International authorship and collaboration in bioRxiv preprints

Richard J. Abdill[1], Elizabeth M. Adamowicz[1], Ran Blekhman[1,2*]

1 – Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN, USA
2 – Department of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, MN, USA
* – Please direct correspondence to blekhman@umn.edu

Abstract

As preprints become more integrated into the conventional avenues of scientific communication, it is critical to understand who is being included and who is not. However, little is known about which countries are participating in the phenomenon or how they collaborate with each other. Here, we present an analysis of 67,885 preprints posted to bioRxiv from 2013 through 2019 that includes the first comprehensive dataset of country-level affiliations for all preprint authors. We find the plurality of preprints (37%) come from the United States, more than three times as many as the next-most prolific country, the United Kingdom (10%). We find some countries are overrepresented on bioRxiv relative to their overall scientific output: The U.S. and U.K. are again at the top of the list, with other countries such as China, India and Russia showing much lower levels of bioRxiv adoption despite comparatively high numbers of scholarly publications. We describe a subset of "contributor countries" including Uganda, Croatia, Thailand, Greece and Kenya, which appear on preprints almost exclusively as part of international collaborations and seldom in the senior author position. Lastly, we find multiple journals that disproportionately favor preprints from some countries over others, a dynamic that almost always benefits manuscripts with a senior author affiliated with the United States.

Introduction

Biology preprints are being shared online at an unprecedented rate (Narock and Goldstein 2019; Abdill and Blekhman 2019b). Since 2013, more than 73,000 preprints have been posted to bioRxiv.org, the largest preprint server in the life sciences, including 29,178 in 2019 alone (Abdill and Blekhman 2019a). In addition to their rising popularity among researchers seeking to share their work outside the traditional pipelines of peer-reviewed journals, preprints provide authors with numerous potential benefits: Preprints may receive more citations after publication (Fu and Hughey 2019; Fraser et al. 2020), and journals proactively search preprint servers to solicit submissions (Barsh et al. 2016; Vence 2017). Programs such as In Review (https://researchsquare.com) and Review Commons (https://www.reviewcommons.org) coordinate with journals for peer review of preprints, and in late 2019 the journal *eLife* announced a "Preprint Review" program in which bioRxiv preprints submitted to *eLife* would be guaranteed to be sent out for peer review (Eisen 2019). A growing number of programs are being launched to encourage the use of preprints, and, in the cases of Review Commons and *eLife*, the use of bioRxiv specifically. However, very little is known about who is benefiting from this attention, who remains left out, and how the technical and professional challenges of this new publishing paradigm impact different groups (Penfold and Polka 2020). Despite all the recent research about preprints, one critical question remains: Where do they come from? More specifically, which

countries are participating in the preprint ecosystem, how are they working with each other, and what happens when they do?

To answer these questions, we looked at country-level participation and outcomes. Academic publishing has grappled for decades with hard-to-quantify concerns about unspoken (and occasionally unconscious) factors of success that are not directly linked to research quality. Studies have found bias in favor of wealthy, English-speaking countries in citation count (Akre et al. 2011) and the acceptance of both papers (Saposnik et al. 2014; Okike et al. 2008) and conference abstracts (Ross et al. 2006). There have also long been concerns regarding how the peer review process is influenced by institutional prestige, among other factors (Lee et al. 2013). Preprints have been praised as a democratizing influence on scientific communication (Berg et al. 2016), and the unlinking of research dissemination from peer review may dramatically alter the publishing landscape. Research suggests U.S. authors are overrepresented on bioRxiv compared to published literature (Fraser et al. 2020), but the scientific community lacks a more specific understanding of who is availing themselves of preprint-based research dissemination opportunities. Here, we aim to answer these questions by analyzing a dataset of all preprints posted to bioRxiv through 2019. After collecting author-level metadata for each preprint, we determined each author's institutional affiliation to summarize authorship measurements at national levels.

Results

Preprint origins

We retrieved author data for 67,885 preprints for which the most recent version was posted before January 1, 2020. First, we attributed each preprint to a single country, using the affiliation of the last individual in the author list, considered by convention in the life sciences to be the "senior author" who supervised the work (see **Methods**). 25,305 manuscripts (37.3%) have a senior author from the United States, followed by 6,845 manuscripts (10.1%) from the United Kingdom (**Fig. 1a**). North America, Europe and Australia dominate the top spots, though China (3.6%), Japan (1.8%) and India (1.6%) are the sources of more than 1,100 preprints each (**Fig. 1b**). Brazil, with 646 manuscripts, has the 15th-most preprints and is the first South American country on the list, followed by Argentina (151 preprints) in 32nd place. South Africa (179 preprints) is the first African country on the list, in 28th place, followed by Ethiopia (57 preprints) in 41st place (**Supplementary Table 1**). Interestingly, both South Africa and Ethiopia were found to have high opt-in rates for a program operated by PLOS journals that enabled submissions to be sent directly to bioRxiv ("Trends in Preprints" 2019).

These attributions were made using the author listed last on each preprint, but we found similar results when we looked at which countries were most highly represented based on authorship at any position (**Table 1**). Overall, U.S. authors appear on the most bioRxiv preprints—33,968 manuscripts (50.0%) include at least one U.S. author (**Fig. 1c**).

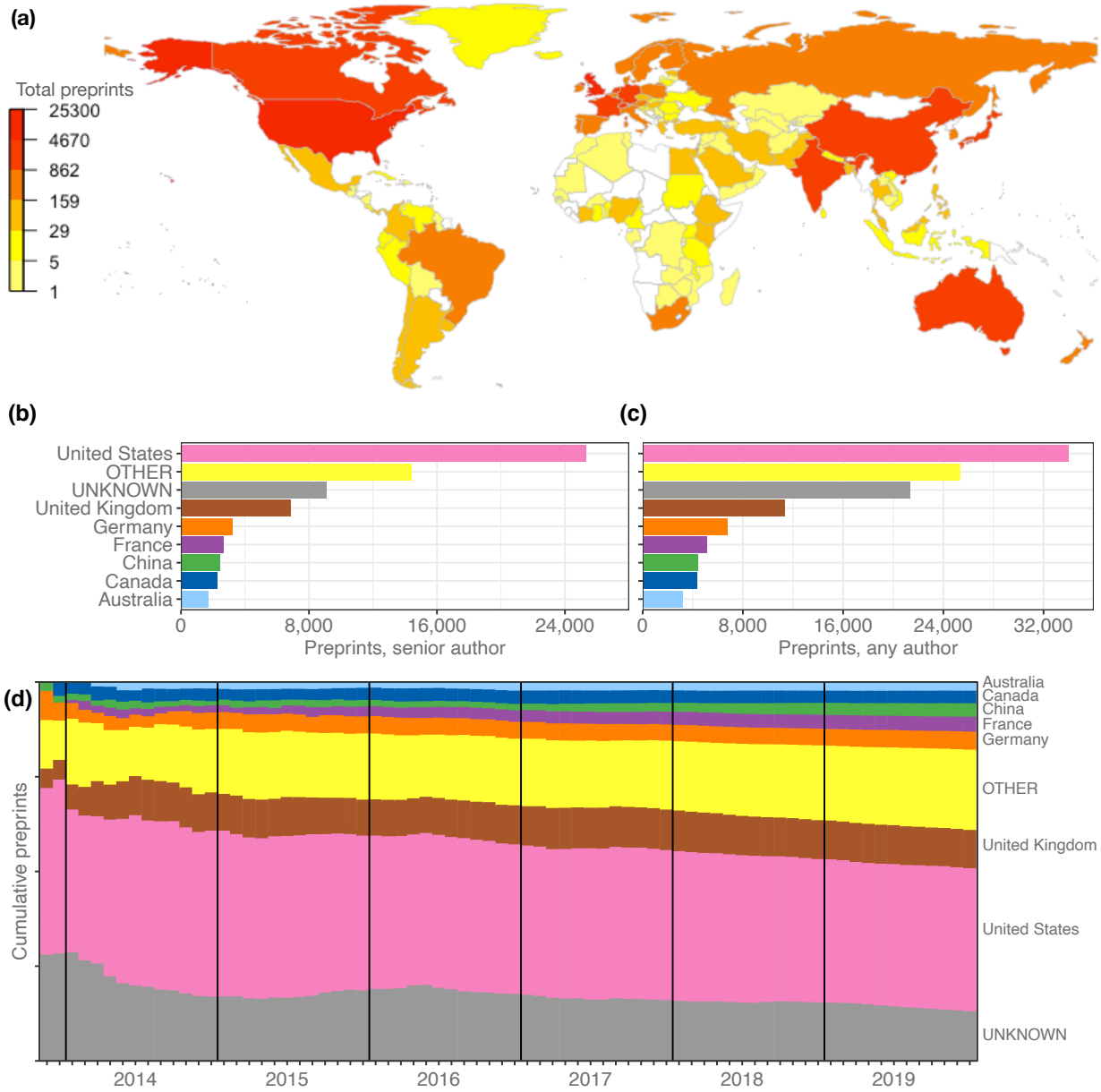

Figure 1. Preprints per country. (a) A heat map indicating the number of preprints per country, based on the institutional affiliation of the senior author. The color coding uses a log scale that splits the full range of preprint counts into six colors. **(b)** The total preprints attributed to the seven most prolific countries. The x-axis indicates total preprints listing a senior author from a country; the y-axis indicates the country. The "OTHER" category includes preprints from all countries not listed in the plot. **(c)** Similar to panel b, but showing the total preprints listing at least one author from the country in any position, not just the senior position. **(d)** Proportion of total senior-author preprints from each country (y-axis) over time (x-axis), starting in November 2013 and continuing through December 2019. Each colored segment indicates the proportion of total preprints attributed to a single country, as of the end of the month indicated on the x-axis. Colors indicate countries, using the same scale as panels B and C.

