International authorship and collaboration across 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 integrated into conventional avenues of scientific communication, it's critical to understand who is included and who is not. However, little is known about which countries are participating or how they collaborate with each other. Here, we present an analysis of all 67,885 preprints posted on bioRxiv (through 2019) using the first comprehensive dataset of country-level preprint author affiliations. We find the plurality of preprints (39%) come from the United States, and that countries such as the U.S. and U.K. are overrepresented on bioRxiv relative to their overall scientific output, while countries including China, Russia, and Turkey show lower bioRxiv adoption. We describe a set of "contributor countries" including Uganda, Croatia and Thailand that appear almost exclusively as non-senior authors on international collaborations. Lastly, we find multiple journals that publish a disproportionate number of preprints from some countries, a dynamic that almost always benefits manuscripts from the U.S.

Introduction

Biology preprints are being shared online at an unprecedented rate (Narock and Goldstein 2019; Abdill and Blekhman 2019b). Since 2013, more than 87,000 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 May 2020 the journal *eLife* started a program in which bioRxiv preprints submitted to *eLife* would be guaranteed to be sent out for peer review ("New from eLife: Invitation to submit to Preprint Review" 2020). A growing number of programs are incorporating preprints and, in some cases, the use of bioRxiv specifically. However, very little is known about who is benefiting from this attention and how the technical and professional challenges of this new publishing paradigm affect different groups (Penfold and Polka 2020).

Academic publishing has grappled for decades with hard-to-quantify concerns about 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 editorial decisions (Nuñez et al. 2019; Saposnik et al. 2014; Okike et al. 2008; Ross et al. 2006), and there have long been concerns regarding how peer review is influenced by factors such as institutional prestige (Lee et al. 2013). Preprints have been praised as a democratizing influence on scientific communication (Berg et al. 2016), but a 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? 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 used each author's institutional affiliation to summarize country-level participation and outcomes.

Results

Country-level bioRxiv participation over time

We retrieved author data for 67,885 preprints for which the most recent version was posted before 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**). North America, Europe and Australia dominate the top spots (**Figure 1a**): 26,598 manuscripts (39.2%) have a senior author from the United States, followed by 7,151 manuscripts (10.5%) from the United Kingdom (**Figure 1b**), though China (4.1%), Japan (1.9%) and India (1.8%) are the sources of more than 1,200 preprints each (**Table 1**). Brazil, with 704 manuscripts, has the 15th-most preprints and is the first South American country on the list, followed by Argentina (163 preprints) in 32nd place. South Africa (182 preprints) is the first African country on the list, in 29th place, followed by Ethiopia (57 preprints) in 42nd place (**Figure 1–source data 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—34,676 manuscripts (51.1%) include at least one U.S. author (**Figure 1c**).

