2018 are excluded from this calculation, the difference between published
299 and unpublished preprints shrinks, but is still significant (**Table 2**; Mann–Whitney U test,
300 $p < 2.2e-16$). Though preprints posted in 2018 received more downloads *in* 2018 than
301 preprints posted in previous years did (**Figure S3**), it appears they have not yet had
302 time to accumulate as many downloads as papers from previous years (**Figure S4**).

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

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

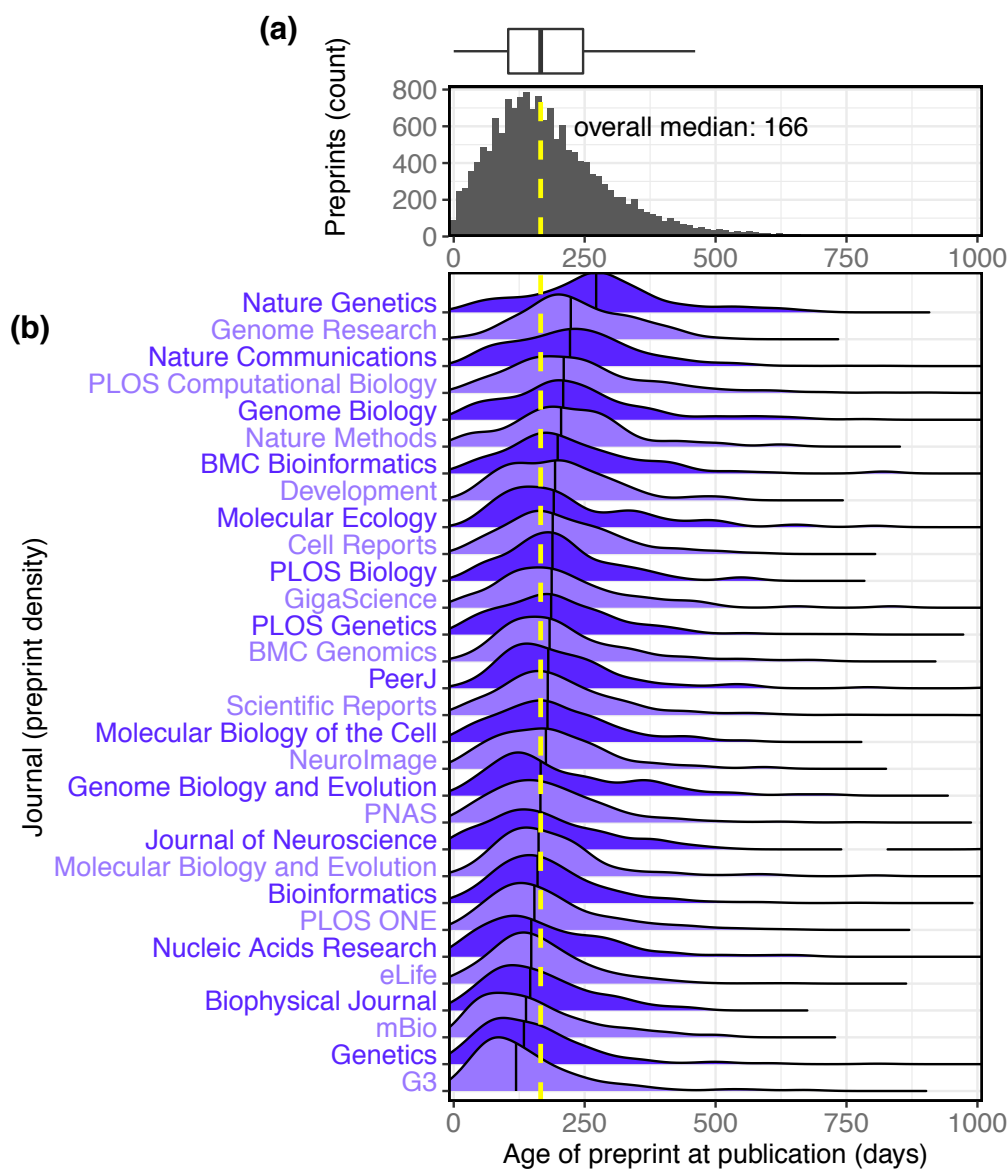
306 **Supplementary file:** *downloads_publication_status.csv*

307 We also retrieved the publication date for all published preprints using the
308 Crossref "Metadata Delivery" API (Crossref 2018). This, combined with the bioRxiv
309 data, gives us a comprehensive picture of the interval between the date a preprint is first
310 posted to bioRxiv and the date it is published by a journal: These data show the median

311 interval is 166 days, or about 5.5 months. 75 percent of preprints are published within
312 247 days of appearing on bioRxiv, and 90 percent are published within 346 days
313 (**Figure 6a**). The median interval we found at the end of November 2018 (166 days) is a
314 23.9 percent increase over the 134-day median interval reported by bioRxiv in mid-2016
315 (Inglis and Sever 2016).