Over time, the country-level proportions on bioRxiv have remained remarkably stable (**Fig. 1d**), even as the number of preprints grew exponentially: At the end of 2015, Germany accounted for 4.5% of bioRxiv's 2,460 manuscripts. At the end of 2019, Germany was responsible for 4.8% of 67,885 preprints. However, the proportion of preprints from countries outside the top seven contributing countries is growing slowly (**Fig. 1d**): At the end of 2015, these countries accounted for 17.6 percent of preprints. By the end of 2019, that number had grown to 21.2 percent, when bioRxiv hosted preprints from senior authors affiliated with 135 countries.

Country	Preprints, senior author (proportion)	Preprints, any author (proportion)
United States	25,305 (37.3%)	33,968 (50.0%)
(Unknown)	9,077 (13.4%)	21,332 (31.4%)
United Kingdom	6,845 (10.1%)	11,335 (16.7%)
Germany	3,234 (4.8%)	6,772 (10.0%)
France	2,669 (3.9%)	5,088 (7.5%)
China	2,419 (3.6%)	4,400 (6.5%)
Canada	2,245 (3.3%)	4,294 (6.3%)
Australia	1,704 (2.5%)	3,213 (4.7%)
Switzerland	1,234 (1.8%)	2,656 (3.9%)
Netherlands	1,200 (1.8%)	2,701 (4.0%)
Japan	1,195 (1.8%)	2,242 (3.3%)
India	1,107 (1.6%)	1,656 (2.4%)

Table 1. Preprints per country. All 11 countries with more than 1,000 preprints attributed to a senior author affiliated with that country. The percentages in the "Preprints, any author" column sum to more than 100 percent because preprints may be counted for more than one country. A full list of countries is provided in **Supplemental Table 1**.

We noted that some patterns may be obscured by countries that had hundreds or thousands of times as many preprints as other countries, so we re-evaluated these ranks after adjusting for overall scientific output (**Fig. 2**). This was measured by the number of "citable documents" associated with each country from 2014 through 2018 in the SCImago Journal & Country Rank portal ("Scimago Journal & Country Rank" n.d.). For all countries with at least 3,000 citable documents and 50 preprints, we generated a productivity-adjusted score, termed "bioRxiv adoption," by taking the proportion of preprints with a senior author from that country and dividing it by that country's proportion of citable documents from 2014–2018. **Fig. 2a** illustrates this relationship: Given a country's total citable documents and total preprints, the diagonal line represents an adoption score of 1.0, which would indicate that a country's share of bioRxiv preprints is identical to its share of general scholarly outputs; a score of 2.0 would indicate that its share of preprints is twice as high as its share of other scholarly outputs (See **Discussion** for more about this measurement.)

The U.S. posted 25,305 preprints and published about 2.8 million citable documents, for a bioRxiv adoption score of 2.15 (**Fig. 2b**). Seven of the nine countries with adoption scores above 1.0 were

from North America and Europe, but Israel has the third-highest score (1.46) based on its 565 preprints. Ethiopia has the fifth-highest bioRxiv adoption (1.19): Though only 57 preprints list a senior author with an affiliation in Ethiopia, the country had a total of 11,624 citable documents published between 2014 and 2018 (**Supplementary Table 2)**. In other words, 4.9 out of every 1,000 Ethiopian research outputs is on bioRxiv, compared to 8.9 out of every 1,000 American research outputs.

By comparison, some countries are present on bioRxiv at much lower frequencies than would be expected, given their participation in scientific publishing in general (**Fig. 2c**): Turkey, for example, published 201,860 citable documents from 2014 through 2018 but was the senior author on only 71 preprints, for a bioRxiv adoption score of 0.09. Russia (241 preprints), Malaysia (72 preprints), Iran (116 preprints) and Greece (54 preprints) all have adoption scores below 0.18. The largest country with a low adoption score is China (2,506,694 citable documents; 2,419 preprints; bioRxiv adoption=0.23), which published more than 15 percent of the world's citable documents (according to SCImago) but was the source of only 3.6 percent of preprints (**Fig. 2b**).

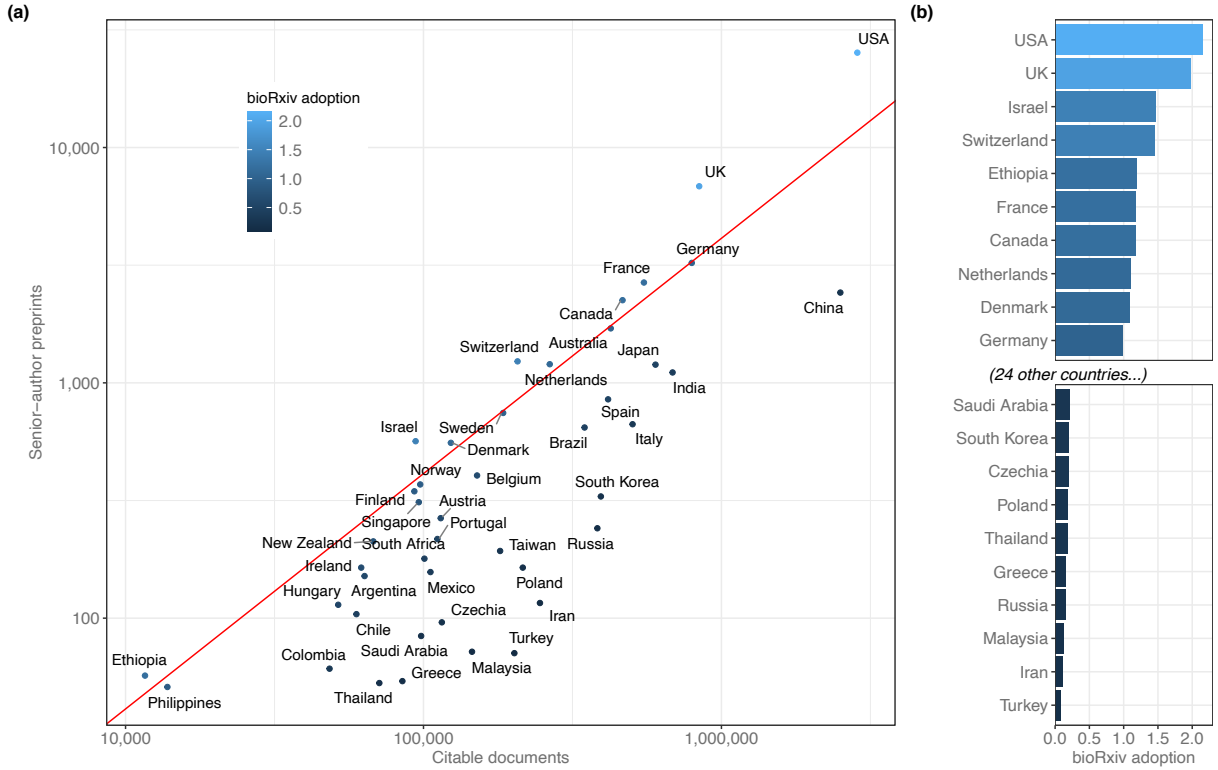

Figure 2. BioRxiv adoption per country. (a) Correlation between two scientific output metrics. Each point is a country; the x-axis (log scale) indicates the total citable documents attributed to that country from 2014–2018, and the y-axis (also log scale) indicates total senior-author preprints attributed to that country overall. The red line demarcates a "bioRxiv adoption" score of 1.0, indicating that a country is as represented on bioRxiv as they are in published literature. Countries to the left of this line have a bioRxiv adoption score greater than 1.0. **(b)** The countries with the 10 highest and 10 lowest bioRxiv adoption scores. The x-axis indicates each country's adoption score, and the y-axis lists each country in order. Data from this

figure is available in **Supplementary Table 2**. All panels include only countries with at least 50 preprints.

Collaboration

After analyzing preprints using senior authorship, we also evaluated interactions *within* manuscripts to better understand collaborative patterns found on bioRxiv. We found the number of authors per paper increased from 3.08 in 2014 to 4.26 at the end of 2019 (**Fig. S1**). The monthly average authors per preprint has increased linearly with time (Pearson's r=0.9488, $p=8.93\times10^{-38}$), a pattern that has also been observed (at a less dramatic rate) in published literature (Adams et al. 2005; Wuchty, Jones, and Uzzi 2007; Bordons, Aparicio, and Costas 2013). Examining the number of countries represented in each preprint (**Fig. S1**), we found that 24,011 preprints (35.4%) included authors from two or more countries; 2,867 preprints (4.2%) were from four or more countries, and one preprint, "Fine-mapping of 150 breast cancer risk regions identifies 178 high confidence target genes," listed 319 authors from 39 countries, the most countries listed on any single preprint. The mean number of countries represented per preprint is 1.836, which has remained fairly stable since 2014 despite steadily growing author lists overall (**Fig. S1**).