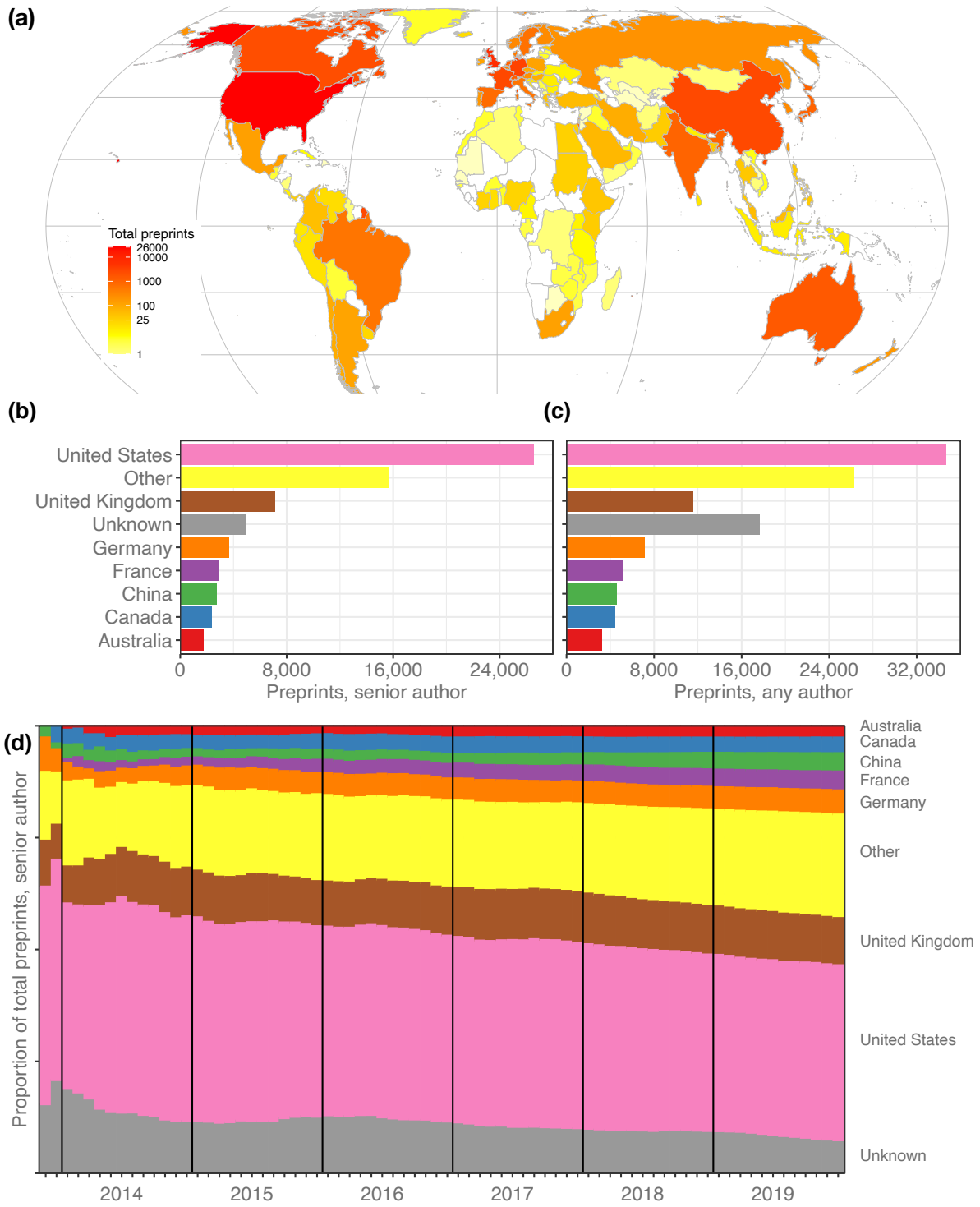

Figure 1. Preprints per country. (a) A heat map indicating the number of preprints per country, based on the institutional affiliation of the senior author. The color coding uses a log scale that splits the full range of preprint counts into six colors. **(b)** The total preprints attributed to the seven 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.

Figure 1–figure supplement 1. Preprint-level collaboration.

Figure 1–figure supplement 2. Preprints with no country assignment.

Figure 1–source data 1. Preprints per country. *supp_table01.csv.*

Figure 1–source data 2. Preprint counting methods at the country level. *supp_table07.csv*.

Over time, the country-level proportions on bioRxiv have remained remarkably stable (**Figure 1d**), even as the number of preprints grew exponentially: At the end of 2015, Germany accounted for 4.7% of bioRxiv's 2,460 manuscripts. At the end of 2019, Germany was responsible for 5.4% of 67,885 preprints. However, the proportion of preprints from countries outside the top seven contributing countries is growing slowly (**Figure 1d**): At the end of 2015, these countries accounted for 19.4 percent of preprints. By the end of 2019, that number had grown to 23.1 percent, when bioRxiv hosted preprints from senior authors affiliated with 136 countries.

Country	Preprints, senior author (proportion)	Preprints, any author (proportion)
United States	26,598 (39.2%)	34,676 (51.1%)
United Kingdom	7,151 (10.5%)	11,578 (17.1%)
(Unknown)	4,985 (7.3%)	17,635 (26.0%)
Germany	3,668 (7.3%)	7,157 (10.5%)
France	2,863 (4.2%)	5,218 (7.7%)
China	2,778 (4.1%)	4,609 (6.8%)
Canada	2,380 (3.5%)	4,409 (6.5%)
Australia	1,755 (2.6%)	3,260 (4.8%)
Switzerland	1,364 (2.0%)	2,779 (4.1%)
Netherlands	1,291 (1.9%)	2,764 (4.1%)
Japan	1,263 (1.9%)	2,287 (3.4%)
India	1,212 (1.8%)	1,769 (2.6%)

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 **Figure 1–source data 1**.