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

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



333

334

335

336

337

338

339

340

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

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

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

348 Discussion

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

363 bioRxiv preprints (Schmid 2016); despite a six-fold increase in the number of published
364 preprints since then, 23 of the top 30 journals found in their results are also in the top 30
365 journals we found.

366 For readers, we show that more than 2,000 new papers are being posted every
367 month, making bioRxiv an increasingly vital source of information for those seeking to
368 stay on top of the most recent research in their fields. This tracks closely with a widely
369 referenced summary of submissions to preprint servers ("Monthly Statistics for October
370 2018" 2018) generated monthly by PrePubMed (<http://www.prepubmed.org>), and
371 expands on submission data collected by researchers using custom web scrapers of
372 their own (Stuart 2016, 2017; Holdgraf 2016). Preprint usage in neuroscience is
373 expanding exceptionally quickly (**Figure 1a**), and collections including bioinformatics,
374 evolutionary biology, and microbiology are growing at a rapid pace (**Figure 1d**). There is
375 also enough data to provide some evidence against the perception that research in
376 preprint is less rigorous than papers appearing in journals ("Methods, preprints and
377 papers" 2017; Vale 2015). In short, the majority of bioRxiv preprints *do* appear in
378 journals eventually, and potentially with very few differences: A 2016 analysis of
379 published preprints that had first been posted to arXiv.org found that "the vast majority
380 of final published papers are largely indistinguishable from their pre-print versions"
381 (Klein et al. 2016).

382 For preprints that are eventually published, we found the median lag time
383 between posting to bioRxiv and publication in a journal is 166 days (**Figure 6a**), and
384 that 75 percent of preprints are published after 247 days on bioRxiv—more than 8
385 months. While this number may seem surprisingly short to researchers, it also provides

386 a lengthy head start to readers looking for the most up-to-date research. The distribution
387 of time to publication is similar to the results from Larivière et al. (2014) showing
388 preprints on arXiv were most frequently published within a year of being posted there,
389 and to a later study examining bioRxiv preprints that found "the probability of publication
390 in the peer-reviewed literature was 48% within 12 months" (Serghiou and Ioannidis
391 2018). Another study published in spring 2017 found that 33.6 percent of preprints from
392 2015 and earlier had been published (Schloss 2017); our data through November 2018
393 show that 68.2 percent of preprints from 2015 and earlier have been published. Multiple
394 studies have examined the interval between submission and publication at individual
395 journals (e.g. Himmelstein 2016a; Royle 2015; Powell 2016), but the incorporation of
396 information about preprints is not as common. We believe this is the first time granular
397 publication rates and timeline statistics have been reported for bioRxiv.

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

409 this dynamic?

410 Researchers have used existing resources and custom scripts to answer
411 questions like these. Himmelstein (2016b) found that only 17.8 percent of bioRxiv
412 papers had an "open license," for example, and another study examined the
413 relationship between Facebook "likes" of preprints and "traditional impact indicators"
414 such as citation count, but found no correlation for papers on bioRxiv (Ringelhan et al.
415 2015). Since most bioRxiv data is not programmatically accessible, many of these
416 studies had to begin by scraping data from the bioRxiv website itself. There have been
417 stated plans to change transition to a new, open-source system (Callaway 2017), but
418 the database and API developed here (<https://rxivist.org>) bring bioRxiv data one step
419 closer to parity with the programmatic interface available for arXiv ("arXiv API" 2018).
420 The Rxivist API allows users to request the details of any preprint or author on the
421 bioRxiv website, and the database snapshots enable bulk querying of preprints using
422 SQL, C, and several other languages ("Procedural Languages" 2019) at a level of
423 complexity currently unavailable using the standard bioRxiv web interface. Using these
424 resources, researchers can now perform detailed and robust bibliometric analysis of the
425 website with the largest collection of preprints in biology, the one that, beginning in
426 September 2018, held more biology preprints than all other major preprint servers
427 combined (Anaya 2018).

428 In addition to our analysis here focused on big picture trends related to bioRxiv,
429 the Rxivist website provides many additional features that may interest preprint readers.
430 Its primary feature is sorting and filtering preprints based by download count or
431 mentions on Twitter, to help users find preprints in particular categories that are being

432 discussed either in the short term (Twitter) or over the span of months (downloads).
433 Several other sites have attempted to use social interaction data to "rank" preprints,
434 though none incorporate bioRxiv download metrics. The "Assert" web application
435 (<https://assert.pub>) ranks preprints from multiple repositories based on data from Twitter
436 and GitHub. The "PromisingPreprints" Twitter bot (<https://twitter.com/PromPreprint>)
437 accomplishes a similar goal, posting links to bioRxiv preprints that receive an
438 exceptionally high social media attention score ("How Is the Altmetric Attention Score
439 Calculated?" 2018) from Altmetric (<https://www.altmetric.com>) in their first week on
440 bioRxiv (De Coster 2017). Arxiv Sanity Preserver (<http://www.arxiv-sanity.com>) provides
441 rankings of arXiv.org preprints based on Twitter activity, though its implementation of
442 this scoring (Karpathy 2018) is more opinionated than that of Rxivist. Other websites
443 perform similar curation, but based on user interactions within the sites themselves:
444 SciRate (<https://scirate.com>), Paperkast (<https://paperkast.com>) and upvote.pub allow
445 users to vote on articles that should receive more attention (van der Silk et al. 2018;
446 Özturan 2018), though upvote.pub is no longer online ("Frontpage" 2018). By
447 comparison, Rxivist doesn't rely on user interaction—by pulling "popularity" metrics from
448 Twitter and bioRxiv, we aim to decouple the quality of our data from the popularity of the
449 website itself.