We then looked at countries appearing on at least 50 international preprints to examine basic patterns in international collaboration. We found many countries with comparatively low output contributed almost exclusively to international collaborations: For example, researchers listing an affiliation in Vietnam appear on 76 preprints; 73 (96.1%) include at least one researcher from another country. Similarly, Uganda, Tanzania, Croatia, Ecuador and Peru also have international collaboration rates of greater than 90%.

Upon closer examination, we found these countries were part of a larger group, which we call "contributor countries," that (1) appear mostly on preprints with authors from other countries, but (2) seldom as the senior author. For this analysis, we defined a contributor country as one that has contributed to at least 50 international preprints but appears in the senior author spot of less than 20 percent of them. (We excluded countries with less than 50 preprints to minimize the effect of dynamics that could be explained by countries with just one or two labs that frequently worked with international collaborators.) 18 countries met these criteria (**Fig. 3a**). Of these, Uganda had the lowest international senior-author rate: Of the 84 international preprints that include an author with an affiliation in Uganda, only 5 preprints (6.0%) include a senior author from Uganda. Other countries with low senior-author rates include Vietnam (8.2%), Tanzania (8.2%) and Croatia (9.7%). By comparison, the highest international senior-author rate was observed for the United States, which appears as senior author on 47.2% of all international preprints it contributes to (**Fig. 3b**).

In addition to a high percentage of international collaborations and a low percentage of senior-author preprints, another characteristic of contributor countries is a comparatively low number of preprints overall. To define this subset of countries more clearly, we examined whether there was a relationship between any of the three factors we identified, but across all countries with at least 30 international preprints, rather than only among contributors. We found consistent patterns for all three (see **Methods**): First, countries with fewer international collaborations also tend to appear as senior author on a smaller proportion of those preprints (Spearman's $\rho=0.616$, $p=1.513\times10^{-9}$;

Fig. S2a). We also observed a negative correlation between *total* international collaborations and international collaboration *rate*—that is, the proportion of preprints a country contributes to that include at least one contributor from another country (**Fig. S2b**; Spearman's ρ=-0.543, p=2.408×10^{-7}). This indicates that countries with mostly international preprints (**Fig. 3c**) also tended to have *fewer* international collaborations (**Fig. 3d**) than other countries. Finally, we found a negative correlation between international collaboration rate and the proportion of international preprints for which a country appears as senior author (Spearman's ρ=-0.492, p=4.114×10^{-6}; **Fig. S2c**), demonstrating that countries that appear mostly on international preprints (**Fig. 3c**) are less likely to appear as senior author of those preprints (**Fig. 3b**). Similar patterns have been observed in previous studies: González-Alcaide et al. (2017) found countries ranked lower on the Human Development Index participated more frequently in international collaborations, and a review of oncology papers found that researchers from low- and middle-income countries collaborated on randomized control trials, but rarely as senior author (Wong et al. 2014).

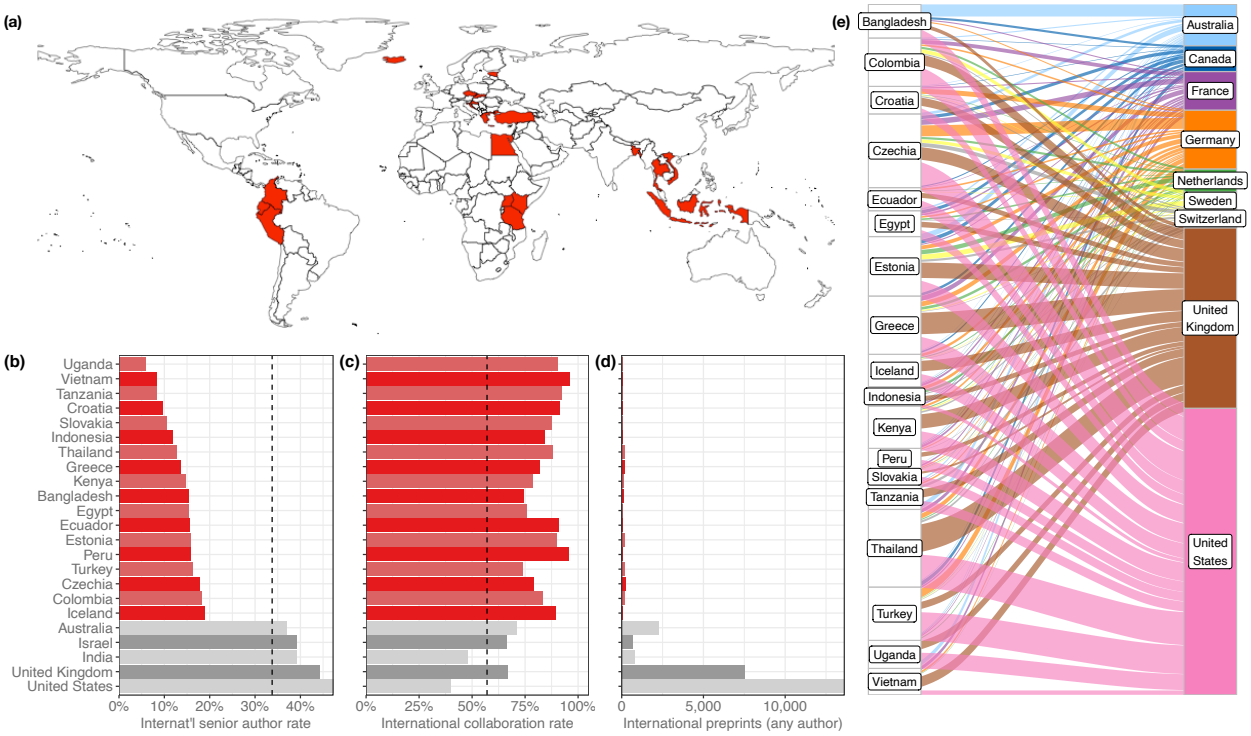

Figure 3. Contributor countries. (a) World map indicating (in red) the location of contributor countries, defined as all countries listed on at least 50 international preprints, but as senior author on less than 20% of them. **(b)** Bar plot indicating the international senior author rate (y-axis) by country (x-axis)—that is, of all international preprints with a contributor from that country, the percentage of them that include a senior author from that country. All 18 contributor countries are listed in red, with the five countries with the highest senior-author rates (in grey) for comparison. **(c)** A bar plot with the same y-axis as panel (b). The x-axis indicates the international collaboration rate, or the proportion of preprints with a contributor from that country that also include at least one author from another country. **(d)** is a bar plot indicating the total international preprints featuring at least one author from that country (the median value per country is 19). Expanded data from this figure is available as **Supplemental Table 3**. **(e)** On the left are the 18 contributor

countries. On the right are the countries that appear in the senior author position of preprints that were co-authored with contributor countries. (Supervising countries with 25 or fewer preprints with contributor countries were excluded from the figure.) The width of the ribbons connecting contributor countries to senior-author countries indicates the number of preprints supervised by the senior-author country that included at least one author from the contributor country.

After generating a list of preprints with authors from contributor countries, we examined which countries appeared most frequently in the senior author spot of those preprints (**Fig. 3e**). Among the 1,824 preprints with an author from a contributor country, 521 (28.6%) had senior authors listing an affiliation in the United States (**Supplementary Table 3**). The United Kingdom was listed as senior author on the next-most preprints with contributor countries, at 328 (18.0%), followed by Germany (5.9%) and France (3.8%). Given the large differences in preprint authorship between countries, we tested which of these senior-author relationships was disproportionately large. After multiple-test correction using the Benjamini–Hochberg procedure, we found seven links between contributor countries and senior-author countries that were significant (**Supplementary Table 4**). The strongest link is between Bangladesh and Australia: Of the 83 preprints with a contributor from Bangladesh, Australia appears as the senior author on 22 of them (Fisher's exact test, $q=2.60\times10^{-12}$). The United States is also frequently senior author on preprints with a contributor in Turkey (52 of 83 preprints, q=0.012). The remaining five links were between a contributor country and the United Kingdom, which appears as senior author with disproportionate frequency on preprints with authors in Thailand ($q=4.73\times10^{-5}$), Greece (q=0.0016), Kenya (q=0.012), Vietnam (q=0.012) and Iceland (q=0.040).

Outcomes

After quantifying which countries were posting preprints, we also examined whether there were differences in preprint outcomes between countries. We obtained monthly download counts for all preprints, as well as publication status, the publishing journal, and date of publication for all preprints flagged as "published" on bioRxiv (see **Methods**). We then evaluated country-level patterns for the 35 countries with at least 100 senior-author preprints.