Preprint adoption relative to overall scientific output

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 (**Figure 2a**). The corrected measurement, which we call "bioRxiv adoption," is the proportion of preprints from each country divided by the proportion of worldwide research outputs from that country (see **Methods**). The U.S. posted 26,598 preprints and published about 3.5 million citable documents, for a bioRxiv adoption score of 2.31 (**Figure 2b**). Nine of the 12 countries with adoption scores above 1.0 were from North America and Europe, but Israel has the third-highest score (1.67) based on its 640 preprints. Ethiopia has the 10th-highest bioRxiv adoption (1.08): Though only 57 preprints list a senior author with an affiliation in Ethiopia, the country had a total of 15,820 citable documents published between 2014 and 2019 (**Figure 2–source data 1)**. In other words, 3.6 out of every 1,000 Ethiopian research outputs is on bioRxiv, compared to 7.7 out of every 1,000 citable documents from the United States.

By comparison, some countries are present on bioRxiv at much lower frequencies than would be expected, given their overall participation in scientific publishing (**Figure 2b**): Turkey published 249,086 citable documents from 2014 through 2019 but was the senior author on only 80 preprints, for a bioRxiv adoption score of 0.10. Russia (283 preprints), Iran (123 preprints) and Malaysia (78 preprints) all have adoption scores below 0.18. The largest country with a low adoption score is China (3,176,571 citable documents; 2,778 preprints; bioRxiv adoption=0.26), which published more than 15 percent of the world's citable documents (according to SCImago) but was the source of only 4.1 percent of preprints (**Figure 2a**).

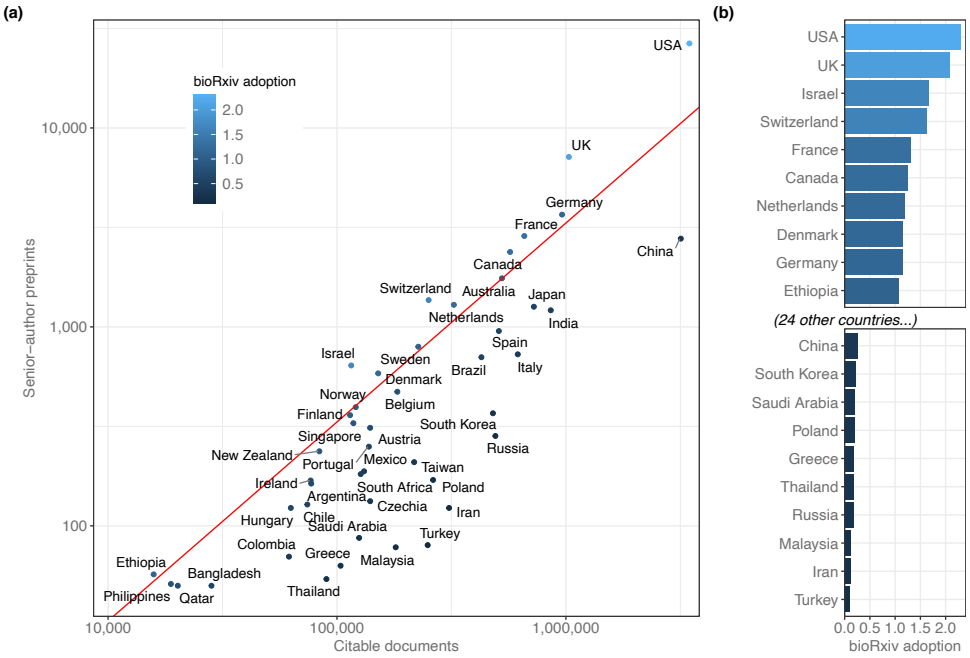

Figure 2. BioRxiv adoption per country. (a) Correlation between two scientific output metrics. Each point is a country; the x-axis (log scale) indicates the total citable documents attributed to that country from 2014–2019, and the y-axis (also log scale) indicates total senior-author preprints attributed to that country overall. The red line demarcates a "bioRxiv adoption" score of 1.0, which indicates that a country's share of bioRxiv preprints is identical to its share of general scholarly outputs. Countries to the left of this line have a bioRxiv adoption score greater than 1.0. 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.) **(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. All panels include only countries with at least 50 preprints.