450 In summary, our approach provides multiple perspectives on trends in biology
451 preprints: (1) the Rxivist.org website, where readers can prioritize preprints and
452 generate reading lists tailored to specific topics, (2) a dataset that can provide a
453 foundation for developers and bibliometric researchers to build new tools, websites, and
454 studies that can further improve the ways we interact with preprints, and (3) an analysis

455 that brings together a comprehensive summary of trends in bioRxiv preprints and an
456 examination of the crossover points between preprints and conventional publishing.

457 Methods

458 Data availability

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

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

469 The Rxivist website

470 We attempted to put the Rxivist data to good use in a relatively straightforward
471 web application. Its main offering is a ranked list of all bioRxiv preprints that can be
472 filtered by areas of interest. The rankings are based on two available metrics: either the
473 count of PDF downloads, as reported by bioRxiv, or the number of Twitter messages
474 linking to that preprint, as reported by Crossref (<https://crossref.org>). Users can also

475 specify a timeframe for the search—for example, one could request the most
476 downloaded preprints in microbiology over the last two months, or view the preprints
477 with the most Twitter activity since yesterday across all categories. Each preprint and
478 each author is given a separate profile page, populated only by Rxivist data available
479 from the API. These include rankings across multiple categories, plus a visualization of
480 where the download totals for each preprint (and author) fall in the overall distribution
481 across all 37,000 preprints and 170,000 authors.

482 The Rxivist API and dataset

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

494 While the analysis presented here deals mostly with overall trends on bioRxiv,
495 the primary entity of the Rxivist API is the individual research preprint, for which we
496 have a straightforward collection of metadata: title, abstract, DOI (digital object

497 identifier), the name of any journal that has also published the preprint (and its new
498 DOI), and which collection the preprint was submitted to. We also collected monthly
499 traffic information for each preprint, as reported by bioRxiv. We use the PDF download
500 statistics to generate rankings for each preprint, both site-wide and for each collection,
501 over multiple timeframes (all-time, year to date, etc.). In the API and its underlying
502 database schema, "authors" exist separately from "preprints" because an author can be
503 associated with multiple preprints. They are recorded with three main pieces of data:
504 name, institutional affiliation and a unique identifier issued by ORCID. Like preprints,
505 authors are ranked based on the cumulative downloads of all their preprints, and
506 separately based on downloads within individual bioRxiv collections. Emails are
507 collected for each researcher, but are not necessarily unique (See below).

508 Data acquisition

509 **Web crawler design.** To collect information on all bioRxiv preprints, we
510 developed an application that pulled preprint data directly from the bioRxiv website. The
511 primary issue with managing this data is keeping it up to date: Rxivist aims to essentially
512 maintain an accurate copy of a subset of bioRxiv's production database, which means
513 routinely running a web crawler against the website to find any new or updated content
514 as it is posted. We have tried to find a balance between timely updates and observing
515 courteous web crawler behavior; currently, each preprint is re-crawled once every two to
516 three weeks to refresh its download metrics and publication status. The web crawler
517 itself uses Python 3 and requires two primary modules for interacting with external
518 services: Requests-HTML (Reitz 2018) is used for fetching individual web pages and

519 pulling out the relevant data, and the psychopg2 module (Di Gregorio et al. 2018) is used
520 to communicate with the PostgreSQL database that stores all of the Rxivist data
521 (PostgreSQL Global Development Group 2017). PostgreSQL was selected over other
522 similar database management systems because of its native support for text search,
523 which, in our implementation, enables users to search for preprints based on the
524 contents of their titles, abstracts and author list. The API, spider and web application are
525 all hosted within separate Docker containers (Docker Inc. 2018), a decision we made to
526 simplify the logistics required for others to deploy the components on their own: Docker
527 is the only dependency, so most workstations and servers should be able to run any of
528 the components.

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

537 **Consolidation of author identities.** Authors are most reliably identified across
538 multiple papers using the bioRxiv feature that allows authors to specify an identifier
539 provided by ORCID (<https://orcid.org>), a nonprofit that provides a voluntary system to
540 create unique identification numbers for individuals. These ORCID ("Open Researcher
541 and Contributor ID") numbers are intended to serve approximately the same role for

542 authors that DOI numbers do for papers (Haak 2012), providing a way to identify
543 individuals whose other information may change over time. 29,559 bioRxiv authors, or
544 17.4 percent, have an associated ORCID. If an individual included in a preprint's list of
545 authors doesn't have an ORCID already recorded in the database, authors are
546 consolidated if they have an identical name to an existing Rxivist author.