Overall, the median number of PDF downloads per preprint is 336 (**Fig. 4a**). Among countries with at least 100 preprints, Austria has the highest median downloads per preprint, with 385.5, followed by the United States (369) and Denmark (368.5). Taiwan has the lowest median, at 196 downloads. Next-fewest is Argentina (205), Brazil (220) and a tie at 235 downloads between Mexico and South Korea. Across all countries with at least 100 preprints, there was a weak correlation between total preprints attributed to a country and the median downloads per preprint (Spearman's $\rho=0.484$, p=0.00323) (**Fig. 4b**), and another correlation between median downloads per preprint and the country's publication rate (Spearman's $\rho=0.725$, $p=8.43\times10^{-7}$) (**Fig. 4c**).

Next, we examined country-level publication rates by assigning preprints posted prior to 2019 to countries using the affiliation of the senior author, then measuring the proportion of those preprints flagged as "published" on the bioRxiv website. Overall, 62.6 percent of pre-2019 preprints were published (**Supplementary Table 5**). Ireland had the highest publication rate (**Fig. 4d)**, with 48 of their 65 preprints (73.9%) published before March 2020, followed by New Zealand (90 of 127,

70.9%) and Switzerland (455 of 651, 69.9%). Among countries with at least 350 preprints prior to 2019, Switzerland had the highest publication rate, followed by Germany (1104 of 1630, 67.7%), the Netherlands (414 out of 620, 66.8%) and France (898 of 1350, 66.5%). The lowest publication rates were observed for Iran (26 of 60, 43.3%) and China (508 of 1155, 44.0%); South Korea, India, Brazil and Taiwan all had publication rates below 50 percent.

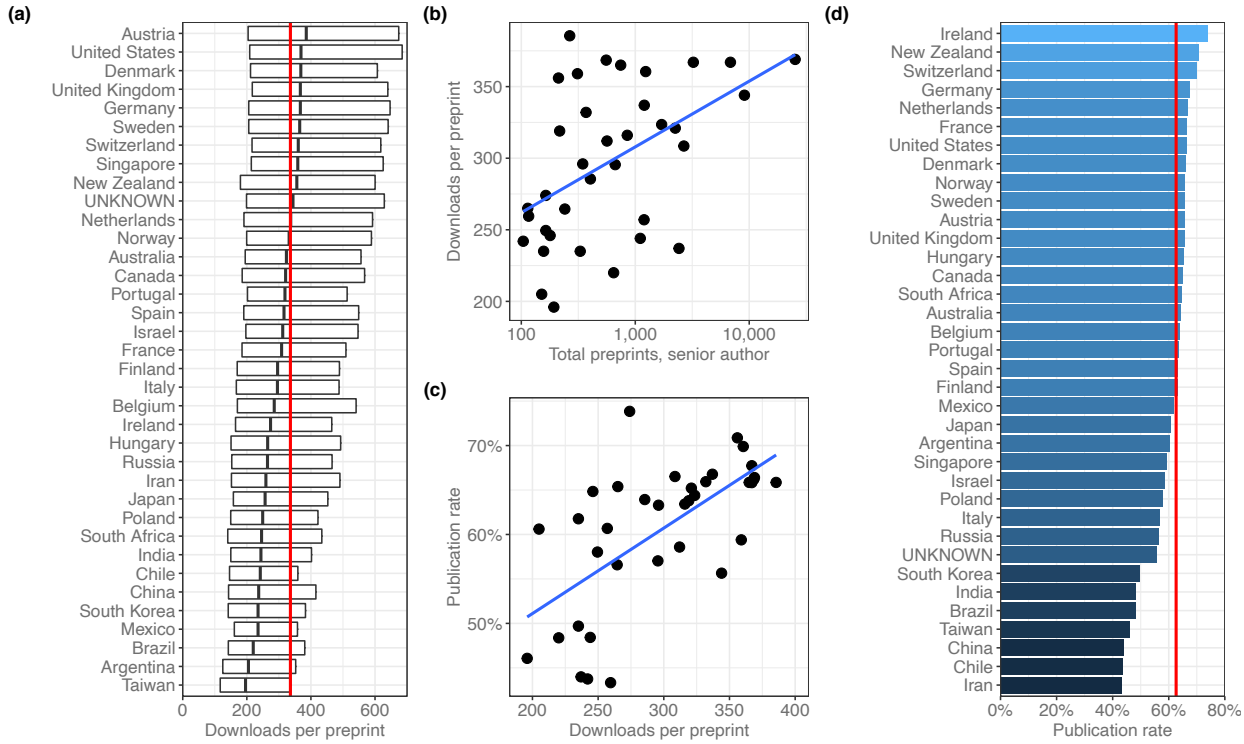

Figure 4. Preprint outcomes. All panels include countries with at least 100 senior-author preprints. **(a)** A box plot indicating the number of downloads per preprint for each country. The dark line in the middle of the box indicates the median, and the ends of each box indicate the first and third quartiles, respectively. "Whiskers" and outliers were omitted from this plot for clarity. The red line indicates the overall median. **(b)** A plot showing the relationship between total preprints and downloads. Each point represents a single country. The x-axis indicates the total number of senior-author preprints attributed to the country. The y-axis indicates the median number of downloads for those preprints. **(c)** A plot showing the relationship between downloads and publication rate. Each point represents a single country. The x-axis indicates the median number of downloads for all preprints listing a senior author affiliated with that country. The y-axis indicates the proportion of preprints posted before 2019 that have been published. **(d)** A bar plot indicating the proportion of preprints posted before 2019 that are now flagged as "published" on the bioRxiv website. The x-axis (and color scale) indicates the proportion, and the y-axis lists each country. The red line indicates the overall publication rate.

After evaluating the country-level publication rates, we examined which journals were publishing these preprints and whether there were any meaningful country-level patterns (**Fig. 5**). We quantified how many senior-author preprints from each country were published in each journal and used the χ^2 test (with Yates's correction for continuity) to examine whether a journal published

a disproportionate number of preprints from a given country, based on how many preprints from that country were published overall. To minimize the effect of journals with differing review times, we limited the analysis to preprints posted before 2019, resulting in a total of 23,102 published preprints.

After controlling the false-discovery rate using the Benjamini–Hochberg procedure, we found 53 significant links between journals and countries (**Fig. 5a**; including journal–country links with at least 15 preprints). Nine countries had links to journals that published a disproportionate number of their preprints, but the United States had far more than any other country. 30 of the 53 significant links were between a journal and the United States: The U.S. is listed as the senior author on 39.6% of published preprints, but accounts for 69.6% of all bioRxiv preprints published in *Cell*, 67.7% of preprints published in *Science*, and 58.5% of those published in *Proceedings of the National Academy of Sciences* (*PNAS*) (**Fig. 5b**).