Figure 2–source data 1. Country productivity and bioRxiv adoption. *supp_table02.csv*.

Patterns and imbalances in international 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 4.56 in 2019 (**Figure 1–figure supplement 1**). The monthly average authors per preprint has increased linearly with time (Pearson's r=0.949, $p=8.73\times10^{-38}$), 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 (**Figure 1–figure supplement 1**), we found that 24,927 preprints (36.7%) included authors from two or more countries; 3,041

preprints (4.5%) were from four or more countries, and one preprint, "Fine-mapping of 150 breast cancer risk regions identifies 178 high confidence target genes," listed 343 authors from 38 countries, the most countries listed on any single preprint. The mean number of countries represented per preprint is 1.612, which has remained fairly stable since 2014 despite steadily growing author lists overall (**Figure 1–figure supplement 1**).

We then looked at countries appearing on at least 50 international preprints to examine basic patterns in international collaboration. We found that a number of countries with comparatively low output contributed almost exclusively to international collaborations: for example, of the 76 preprints that had an author with an affiliation in Vietnam, 73 (96%) had an author from another country. 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 position 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.) 17 countries met these criteria (**Figure 3–figure supplement 1**): for example, of the 84 international preprints that had an author with an affiliation in Uganda, only 5 (6%) had an author from Uganda in the senior author position. This percentage was also less than 12% for Vietnam, Tanzania, Slovakia and Indonesia. By comparison, this percentage was 48.7% for the United States (**Figure 3a**).

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 across all countries with at least 50 international preprints. 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 (**Figure 3–figure supplement 2a**). 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 (**Figure 3–figure supplement 2b**). This indicates that countries with mostly international preprints (**Figure 3b**) also tended to have *fewer* international collaborations (**Figure 3c**) 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 (**Figure 3–figure supplement 2c**), demonstrating that countries that appear mostly on international preprints (**Figure 3b**) are less likely to appear as senior author of those preprints (**Figure 3**). 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).