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

558 Inconsistent naming patterns cause another chronic issue that is harder to detect
559 and account for. For example, at one point thousands of duplicate authors were indexed
560 in the Rxivist database with various versions of the same name—including a full middle
561 name, or a middle initial, or a middle initial with a period, and so on—which would all
562 have been recorded as separate people if they did not all share an ORCID, to say
563 nothing of authors who occasionally skip specifying a middle initial altogether.
564 Accommodations could be made to account for inconsistencies such as these (using

565 institutional affiliation or email address as clues, for example), but these methods also
566 have the potential to increase the opposite problem of incorrectly combining different
567 authors with similar names who intentionally introduce slight modifications such as a
568 middle initial to help differentiate themselves. One allowance was made to normalize
569 author names: When the web crawler searches for name matches in the database,
570 periods are now ignored in string matches, so "John Q. Public" would be a match with
571 "John Q Public." The other naming problem we encountered was of the opposite
572 variety: multiple authors with identical names (and no ORCID). For example, the Rxivist
573 profile for author "Wei Wang" is associated with 40 preprints and 21 different email
574 addresses but is certainly the conglomeration of multiple researchers. A study of more
575 than 30,000 Norwegian researchers found that when using full names rather than
576 initials, the rate of name collisions was 1.4 percent (Aksnes 2008).

577 **Retrieval of publication date information.** Publication dates were pulled from
578 the Crossref Metadata Delivery API (Crossref 2018) using the publication DOI numbers
579 provided by bioRxiv. Dates were found for all but 31 (0.2%) of the 15,797 published
580 bioRxiv preprints. Because journals measure "publication date" in different ways,
581 several metrics were used. If a "published—online" date was available from Crossref
582 with a day, month and year, then that was recorded. If not, "published—print" was used,
583 and the Crossref "created" date was the final option evaluated. Requests for which we
584 received a 404 response were assigned a publication date of 1 Jan 1900, to prevent
585 further attempts to fetch a date for those entries. These results were filtered out of the
586 analysis. There was no practical way to validate the nearly 16,000 values retrieved, but
587 anecdotal evaluation reveals some inconsistencies: For example, the preprint with the

588 longest interval before publication (1,371 days) has a publication date reported by
589 Crossref of 1 Jul 2018, when it appeared in *IEEE/ACM Transactions on Computational*
590 *Biology and Bioinformatics* 15(4). However, the IEEE website lists a date of 15 Dec
591 2015, two and a half years earlier, as that paper's "publication date," which they define
592 as "the very first instance of public dissemination of content." Since every publisher is
593 free to make their own unique distinctions, these data are difficult to compare at a
594 granular level.

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

611 Data preparation

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

629 Before evaluating data from the table linking published preprints to journals and
630 their post-publication DOI ("article_publications"), journal names were consolidated to
631 avoid under-counting journals with spelling inconsistencies. First, capitalization was
632 stripped from all journal titles, and inconsistent articles ("The Journal of..." vs. "Journal
633 of..."; "and" vs. "&" and so on) were removed. Then, the list of journals was reviewed by

634 hand to remove duplication more difficult to capture automatically: "PNAS" and
635 "Proceedings of the National Academy of Sciences," for example. Misspellings were
636 rare, but one publication in "integrative biology" did appear. See *figures.md* in the
637 project's [GitHub](https://github.com/blekhmanlab/rxivist) repository
638 (<https://github.com/blekhmanlab/rxivist/blob/master/paper/figures.md>) for a full list of
639 corrections made to journal titles. We also evaluated preprints for publication in
640 "predatory journals," organizations that use irresponsibly low academic standards to
641 bolster income from publication fees (Xia et al. 2015). A search for 1,345 journals based
642 on the list compiled by Stop Predatory Journals (<https://predatoryjournals.com>) showed
643 that bioRxiv lists zero papers appearing in those publications ("List of Predatory
644 Journals" 2018).

645 Data analysis

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

655 accompanying R code should create figures containing the exact data as displayed
656 here.

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

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

674 **Calculation of publication intervals.** There are 15,797 distinct preprints with an
675 associated date of publication in a journal, a corpus too large to allow detailed manual
676 validation across hundreds of journal websites. Consequently, these dates are only as
677 accurate as the data collected by Crossref from the publishers. We attempted to use the

678 earliest publication date, but researchers have found that some publishers may be
679 intentionally manipulating dates associated with publication timelines (Royle 2015),
680 particularly the gap between online and print publication, which can inflate journal
681 impact factor (Tort et al. 2012). Intentional or not, these gaps may be inflating the time
682 to press measurements of some preprints and journals in our analysis. In addition, there
683 are 66 preprints (0.42 percent) that have a publication date that falls before the date it
684 was posted to bioRxiv; these were excluded from analyses of publication interval.