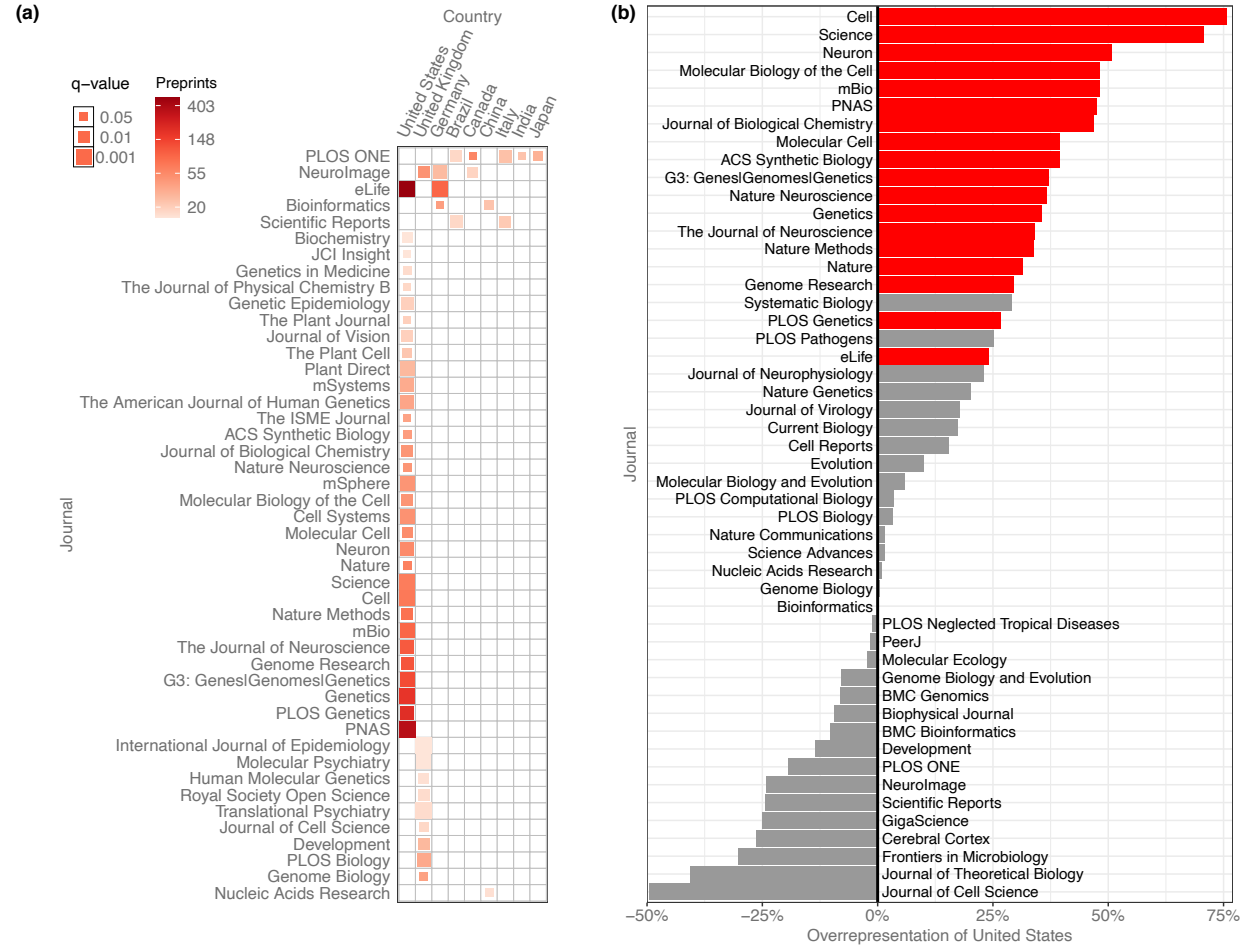

Figure 5. Overrepresentation of U.S. preprints. (a) A heat map indicating all disproportionately strong ($q < 0.05$) links between countries and journals, for journals that have published at least 15 preprints from that country. Columns each represent a single country, and rows each represent a single journal. Colors indicate the raw number of preprints published, and the size of each square indicates the statistical significance of that link—larger squares represent smaller q-values. **(b)** A bar plot indicating the degree to which U.S. preprints are over- or under-represented in a journal's published bioRxiv preprints. The y-axis lists all the

journals that would be expected to have published at least 30 preprints with a U.S. senior author, based on the proportion of published preprints from the U.S. and the total number of preprints published by that journal. The x-axis indicates the overrepresentation of U.S. preprints compared to the expected number: For example, a value of "0%" would indicate the journal published the same proportion of U.S. preprints as all journals combined. A value of "100%" would indicate the journal published twice as many U.S. preprints as expected, based on the overall representation of the U.S. among published preprints. The red bars indicate which of these relationships were significant using Benjamini–Hochberg-adjusted results from χ^2 tests. All results are available in **Supplementary Table 6**.

Methods

Ethical statement. This study was submitted to the University of Minnesota Institutional Review Board (study #00008793), which determined the work did not qualify as human subjects research and did not require IRB oversight.

Preprint metadata. We used existing data from the Rxivist web crawler (Abdill and Blekhman 2019c) to build a list of URLs for every preprint on bioRxiv.org. We then used this list as the input for a new tool that collects author data: We recorded a separate entry for each author of each preprint, and stored name, email address, affiliation, ORCID identifier, and the date of the most recent version of the preprint that has been indexed in the Rxivist database. While the original web crawler performs author consolidation during the paper index process (i.e. "Does this new paper have any authors we already recognize?"), this new tool creates a new entry for each preprint; we make no connections for authors across preprints in this analysis, and infer author country separately for every author of every paper. It is also important to note that for longitudinal analyses of preprint trends, each preprint is associated with the date on *its most recent version*, which means a paper first posted in 2015, but then revised in 2017, would be listed in 2017. The final version of the preprint metadata was collected in the final weeks of January 2020—because preprints were filtered using the most recent known date, those posted before 2020, but revised in the first month of 2020, were not included in the analysis. In addition, 95 preprints were excluded because the bioRxiv website repeatedly returned errors when we tried to collect the metadata, leaving a total of 67,885 preprints in the analysis. Of these, there were 2,409 manuscripts (3.6%) for which we were unable to scrape affiliation data for at least one author, including 137 preprints with no affiliation information for any author. These preprints were included in the analysis, but all missing affiliation strings were placed in the "unknown" institution classification.

bioRxiv maintains an application programmatic interface (API) that provides machine-readable data about their holdings. However, the information it exposes about authors and their affiliations is not as complete as the information available from the website itself, and only the corresponding author's institutional affiliation is included ("bioRxiv API (beta)" n.d.). Therefore, we used the more complete data in the Rxivist database (Abdill and Blekhman 2019b), which includes affiliations for all authors.

All data on published preprints was pulled directly from bioRxiv. However, it is also possible, if not likely, that the publication of many preprints goes undetected by its system. Fraser et al. (2020)

developed a method of searching for published preprints in Scopus and Crossref databases and found most had already been picked up by bioRxiv's detection process, though bioRxiv states that preprints published with new titles or authors can go undetected ("About bioRxiv" n.d.), and preliminary data suggests this may affect thousands of preprints (Abdill and Blekhman 2019b). How these effects differ by country of origin remains unclear—perhaps authors from some countries are more likely to have their titles changed by journal editors, for example—but bias at the country level may also be more pronounced for other reasons. The assignment of Digital Object Identifiers (DOIs) to papers provides a useful proxy for participation in the "western" publishing system. Each published bioRxiv preprint is listed with the DOI of its published version, but DOI assignment is not yet universally adopted. Boudry and Chartron (2017) examined papers from 2015 indexed by PubMed and found DOI assignment varied widely based on the country of the publisher. 96% of publications in Germany had a DOI, for example, plus 98% of U.K. publications and more than 99% of Brazilian publications. However, only 31% of papers published in China had DOIs, and just 2% (33 out of 1582) of papers published in Russia. Boudry and Chartron (2017) included the 50 most productive countries in their analysis; of these, we found no relationship between a country's preprint publication rate and the rate at which publishers in that country assigned DOIs (Pearson's $r=0.168$, $p=0.245$).

Attribution of preprints. Throughout the analysis, we define the "senior author" for each preprint as the author appearing last in the author list. In addition to being a longstanding practice in biomedical literature (Riesenberg and Lundberg 1990; Buehring, Buehring, and Gerard 2007), a 2003 study found that 91 percent of publications indicated a corresponding author that was in the first- or last-author position (Mattsson, Sundberg, and Laget 2011). Among the 56,002 preprints for which the country was known for the first and last author, 7,239 (12.9%) preprints included a first author associated with a different country than the senior author.

When examining international collaboration, we also considered whether more nuanced methods of distributing credit would be more informative. Our primary approach—assigning each preprint to the one country appearing in the senior author spot—is considered *straight counting* (Gauffriau et al. 2008). We repeated the process using *complete-normalized counting* (**Supplementary Table 7**), which splits a single credit among all authors of a preprint. So, for a preprint with 10 authors, if six authors are affiliated with an institution in the United Kingdom, the U.K. would receive 0.6 "credits" for that preprint. We found the complete-normalized preprint counts to be almost identical to the counts distributed based on straight counting (Pearson's $r=0.9998$, $p=3.27\times10^{-306}$). While there are numerous proposals for proportioning differing levels of recognition to authors at different positions in the author list (e.g. Hagen 2013; Kim and Diesner 2015), the close link between the complete-normalized count and the count based on senior authorship indicates that senior authors are at least an accurate proxy for the overall number of individual authors, at the country level.

When computing the average authors per paper, the harmonic mean is used to capture the average "contribution" of an author, as in Glänzel and Schubert (2005)—in short, this shows that authors were responsible for about one-third of a preprint in 2014, but less than one-fourth of a preprint as of 2019.

Data collection and management. All bioRxiv metadata was collected in a relational PostgreSQL database (PostgreSQL Global Development Group 2017). The main table, "article_authors," recorded one entry for each author of each preprint, with the author-level metadata described above. Another table associated each unique affiliation string with an inferred institution (see **Institutional affiliation assignment** below), with other tables linking institutions to countries and preprints to publications. (See **Supplemental materials** for a full description of the database schema.) Analysis was performed by querying the database for different combinations of data and outputting them into CSV files for analysis in R (R Core Team 2019). For example, data on "authors per preprint" was collected by associating all the unique preprints in the "article_authors" table with a count of the number of entries in the table for that preprint. Similar consolidation was done at many other levels as well—for example, since each author is associated with an affiliation string, and each affiliation string is associated with an institution, and each institution is associated with a country, we can build queries to evaluate properties of preprints grouped by country.

Contributor countries. The analysis described in the "Collaboration" section measured correlations between three country-level descriptors, calculated for all countries that contributed to more than 30 international preprints:

1. **International collaborations.** The total number of international preprints including at least one author from that country.
2. **International collaboration rate.** Of all preprints listing an author from that country, the proportion of them that includes at least one author from another country.
3. **International senior-author rate.** Of all the international collaborations associated with a country, the proportion of them for which that country was listed as the senior author.

We examined disproportionate links between contributor countries and senior-author countries by performing one-tailed Fisher's exact tests between each contributor country and each senior-author country, to test the null hypothesis that there is no association between the classifications "preprints with an author from the contributor country" and preprints with a senior author from the senior-author country." To minimize the effect of partnerships between individual researchers affecting country-level analysis, the senior-author country list included only countries with at least 25 senior-author preprints that include a contributor country, and we only evaluated links between contributor countries and senior-author countries that included at least 5 preprints.