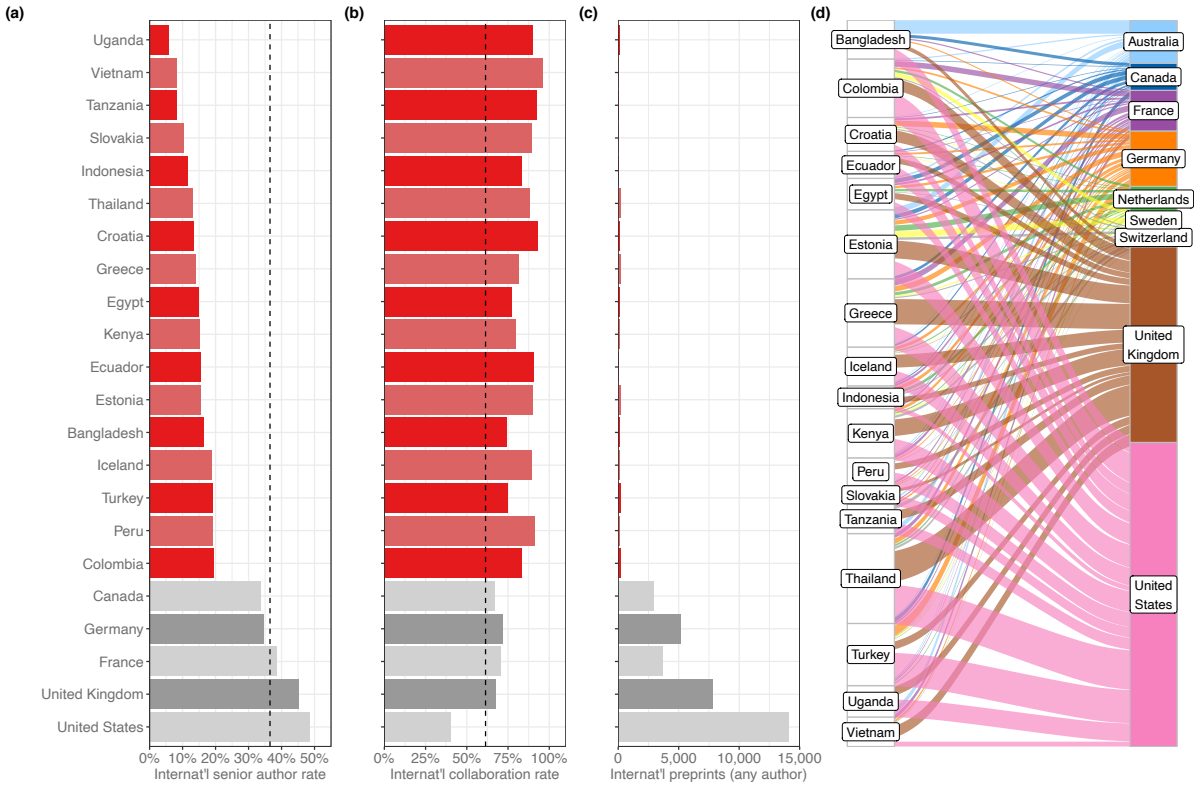

Figure 3. Contributor countries. (a) Bar plot indicating the international senior author rate (y-axis) by country (x-axis)—that is, of all international preprints with a contributor from that country, the percentage of them that include a senior author from that country. All 17 contributor countries are listed in red, with the five countries with the highest senior-author rates (in grey) for comparison. **(b)** A bar plot with the same y-axis as panel (a). 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. **(c)** is a bar plot indicating the total international preprints featuring at least one author from that country (the median value per country is 19). **(d)** On the left are the 17 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. Statistically significant links were found between four combinations of supervising countries and contributors: Australia and Bangladesh (Fisher's exact test, $q=1.01\times10^{-11}$), the U.K. and Thailand ($q=9.54\times10^{-4}$), the U.K. and Greece ($q=6.85\times10^{-3}$), Australia

and Vietnam (q=0.049). All p-values reflect multiple-test correction using the Benjamini–Hochberg procedure.

Figure 3–figure supplement 1. Map of contributor countries.

Figure 3–figure supplement 2. International collaboration correlations.

Figure 3–source data 1. Combinations of senior authors with collaborator countries. *supp_table03.csv*.

Figure 3–source data 2. Links between contributor countries and the senior-author countries they write with. *supp_table04.csv.*

Figure 3–source data 3. International collaboration. *supp_table08.csv.*

After generating a list of preprints with authors from contributor countries, we examined which countries appeared most frequently in the senior author position of those preprints (**Fig. 3d**). Among the 2,133 preprints with an author from a contributor country, 494 (23.2%) had a senior author listing an affiliation in the United States (**Figure 3–source data 1**). The United Kingdom was listed as senior author on the next-most preprints with contributor countries, at 318 (14.9%), followed by Germany (4.2%) and France (3.1%). Given the large differences in preprint authorship between countries, we tested which of these senior-author relationships was disproportionately large. Using Fisher's exact test (see **Methods**), we found four links between contributor countries and senior-author countries that were significant (**Figure 3–source data 2**). The strongest link is between Bangladesh and Australia: Of the 82 international preprints with an author from Bangladesh, 22 list a senior author with an affiliation in Australia. Authors in Vietnam appear with disproportionate frequency on preprints with a senior author in Australia as well (9 of 67 preprints). The other two links are to senior authors in the United Kingdom, with contributing authors from Thailand (50 of 187 preprints) and Greece (41 of 155 preprints).

Differences in preprint downloads and publication rates

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 36 countries with at least 100 senior-author preprints.

When evaluating downloads per preprint, we used only download numbers from each preprint's first six months online, which would capture the majority of downloads for most preprints (Abdill & Blekhman 2019b) while minimizing the effect of the "long tail" of downloads that would be longer for countries that were earlier adopters. Using this measurement, the median number of PDF downloads per preprint is 210 (**Figure 4a**). Among countries with at least 100 preprints, Austria has the highest median downloads per preprint, with 261.5, followed by Germany (235.0), Switzerland (233.0) and the United States (233.0). Argentina has the lowest median, at 138.5 downloads. Next-fewest is Taiwan (142) and a tie at 145 downloads between Brazil and Russia.