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

691 Acknowledgements

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

700 Competing interests

701 The authors declare no competing interests.

702 References

- 703 Abutaleb, Yasmeeen, "Facebook's CEO and wife to give 99 percent of shares to their
704 new foundation." Reuters, 1 Dec 2015. [https://www.reuters.com/article/us-](https://www.reuters.com/article/us-markzuckerberg-baby/facebooks-ceo-and-wife-to-give-99-percent-of-shares-to-their-new-foundation-idUSKBN0TK5UG20151202)
705 [markzuckerberg-baby/facebooks-ceo-and-wife-to-give-99-percent-of-shares-to-](https://www.reuters.com/article/us-markzuckerberg-baby/facebooks-ceo-and-wife-to-give-99-percent-of-shares-to-their-new-foundation-idUSKBN0TK5UG20151202)
706 [their-new-foundation-idUSKBN0TK5UG20151202](https://www.reuters.com/article/us-markzuckerberg-baby/facebooks-ceo-and-wife-to-give-99-percent-of-shares-to-their-new-foundation-idUSKBN0TK5UG20151202)
- 707 Aksnes, Dag W. 2008. "When different persons have an identical author name. How
708 frequent are homonyms?" *Journal of the Association for Information Science and*
709 *Technology* 59: 838-841. doi: 10.1002/asi.20788
- 710 Anaya, Jordan. 2018. PrePubMed: analyses (version 674d5aa).
711 https://github.com/OmnesRes/prepub/tree/master/analyses/preprint_data.txt
712 "arXiv API," arXiv (accessed 18 Dec 2018). <https://arxiv.org/help/api/index>
- 713 Barsh, Gregory S., Casey M. Bergman, Christopher D. Brown, Nadia D. Singh, and
714 Gregory P. Copenhagen. 2016. "Bringing PLOS Genetics Editors to Preprint
715 Servers," *PLOS Genetics* 12(12): e1006448. doi: 10.1371/journal.pgen.1006448
- 716 Berg, Jeremy M., Needhi Bhalla, Philip E. Bourne, Martin Chalfie, David G. Drubin,
717 James S. Fraser, Carol W. Greider, Michael Hendricks, Chonnetia Jones,
718 Robert Kiley, Susan King, Marc W. Kirschner, Harlan M. Krumholz, Ruth
719 Lehmann, Maria Leptin, Bernd Pulverer, Brooke Rosenzweig, John E. Spiro,
720 Michael Stebbins, Carly Strasser, Sowmya Swaminathan, Paul Turner, Ronald
721 D. Vale, K. VijayRaghavan, and Cynthia Wolberger. 2016. "Preprints for the life
722 sciences," *Science* 352(6288), pp. 899–901. doi: 10.1126/science.aaf9133
- 723 Callaway, Ewen. 2013. "Preprints come to life," *Nature* 503, p. 180. doi:
724 10.1038/503180a
- 725 ———. 2017. "BioRxiv preprint server gets cash boost from Chan Zuckerberg
726 Initiative," *Nature* 545(18). doi: 10.1038/nature.2017.21894
- 727 Champieux, Robin. "Gathering Steam: Preprints, Librarian Outreach, and Actions for
728 Change," The Official PLOS Blog, 15 Oct 2018 (accessed 18 Dec 2018).
729 [https://blogs.plos.org/plos/2018/10/gathering-steam-preprints-librarian-outreach-](https://blogs.plos.org/plos/2018/10/gathering-steam-preprints-librarian-outreach-and-actions-for-change/)
730 [and-actions-for-change/](https://blogs.plos.org/plos/2018/10/gathering-steam-preprints-librarian-outreach-and-actions-for-change/)
- 731 Cobb, Matthew. 2017. "The prehistory of biology preprints: A forgotten experiment from
732 the 1960s," *PLOS Biology* 15(11): e2003995. doi: 10.1371/journal.pbio.2003995
- 733 De Coster, Wouter. "A Twitter bot to find the most interesting bioRxiv preprints,"
734 Gigabase or gigabyte, 8 Aug 2017 (accessed 11 Dec 2018).