BioRxiv adoption. When evaluating bioRxiv participation, we corrected for overall research output, as documented by SCImago Journal & Country Rank portal ("Scimago Journal & Country Rank" n.d.) articles, conference papers, and reviews in Scopus-indexed journals ("SJR - Help" n.d., "Scimago Journal & Country Rank" n.d.) This is not an ideal reference: The SCImago data does not include 2019 outputs yet and is not specific to life sciences research. However, we used this because it had consistent data for all countries in our dataset; assuming there were no dramatic changes in overall output in 2019, the inclusion of more years should not change the bioRxiv adoption score. Another shortcoming of combining data SCImago and the Research Organization Registry (see below) is that they use different criteria for the inclusion of separate states. In most cases, SCImago provides more specific distinctions than ROR: For example, Puerto Rico is listed separately from the United States in the SCImago dataset, but not in the ROR dataset. We did not alter these distinctions—as a result, nations with disputed or complex borders may have slightly inflated bioRxiv adoption scores. For example, preprints attributed to institutions in Hong Kong

are counted in the total for China, but the 85,146 citable documents from Hong Kong in the SCImago dataset are not included in the China total.

Visualization. All figures were made with R and the ggplot2 package (Wickham 2016), with colors from the RColorBrewer package (Neuwirth 2014; Woodruff and Brewer 2017). The world map in Figure 1 was generated using the rworldmap package (South 2011). Code to reproduce all figures is available on GitHub (https://github.com/blekhmanlab/biorxiv_countries).

Institutional affiliation assignment. We used the Research Organization Registry (ROR) API to translate bioRxiv affiliation strings into canonical institution identities (Research Organization Registry 2019). We launched a local copy of the database using their included Docker configuration and linked it to our web crawler's container, to allow the two applications to communicate. We then pulled a list of every unique affiliation string observed on bioRxiv and submitted them to the ROR API. We used the response's "chosen" field, indicating the ROR application's confidence in the assignment, to dictate whether the assignment was recorded. Any affiliation strings that did not have an assigned result were put into a separate "unknown" category. As with any study of this kind, we are limited by the quality of available metadata. Though we are able to efficiently scrape data from bioRxiv, data provided by authors can be unreliable or ambiguous. There are 465 preprints, for example, in which multiple or all authors on a paper are listed with the same ORCID, ostensibly a unique personal identifier, including seven preprints for which 30 or more authors were listed under the same ORCID. We are also limited by the content of the ROR system: Though there are tens of thousands of institutions in the dataset ("About" 2020) and its basis, the Global Research Identifier Database (GRID), has extensive coverage around the world ("Statistics" n.d.), the translation of affiliation strings is likely more effective for regions that have more extensive coverage.

Country-level accuracy of ROR assignments. Across 67,885 total preprints, we found 488,660 total author entries (one for each author of each preprint). These entries each included one of 136,456 distinct affiliation strings, each of which was processed by the ROR API. We wanted to measure the accuracy of these assignments. First, we took a random sample of 100 distinct affiliation strings and found the institution-level error rate to be 9 percent. This yielded a sample size of 488 affiliation strings at p=0.05, with 80 percent power to detect an improvement in error rate from 0.09 to 0.045 (Whitley and Ball 2002). Of the output recorded directly from the ROR API, we found 61 out of 488 (12.5%) sampled affiliations had been assigned to the wrong institution, and 38 of 488 (7.8%) had been assigned to the wrong country (**Table 5**). To improve these rates, we made the manual adjustments described below.

We evaluated the affiliation strings classified in the "unknown" category. We did this by first examining affiliation strings associated with ten or more authors. (The highest number of authors listing an "unknown" affiliation string was 364, but the median was 1, and the mean was 2.8.) For these affiliation strings, we broke each string into a list of comma-separated elements. We then attempted to match the last element from each string list to the ROR institution list. For affiliation strings where this was unsuccessful, we then identified each institution from the affiliation string by hand. Several shortcuts were used to do this, including: identifying institutions at other positions within affiliation strings, rather than the end (e.g. the affiliation string "Université de Tours, EA2106, Biomolécules et Biotechnologies Végétales, Tours" was assigned the institution

"Université de Tours"); defining acronyms present in affiliation strings and matching them to ROR-listed institutions (e.g. "Veterans Affairs Connecticut Healthcare System" and "VA Connecticut Healthcare System" affiliation strings should match to the same institution); looking up the locations of specific institutes (e.g. the Athinoula A. Martinos Center for Biomedical Imaging, which is at the Massachusetts Institute of Technology); and accounting for variations in institution listing (e.g. the Adam Mickiewicz University and the Adam Mickiewicz University in Poznań refer to the same institution).

We were able to find classifications for some of them, but there were also corrections made that placed more affiliations into the "unknown" category—there is an ROR institution called "Computer Science Department," for example, that contained spurious assignments. Prior to correction, 23,158 (17%) distinct affiliation strings were categorized as "unknown," associated with 71,947 authors. Manual corrections reduced this to 20,299 affiliation strings associated with 49,447 authors, but other corrections moved incorrectly assigned affiliation strings *into* the "unknown" category, so there were ultimately 23,754 affiliation strings in the "unknown" category, associated with 66,544 author entries. While our corrections increased the number of "unknown" affiliations by 596, the number of author entries associated with those affiliations decreased by 5,403.

There were also corrections made to existing institutional assignments, which were important to evaluate because institutional assignments were used to make the country-level inferences about author location. It appears the API struggles with institutions that are commonly expressed as acronyms—affiliation strings including "MIT," for example, was sometimes incorrectly coded not as "Massachusetts Institute of Technology" in the United States, but as "Manukau Institute of Technology" in New Zealand, even when other clues within the affiliation string indicated it was the former. Other affiliation strings were more broadly opaque— "Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB," for example. A full list of manual edits is included in the "manual_edits.sql" and "unknown_corrections.csv" files.

In total, 9,378 institutional assignments were corrected or added, affecting 44,619 author entries. After the corrections were made, we repeated the sampling and evaluation process. We found precision at the institution level increased from 87.5% to 96.1%, an improvement of 8.6% ± 3.4% (**Table 2**). Precision at the country level went from 92.2% to 96.5%, an improvement of 4.3% ± 2.9%.

	Uncorrected	First correction	Improvement
Correct institution	87.5% ± 2.9%	96.1% ± 1.7%	8.6% ± 3.4%
Correct country	92.2% ± 2.4%	96.5% ± 1.6%	4.3% ± 2.9%

Table 2. Precision of institutional assignment to affiliation strings. Margins of error use a 95 percent confidence interval.

Next, we evaluated the country-level effects of our corrections by generating an approximation of precision and recall. An affiliation string that remained unchanged after correction was counted as a "true positive," a string that was removed from a country was counted as a "false positive," and a string that was added to a country by a correction was counted as a "false negative." (We counted

affiliation strings, rather than the total authors *associated* with those strings, to focus on the ROR API's capability to assign institutions regardless of the popularity of a given affiliation.)

Because our corrected dataset was used as the ground truth in this evaluation, countries with low precision reflect those with many corrections assigning affiliation strings *out* of that country, and countries with low recall reflect those that picked up many affiliation strings in the correction.

The country with the lowest recall was the Netherlands (85.1%), which had 2,425 affiliations remain after corrections but also picked up 425 additional ones (**Supplementary Table 8**), mostly corrections for affiliations linked to Radboud University and Wageningen University that were either linked to China or placed in the unknown category. Qatar had a similar recall; it maintained the 102 affiliation strings that were initially assigned but gained 15 more from moving affiliations related to "Weill Cornell Medicine in Qatar" out of the unknown category.

Discussion

Our study represents the first comprehensive, country-level analysis of bioRxiv preprint publication and outcomes. While previous studies have split up papers into "USA" and "everyone else" categories in biology (Fraser et al. 2020) and astrophysics (Schwarz and Kennicutt 2004), our results provide a broad picture of worldwide participation in the largest preprint server in biology. We show that the United States is by far the most highly represented country by number of preprints, followed distantly by the United Kingdom and Germany.

By adjusting preprint counts by each country's overall scientific output, we were able to develop a "bioRxiv adoption" score (**Fig. 2**). The United States and the United Kingdom again had the highest scores, while countries such as Turkey, Iran and Malaysia were underrepresented even after accounting for their comparatively low scientific output. Studies have found countries take very different approaches to research communication. Large-scale differences frequently deal with balancing the sharing of research findings with the protection of commercial interests (Walsh and Huang 2014; Caulfield, Harmon, and Joly 2012; Azmi and Alavi 2013), but open science advocates have argued for years that there can be no "one size fits all" approach to preprints and open-access publication because of the dramatically different country-level incentive structures, cultural practices, and access to resources, funding and infrastructure (Debat and Babini 2020; "Systemic Reforms and Further Consultation Needed to Make Plan S a Success" 2018; Becerril-García 2019; Mukunth 2019). Further research is required to determine what drives certain countries to use bioRxiv and other preprint servers—what incentives are present for biologists in Finland but not Greece, for example—but the current results make it clear that those reading bioRxiv (or soliciting submissions from the platform) are reviewing a biased sample of worldwide scholarship.