735 [https://gigabaseorgigabyte.wordpress.com/2017/08/08/a-twitter-bot-to-find-the-](https://gigabaseorgigabyte.wordpress.com/2017/08/08/a-twitter-bot-to-find-the-most-interesting-biorxiv-preprints/)
736 [most-interesting-biorxiv-preprints/](https://gigabaseorgigabyte.wordpress.com/2017/08/08/a-twitter-bot-to-find-the-most-interesting-biorxiv-preprints/)
737 Crossref Metadata Delivery REST API. Web service (accessed 19 Dec 2018).
738 <https://www.crossref.org/services/metadata-delivery/rest-api/>
739 Delamothe, Tony, Richard Smith, Michael A Keller, John Sack, and Bill Witscher. 1999.
740 "Netprints: the next phase in the evolution of biomedical publishing," *BMJ*
741 319(7224): 1515–6. doi: 10.1136/bmj.319.7224.1515
742 Desjardins-Proulx, Philippe, Ethan P. White, Joel J. Adamson, Karthik Ram, Timothée
743 Poisot, and Dominique Gravel. 2013. "The case for open preprints in biology,"
744 *PLOS Biology* 11(5). doi: 10.1371/journal.pbio.1001563
745 Di Gregorio, Federico, and Daniele Varrazzo. 2018. *psycopg2* (version 2.7.5).
746 <https://github.com/psycopg/psycopg2>
747 Docker Inc. 2018. Docker (version 18.06.1-ce). <https://www.docker.com>
748 Feldman, Sergey, Kyle Lo, and Waleed Ammar. 2018. "Citation Count Analysis for
749 Papers with Preprints," arXiv. <https://arxiv.org/abs/1805.05238>
750 Fowler, Kristine K. 2011. "Mathematicians' Views on Current Publishing Issues: A
751 Survey of Researchers," *Issues in Science and Technology Librarianship* 67. doi:
752 10.5062/F4QN64NM
753 "Frontpage," upvote.pub. Archive.org snapshot, 30 Apr 2018 (accessed 29 Dec 2018).
754 <https://web.archive.org/web/20180430180959/https://upvote.pub/>
755 "Funding Opportunities," Chan Zuckerberg Initiative, accessed 18 Dec 2018.
756 <https://chanzuckerberg.com/science/#funding-opportunities>
757 Haak, Laure. "The O in ORCID," ORCID, 5 Dec 2012 (accessed 30 Nov 2018).
758 <https://orcid.org/blog/2012/12/06/o-orcid>
759 Hartgerink, C.H.J. 2015. "Publication cycle: A study of the public Library of Science
760 (PLOS)," accessed 4 Dec 2018.
761 [https://www.authorea.com/users/2013/articles/36067-publication-cycle-a-study-](https://www.authorea.com/users/2013/articles/36067-publication-cycle-a-study-of-the-public-library-of-science-plos/_show_article)
762 [of-the-public-library-of-science-plos/_show_article](https://www.authorea.com/users/2013/articles/36067-publication-cycle-a-study-of-the-public-library-of-science-plos/_show_article)
763 Haustein, Stefanie. 2018. "Scholarly Twitter Metrics," arXiv.
764 <http://arxiv.org/abs/1806.02201>
765 Himmelstein, Daniel, "The history of publishing delays," Satoshi Village, 10 Feb 2016
766 (accessed 29 Dec 2018). <https://blog.dhimmel.com/history-of-delays/>
767 ———, "The licensing of bioRxiv preprints," Satoshi Village, 5 Dec 2016 (accessed 29
768 Dec 2018). <https://blog.dhimmel.com/biorxiv-licenses/>
769 Holdgraf, Christopher R. "The bleeding edge of publishing, Scraping publication
770 amounts at biorxiv," Predictably Noisy, 19 Dec 2016 (accessed 30 Nov 2018).
771 <https://predictablynoisy.com/scrape-biorxiv>
772 "How Is the Altmetric Attention Score Calculated?" Altmetric Support, 5 Apr 2018
773 (accessed 30 Nov 2018).

774 [https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-](https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated)
775 [altmetric-attention-score-calculated](https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated)
776 Inglis, John R., and Richard Sever, "bioRxiv: a progress report." *ASAPbio*. 12 Feb 2016
777 (accessed 5 Dec 2018). <http://asapbio.org/biorxiv>
778 "ERA Home," The Lancet Electronic Research Archive. Archive.org snapshots, 22 Apr
779 2005 and 30 Jul 2005 (accessed 3 Jan 2019).
780 <https://web.archive.org/web/20050422224839/http://www.thelancet.com/era>
781 "Journal Citation Reports Science Edition." 2018. Clarivate Analytics.
782 Kaiser, Jocelyn. 2017. "The preprint dilemma," *Science* 357(6358):1344–1349. doi:
783 10.1126/science.357.6358.1344
784 Karpathy, Andrej. 2018. Arxiv Sanity Preserver, "twitter_daemon.py" (version
785 8e52b8b). [https://github.com/karpathy/arxiv-sanity-](https://github.com/karpathy/arxiv-sanity-preserver/blob/8e52b8ba59bfb5684f19d485d18faf4b7fba64a6/twitter_daemon.py)
786 [preserver/blob/8e52b8ba59bfb5684f19d485d18faf4b7fba64a6/twitter_daemon.py](https://github.com/karpathy/arxiv-sanity-preserver/blob/8e52b8ba59bfb5684f19d485d18faf4b7fba64a6/twitter_daemon.py)
787 Klein, Martin, Peter Broadwell, Sharon E. Farb, and Todd Grappone. 2016. "Comparing
788 published scientific journal articles to their pre-print versions," *Proceedings of the*
789 *16th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 153–162. doi:
790 10.1145/2910896.2910909
791 Kling, Rob, Lisa B. Spector, and Joanna Fortuna. 2003. "The real stakes of virtual
792 publishing: The transformation of E-Biomed into PubMed central," *Journal of the*
793 *Association for Information Science and Technology* 55(2):127–48. doi:
794 10.1002/asi.10352
795 Larivière, Vincent, Cassidy R. Sugimoto, Benoit Macaluso, Staša Milojević, Blaise
796 Cronin, and Mike Thelwall. 2014. "arXiv E-prints and the journal of record: An
797 analysis of roles and relationships," *Journal of the Association for Information*
798 *Science and Technology* 65(6), pp. 1157–69. doi: 10.1002/asi.23044
799 "List of Predatory Journals," Stop Predatory Journals (accessed 28 Dec 2018).
800 <https://predatoryjournals.com/journals/>
801 Marshall, Eliot. 1999. "PNAS to Join PubMed Central--On Condition," *Science*
802 286(5440):655–6. doi: 10.1126/science.286.5440.655a
803 McConnell, John, and Richard Horton. 1999. "Lancet electronic research archive in
804 international health and eprint server," *The Lancet* 354(9172):2–3. doi:
805 10.1016/S0140-6736(99)00226-3.
806 "Methods, preprints and papers." 2017. *Nature Biotechnology* 35(12). doi:
807 10.1038/nbt.4044
808 "Monthly Statistics for October 2018," PrePubMed, accessed 17 Dec 2018.
809 http://www.prepubmed.org/monthly_stats/
810 "Nature respects preprint servers," 2005. *Nature* 434, p. 257. doi: 10.1038/434257b
811 O’Roak, Brian. "How I learned to stop worrying and love preprints," *Spectrum*, 22 May
812 2018 (accessed 30 Nov 2018). [https://www.spectrumnews.org/opinion/learned-](https://www.spectrumnews.org/opinion/learned-stop-worrying-love-preprints/)
813 [stop-worrying-love-preprints/](https://www.spectrumnews.org/opinion/learned-stop-worrying-love-preprints/)

- 814 Özturan, Doğançan. "Paperkast: Academic article sharing and discussion," 2 Sep 2018
815 (accessed 8 Jan 2019). [https://medium.com/@dogancan/paperkast-academic-](https://medium.com/@dogancan/paperkast-academic-article-sharing-and-discussion-e1aebc6fe66d)
816 [article-sharing-and-discussion-e1aebc6fe66d](https://medium.com/@dogancan/paperkast-academic-article-sharing-and-discussion-e1aebc6fe66d)
- 817 PostgreSQL Global Development Group. 2017. PostgreSQL (version 9.6.6).
818 <https://www.postgresql.org>
- 819 Powell, Kendall. 2016. "Does it take too long to publish research?" *Nature* 530, pp.
820 148–151. doi: 10.1038/530148a
- 821 "Procedural Languages," PostgreSQL Documentation (version 9.4.20), accessed 1 Jan
822 2019. <https://www.postgresql.org/docs/9.4/xplang.html>
- 823 Raff, Martin, Alexander Johnson, and Peter Walter. 2008. "Painful Publishing," *Science*
824 321(5885):36. doi: 10.1126/science.321.5885.36a
- 825 Reitz, Kenneth. 2018. Requests-HTML (version 0.9.0).
826 <https://github.com/kennethreitz/requests-html>
- 827 "Reporting Preprints and Other Interim Research Products," notice number NOT-OD-
828 17-050. National Institutes of Health. 24 Mar 2017 (accessed 7 Jan 2019).
829 <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-050.html>
- 830 Ringelhan, Stefanie, Jutta Wollersheim, and Isabell M. Welp. 2015. "I Like, I Cite? Do
831 Facebook Likes Predict the Impact of Scientific Work?" *PLOS ONE* 10(8):
832 e0134389. doi: 10.1371/journal.pone.0134389
- 833 Rørstad, Kristoffer, and Dag W. Aksnes. 2015. "Publication rate expressed by age,
834 gender and academic position – A large-scale analysis of Norwegian academic
835 staff," *Journal of Informetrics* 9(2). doi: 10.1016/j.joi.2015.02.003
- 836 Royle, Stephen, "What The World Is Waiting For," quantixed, 17 Oct 2014 (accessed
837 29 Dec 2018). <https://quantixed.org/2014/10/17/what-the-world-is-waiting-for/>
838 ———, "Waiting to happen II: Publication lag times," quantixed, 16 Mar 2015
839 (accessed 29 Dec 2018). [https://quantixed.org/2015/03/16/waiting-to-happen-ii-](https://quantixed.org/2015/03/16/waiting-to-happen-ii-publication-lag-times/)
840 [publication-lag-times/](https://quantixed.org/2015/03/16/waiting-to-happen-ii-publication-lag-times/)
- 841 Schloss, Patrick D. 2017. "Preprinting Microbiology," *mBio* 8:e00438-17. doi:
842 10.1128/mBio.00438-17
- 843 Schmid, Marc W. 2016. crawlBiorxiv (version e2af128).
844 <https://github.com/MWSchmid/crawlBiorxiv/blob/master/README.md>
- 845 Schwarz, Greg J., and Robert C. Kennicutt Jr. 2004. "Demographic and Citation
846 Trends in Astrophysical Journal papers and Preprints," arXiv.
847 <https://arxiv.org/abs/astro-ph/0411275>
- 848 Sever, Richard. Twitter Post. 1 Nov 2018, 9:29 AM.
849 <https://twitter.com/cshperspectives/status/1058002994413924352>
- 850 van der Silk, Noon, Aram Harrow, Jaiden Mispy, Dave Bacon, Steven Flammia,
851 Jonathan Oppenheim, James Payor, Ben Reichardt, Bill Rosgen, Christian
852 Schaffner, and Ben Toner. "About," SciRate, accessed 30 Nov 2018.
853 <https://scirate.com/about>