There are two findings that may be particularly informative about the state of open science in biology. First, we present evidence of contributor countries—countries from which authors appear almost exclusively in non-senior roles on preprints led by authors from more prolific countries (**Fig. 3**). While there are many reasons these dynamics could arise, it is worth noting that the current corpus of bioRxiv preprints contains the same familiar disparities observed in published literature (Mammides et al. 2016; Burgman, Jarrad, and Main 2015; Wong et al. 2014; González-

Alcaide et al. 2017). Critically, we found the three characteristics of contributor countries (low international collaboration *count*, high international collaboration *rate*, low international senior author rate) are strongly correlated with each other (**Fig. 3** and **Supplementary Table 9**). When looking at international collaboration using pairwise combinations of these three measurements, countries fall along tidy gradients—which means not only that they can be used to delineate properties of contributor countries, but that if a country fits even one of these criteria, they are more likely to fit the other two as well.

Second, we found numerous country-level differences in preprint outcomes. Differences in downloads per paper have the most straightforward interpretation: If one of the goals of preprinting one's work is to solicit feedback from the community (Sarabipour et al. 2019; Sever et al. 2019), more "reads" of a preprint may represent an increased probability of receiving helpful feedback, or at least increased exposure to other researchers in the field. The sources and implications of these disparities are an open question: What is the effect of Dutch preprints receiving a median of 368.5 downloads per preprint, while Brazilian preprints receive 220? Do preprint authors from the most-downloaded countries (mostly in western Europe) have broader social-media reach than authors in low-download countries such as Chile, Argentina and Taiwan? Are preprints from some countries more likely to be included in newsletters and search alerts? What role does language play? The observed correlation between country-level publication rate and median downloads per paper also reinforces the assertion that preprints from some countries generally fare better, and the observed differences are not solely due to artifacts in bibliometric data. The average preprint from the United States is downloaded 369 times and has a 66.4 percent chance of being published, while South Korean preprints receive 36 percent fewer downloads and have a 25 percent reduction in publication rate. We also found some journals had particularly strong affinities for preprints from some countries over others: Even when accounting for differing publication rates across countries, we found dozens of journal–country links that disproportionately favored countries such as the United States and United Kingdom. While it's possible some of these relationships are coincidental, this finding demonstrates that journals can embrace preprints while still perpetuating some of the imbalances that preprints could be theoretically alleviating.

Our study has several limitations. First, bioRxiv is not the only preprint server hosting biology preprints. For example, arXiv's "Quantitative Biology" category (https://arxiv.org/archive/q-bio) held 18,024 preprints at the end of 2019 ("arXiv Submission Rate Statistics" 2020), and repositories such as Indonesia's INA-Rxiv (https://osf.io/preprints/inarxiv/) hold multidisciplinary collections of country-specific preprints. We chose to focus on bioRxiv for several reasons: Primarily, bioRxiv is the preprint server most broadly integrated into the traditional publishing system (see **Introduction**) (Barsh et al. 2016; Vence 2017; Eisen 2019). In addition, bioRxiv currently holds the largest collection of biology preprints, with metadata available in a format we were already equipped to ingest (Abdill and Blekhman 2019c). Analyzing data from only a single repository also avoids the issue of different websites holding metadata that is mismatched or collected in different ways. Comparing publication rates between repositories would also be difficult, particularly because bioRxiv is one of the few with an automated method for detecting when a preprint has been published. Second, this "worldwide" analysis of preprints is explicitly biased toward English-language publishing. BioRxiv accepts submissions only in English, and the primary motivation for this work was the attention being paid to bioRxiv by organizations based mostly in the U.S. and western Europe. In addition, bibliometrics databases such as Scopus and

Web of Science have well-documented biases in favor of English-language publications (Mongeon and Paul-Hus 2016; Archambault et al. 2006; de Moya-Anegón et al. 2007), which could have an effect on observed publication rates and the bioRxiv adoption scores that depend on scientific output derived from Scopus.

In summary, we find country-level participation on bioRxiv differs significantly from existing patterns in scientific publishing. Preprint outcomes reflect particularly large differences between countries: Comparatively wealthy countries in Europe and North America post more preprints, which are downloaded more frequently, published more consistently, and favored by the largest and most well-known journals in biology. While there are many potential explanations for these dynamics, the quantification of these patterns may help stakeholders make more informed decisions about how they read, write and publish preprints in the future.

Acknowledgements

We thank Alex D. Wade (Chan Zuckerberg Initiative) for his insights on author disambiguation and the members of the Blekhman lab for helpful discussions. We also thank the Research Organization Registry community for curating an extensive, freely available dataset on research institutions around the world.

Funding and competing interests

RB is supported by the National Institutes of General Medicine (R35-GM128716) and a McKnight Land-Grant Professorship from the University of Minnesota. The funders had no role in study design, data collection and analysis, or preparation of the manuscript. RA is a volunteer ambassador for ASAPbio, a nonprofit preprint advocacy organization that is affiliated with Review Commons.

Data availability

There are several online repositories linked to this study:

- The code for the web crawler used to collect the preprint data is available on GitHub at https://github.com/blekhmanlab/biorxiv_countries
- All data used for the analyses is contained in a database snapshot available, along with data and R code to reproduce all figures, via Zenodo at https://doi.org/10.5281/zenodo.3762815
- Supplementary tables are available in CSV format in the same Zenodo repository.
- Supplementary figures, and legends for the supplementary tables, are available in a separate file attached to this manuscript.

References

Abdill, Richard J., and Ran Blekhman. 2019a. “Complete Rxivist Dataset of Scraped bioRxiv Data.” Zenodo. https://doi.org/10.5281/ZENODO.2529922.

———. 2019b. “Tracking the Popularity and Outcomes of All bioRxiv Preprints.” *eLife* 8 (April): e45133.

———. 2019c. “Rxivist.org: Sorting Biology Preprints Using Social Media and Readership Metrics.” *PLOS Biology* 17 (5): e3000269.

“About.” 2020. Research Organization Registry. 2020. https://ror.org/about/.

"About bioRxiv." n.d. bioRxiv. Accessed March 19, 2020. https://www.biorxiv.org/about-biorxiv.

Adams, James D., Grant C. Black, J. Roger Clemmons, and Paula E. Stephan. 2005. "Scientific Teams and Institutional Collaborations: Evidence from U.S. Universities, 1981–1999." *Research Policy* 34 (3): 259–85.

Akre, Olof, Francesco Barone-Adesi, Andreas Pettersson, Neil Pearce, Franco Merletti, and Lorenzo Richiardi. 2011. "Differences in Citation Rates by Country of Origin for Papers Published in Top-Ranked Medical Journals: Do They Reflect Inequalities in Access to Publication?" *Journal of Epidemiology and Community Health* 65 (2): 119–23.

Archambault, Éric, Étienne Vignola-Gagné, Grégoire Côté, Vincent Larivière, and Yves Gingrasb. 2006. "Benchmarking Scientific Output in the Social Sciences and Humanities: The Limits of Existing Databases." *Scientometrics* 68 (3): 329–42.

"arXiv Submission Rate Statistics." 2020. arXiv. 2020. https://arxiv.org/help/stats/2019_by_area/index.

Azmi, Ida Madieha, and Rokiah Alavi. 2013. "Patents and the Practice of Open Science among Government Research Institutes in Malaysia: The Case of Malaysian Rubber Board." *World Patent Information* 35 (3): 235–42.

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.

Becerril-García, Arianna. 2019. "AmeliCA vs Plan S: Same Target, Two Different Strategies to Achieve Open Access." *AmeliCA (blog)*. http://amelica.org/index.php/en/2019/02/10/amelica-vs-plan-s-same-target-two-different-strategies-to-achieve-open-access/.

Berg, Jeremy M., Needhi Bhalla, Philip E. Bourne, Martin Chalfie, David G. Drubin, James S. Fraser, Carol W. Greider, et al. 2016. "Preprints for the Life Sciences." *Science* 352 (6288): 899–901.

"bioRxiv API (beta)." n.d. Accessed January 16, 2020. http://api.biorxiv.org/.

Bordons, María, Javier Aparicio, and Rodrigo Costas. 2013. "Heterogeneity of Collaboration and Its Relationship with Research Impact in a Biomedical Field." *Scientometrics* 96 (2): 443–66.

Boudry, Christophe, and Ghislaine Chartron. 2017. "Availability of Digital Object Identifiers in Publications Archived by PubMed." *Scientometrics* 110 (3): 1453–69.

Buehring, Gertrude Case, Jessica E. Buehring, and Patrick D. Gerard. 2007. "Lost in Citation: Vanishing Visibility of Senior Authors." *Scientometrics* 72 (3): 459–68.

Burgman, Mark, Frith Jarrad, and Ellen Main. 2015. "Decreasing Geographic Bias in Conservation Biology." *Conservation Biology* 29 (5): 1255–56.