- 854 Smaglik, Paul. "E-Biomed Becomes Pubmed Central," *The Scientist*, 27 Sep 1999
855 (accessed 29 Dec 2018). [https://www.the-scientist.com/news/e-biomed-](https://www.the-scientist.com/news/e-biomed-becomes-pubmed-central-56359)
856 [becomes-pubmed-central-56359](https://www.the-scientist.com/news/e-biomed-becomes-pubmed-central-56359)
- 857 Snyder, Solomon H. 2013. "Science interminable: Blame Ben?" *PNAS* 110(7):2428–9.
858 doi: 10.1073/pnas.201300924
- 859 Stuart, Tim, "bioRxiv," 1 Mar 2016 (accessed 2 Jan 2019).
860 <http://timoast.github.io/blog/2016-03-01-biorxiv/>
861 ———, "bioRxiv 2017 update," 4 Oct 2017 (accessed 2 Jan 2019).
862 <http://timoast.github.io/blog/biorxiv-2017-update/>
- 863 Serghiou, Stylianos, and John P.A. Ioannidis. 2018. "Altmetric Scores, Citations, and
864 Publication of Studies Posted as Preprints," *JAMA* 318(4): 402–4. doi:
865 10.1001/jama.2017.21168
- 866 "Submission Guide," bioRxiv, accessed 30 Nov 2018. [https://www.biorxiv.org/submit-a-](https://www.biorxiv.org/submit-a-manuscript)
867 [manuscript](https://www.biorxiv.org/submit-a-manuscript)
- 868 Thelwall, Mike, Stefanie Haustein, Vincent Larivière, and Cassidy R. Sugimoto. 2013.
869 "Do Altmetrics Work? Twitter and Ten Other Social Web Services," *PLOS ONE*
870 8(5): e64841. doi: 10.1371/journal.pone.0064841
- 871 Tort, Adriano B.L., Zé H. Targino, and Olavo B. Amaral. 2012. "Rising Publication
872 Delays Inflate Journal Impact Factors," *PLOS ONE* 7(12): e53374. doi:
873 10.1371/journal.pone.0053374
- 874 Vale, Ronald D. 2015. "Accelerating scientific publication in biology," *PNAS*
875 112(44):13439–46. doi: 10.1073/pnas.1511912112
- 876 Vale, Ronald D., and Anthony A. Hyman. 2016. "Priority of discovery in the life
877 sciences," *eLife* 5(e16931). Doi: 10.7554/eLife.16931
- 878 Varmus, Harold. "E-BIOMED: A Proposal for Electronic Publications in the Biomedical
879 Sciences," National Institutes of Health, 5 May 1999. Archive.org snapshot, 18
880 Oct 2015 (accessed 29 Dec 2018).
881 [https://web.archive.org/web/20151018182443/https://www.nih.gov/about/director/](https://web.archive.org/web/20151018182443/https://www.nih.gov/about/director/pubmedcentral/ebiomedarch.htm)
882 [pubmedcentral/ebiomedarch.htm](https://web.archive.org/web/20151018182443/https://www.nih.gov/about/director/pubmedcentral/ebiomedarch.htm)
- 883 Vence, Tracy. "Journals Seek Out Preprints," *The Scientist*, 18 Jan 2017 (accessed 7
884 Jan 2019). [https://www.the-scientist.com/news-opinion/journals-seek-out-](https://www.the-scientist.com/news-opinion/journals-seek-out-preprints-32183)
885 [preprints-32183](https://www.the-scientist.com/news-opinion/journals-seek-out-preprints-32183)
- 886 Verma, Inder M. 2017. "Preprint servers facilitate scientific discourse," *PNAS* 114(48).
887 doi: 10.1073/pnas.1716857114
- 888 Wang, Zhiqi, Wolfgang Glänzel, and Yue Chen. 2018. "How Self-Archiving Influences
889 the Citation Impact of a Paper: A Bibliometric Analysis of arXiv Papers and Non-
890 arXiv Papers in the Field of Information and Library Science," Leiden, The
891 Netherlands: Proceedings of the 23rd International Conference on Science and
892 Technology Indicators (ISBN: 978-90-9031204-0), pages 323–30.
893 <https://openaccess.leidenuniv.nl/handle/1887/65329>

894 Xia, Jingfeng, Jennifer L. Harmon, Kevin G. Connolly, Ryan M. Donnelly, Mary R.
895 Anderson, and Heather A. Howard. 2015. "Who published in 'predatory'
896 journals?" *Journal of the Association for Information Science and Technology*
897 66(7). doi: 10.1002/asi.23265