Caulfield, Timothy, Shawn He Harmon, and Yann Joly. 2012. "Open Science versus Commercialization: A Modern Research Conflict?" *Genome Medicine* 4 (2): 17.

Debat, Humberto, and Dominique Babini. 2020. "Plan S in Latin America: A Precautionary Note." *Scholarly and Research Communication* 11 (1): 12.

Eisen, Michael. 2019. "Peer Review: New Initiatives to Enhance the Value of eLife's Process." *eLife*, November. https://elifesciences.org/inside-elife/e9091cea/peer-review-new-initiatives-to-enhance-the-value-of-elife-s-process.

Fraser, Nicholas, Fakhri Momeni, Philipp Mayr, and Isabella Peters. 2020. "The Relationship between bioRxiv Preprints, Citations and Altmetrics." *Quantitative Science Studies*, April, 1–39.

Fu, Darwin Y., and Jacob J. Hughey. 2019. "Releasing a Preprint Is Associated with More Attention and Citations for the Peer-Reviewed Article." *eLife* 8 (December): e52646.

Gauffriau, Marianne, Peder Olesen Larsen, Isabelle Maye, Anne Roulin-Perriard, and Markus von Ins. 2008. "Comparisons of Results of Publication Counting Using Different Methods." *Scientometrics* 77 (1): 147–76.

Glänzel, Wolfgang, and András Schubert. 2005. "Analysing Scientific Networks Through Co-Authorship." In *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*, edited by Henk F. Moed, Wolfgang Glänzel, and Ulrich Schmoch, 257–76. Dordrecht: Springer Netherlands.

González-Alcaide, Gregorio, Jinseo Park, Charles Huamaní, and José M. Ramos. 2017. "Dominance and Leadership in Research Activities: Collaboration between Countries of Differing Human Development Is Reflected through Authorship Order and Designation as Corresponding Authors in Scientific Publications." *PLOS One* 12 (8): e0182513.

Hagen, Nils T. 2013. "Harmonic Coauthor Credit: A Parsimonious Quantification of the Byline Hierarchy." *Journal of Informetrics* 7 (4): 784–91.

Kim, Jinseok, and Jana Diesner. 2015. "Coauthorship Networks: A Directed Network Approach Considering the Order and Number of Coauthors." *Journal of the Association for Information Science and Technology* 66 (12): 2685–96.

714 Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. "Bias in Peer Review." *Journal of the*
715 *American Society for Information Science and Technology* 64 (1): 2–17.
716 Mammides, Christos, Uromi M. Goodale, Richard T. Corlett, Jin Chen, Kamaljit S. Bawa, Hetal Hariya, Frith
717 Jarrad, et al. 2016. "Increasing Geographic Diversity in the International Conservation Literature: A Stalled
718 Process?" *Biological Conservation* 198 (June): 78–83.
719 Mattsson, Pauline, Carl Johan Sundberg, and Patrice Laget. 2011. "Is Correspondence Reflected in the Author
720 Position? A Bibliometric Study of the Relation between Corresponding Author and Byline Position."
721 *Scientometrics* 87 (1): 99–105.
722 Mongeon, Philippe, and Adèle Paul-Hus. 2016. "The Journal Coverage of Web of Science and Scopus: A
723 Comparative Analysis." *Scientometrics* 106 (1): 213–28.
724 Moya-Anegón, Félix de, Zaida Chinchilla-Rodríguez, Benjamín Vargas-Quesada, Elena Corera-Álvarez, Francisco
725 José Muñoz-Fernández, Antonio González-Molina, and Victor Herrero-Solana. 2007. "Coverage Analysis of
726 Scopus: A Journal Metric Approach." *Scientometrics* 73 (1): 53–78.
727 Mukunth, Vasudevan. 2019. "India Will Skip Plan S, Focus on National Efforts in Science Publishing." The Wire:
728 Science. October 26, 2019. https://science.thewire.in/the-sciences/plan-s-open-access-scientific-publishing-
729 article-processing-charge-insa-k-vijayraghavan/.
730 Narock, Tom, and Evan B. Goldstein. 2019. "Quantifying the Growth of Preprint Services Hosted by the Center for
731 Open Science." *Publications* 7 (2): 44.
732 Neuwirth, Erich. 2014. "RColorBrewer: ColorBrewer Palettes. R Package Version 1.1-2." *The R Foundation*.
733 https://CRAN.R-project.org/package=RColorBrewer.
734 Okike, Kanu, Mininder S. Kocher, Charles T. Mehlman, James D. Heckman, and Mohit Bhandari. 2008.
735 "Nonscientific Factors Associated with Acceptance for Publication in The Journal of Bone and Joint Surgery
736 (American Volume)." *The Journal of Bone and Joint Surgery. American Volume* 90 (11): 2432–37.
737 Penfold, Naomi C., and Jessica K. Polka. 2020. "Technical and Social Issues Influencing the Adoption of Preprints
738 in the Life Sciences." *PLoS Genetics* 16 (4): e1008565.
739 PostgreSQL Global Development Group. 2017. *PostgreSQL* (version 9.6.6). https://www.postgresql.org.
740 R Core Team. 2019. *R: A Language and Environment for Statistical Computing* (version 3.6.2). http://r-project.org.
741 Research Organization Registry. 2019. *ROR API* (version a3b153c). Github. https://github.com/ror-community/ror-
742 api.
743 Riesenberg, D., and G. D. Lundberg. 1990. "The Order of Authorship: Who's on First?" *JAMA: The Journal of the*
744 *American Medical Association* 264 (14): 1857.
745 Ross, Joseph S., Cary P. Gross, Mayur M. Desai, Yuling Hong, Augustus O. Grant, Stephen R. Daniels, Vladimir C.
746 Hachinski, Raymond J. Gibbons, Timothy J. Gardner, and Harlan M. Krumholz. 2006. "Effect of Blinded Peer
747 Review on Abstract Acceptance." *Journal of the American Medical Association* 295 (14): 1675–80.
748 Saposnik, Gustavo, Bruce Ovbiagele, Stavroula Raptis, Marc Fisher, and S. Claiborne Johnston. 2014. "Effect of
749 English Proficiency and Research Funding on Acceptance of Submitted Articles to Stroke Journal." *Stroke* 45
750 (6): 1862–68.
751 Sarabipour, Sarvenaz, Humberto J. Debat, Edward Emmott, Steven J. Burgess, Benjamin Schwessinger, and Zach
752 Hensel. 2019. "On the Value of Preprints: An Early Career Researcher Perspective." *PLOS Biology* 17 (2):
753 e3000151.
754 Schwarz, Greg J., and Robert C. Kennicutt Jr. 2004. "Demographic and Citation Trends in Astrophysical Journal
755 Papers and Preprints." *arXiv [astro-Ph]*. arXiv. http://arxiv.org/abs/astro-ph/0411275.
756 "Scimago Journal & Country Rank." n.d. Accessed February 15, 2020. https://www.scimagojr.com.
757 Sever, Richard, Ted Roeder, Samantha Hindle, Linda Sussman, Kevin-John Black, Janet Argentine, Wayne Manos,
758 and John R. Inglis. 2019. "bioRxiv: The Preprint Server for Biology." *bioRxiv*, November.
759 https://doi.org/10.1101/833400.
760 "SJR - Help." n.d. Accessed April 7, 2020. https://www.scimagojr.com/help.php.
761 South, Andy. 2011. "Rworldmap: A New R Package for Mapping Global Data." *The R Journal* 3 (1).
762 http://www.econ.upf.edu/~michael/visualdata/RJournal_2011-1_South.pdf.
763 "Statistics." n.d. Global Research Identifier Database. Accessed February 2, 2020. https://www.grid.ac/stats.
764 "Systemic Reforms and Further Consultation Needed to Make Plan S a Success." 2018. European Federation of
765 Academies of Sciences and Humanities. December 12, 2018. https://allea.org/systemic-reforms-and-further-
766 consultation-needed-to-make-plan-s-a-success/.
767 "Trends in Preprints." 2019. PLOS. October 8, 2019. https://plos.org/blog/announcement/trends-in-preprints/.
768 Vence, Tracy. 2017. "Journals Seek out Preprints." *The Scientist*, January 18, 2017. https://www.the-
769 scientist.com/news-opinion/journals-seek-out-preprints-32183.

Walsh, John P., and Hsini Huang. 2014. "Local Context, Academic Entrepreneurship and Open Science: Publication Secrecy and Commercial Activity among Japanese and US Scientists." *Research Policy* 43 (2): 245–60.

Whitley, Elise, and Jonathan Ball. 2002. "Statistics Review 4: Sample Size Calculations." *Critical Care* 6 (4): 335.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wong, Janice C., Kimberly A. Fernandes, Shubarna Amin, Zarnie Lwin, and Monika K. Krzyzanowska. 2014. "Involvement of Low- and Middle-Income Countries in Randomized Controlled Trial Publications in Oncology." *Globalization and Health* 10 (December): 83.

Woodruff, Andy, and Cynthia Brewer. 2017. *Colorbrewer*. Github. https://github.com/axismaps/colorbrewer.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi. 2007. "The Increasing Dominance of Teams in Production of Knowledge." *Science* 316 (5827): 1036–39.