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 (**Figure 4b**), and another correlation between median downloads per preprint and each country's publication rate (**Figure 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 (**Figure 4–source data 1**). 49 of Ireland's 67 preprints have been published (73.1%), the highest publication rate (**Figure 4d)**. New Zealand (100 of 142, 70.4%) and Switzerland (505 of 724, 69.8%) had the next-highest rates. China (588 of 1355, 43.4%) had the lowest publication rate, followed by Iran and Taiwan.

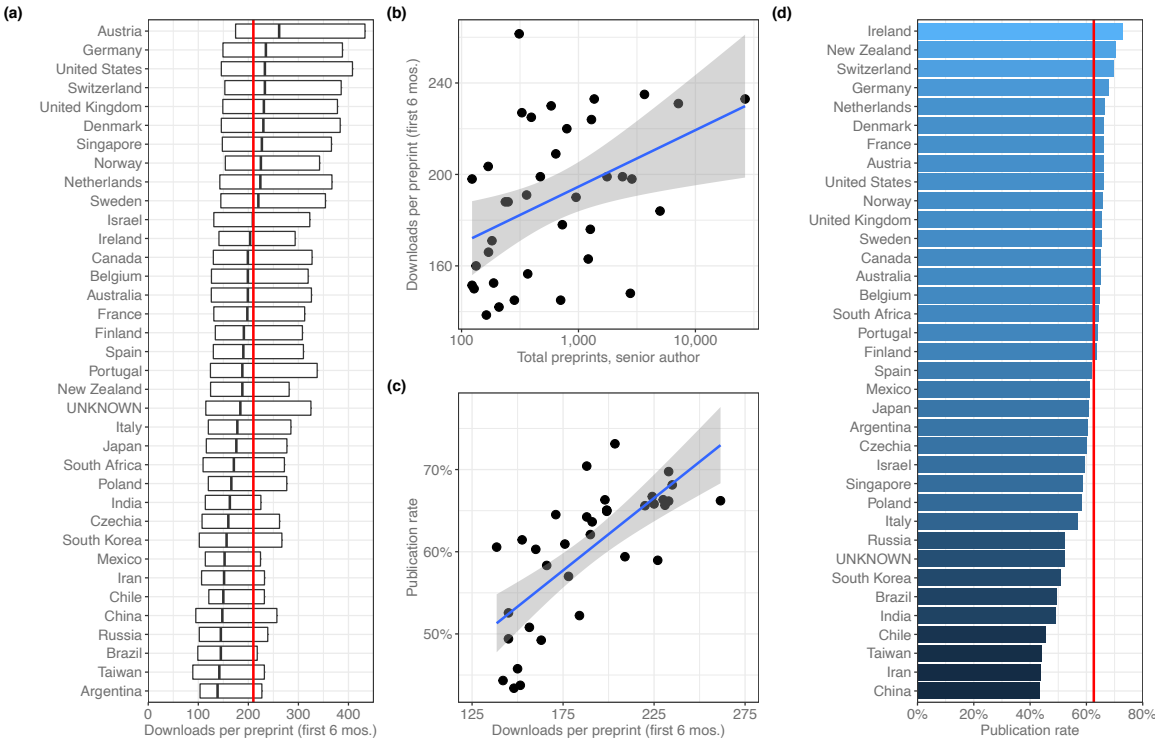

Figure 4. Preprint outcomes. All panels include countries with at least 100 senior-author preprints. **(a)** A box plot indicating the number of downloads per preprint for each country. The dark line in the middle of the box indicates the median, and the ends of each box indicate the first and third 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 (Spearman's ρ=0.485, p=0.00274) 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. The blue lines in panels B and C indicate the best fit to the plotted points, though the Spearman correlation coefficient measures agreement between country ranks in

each category, not the values themselves. **(c)** A plot showing the relationship (Spearman's ρ=0.777, p=2.442×10^{-8}) 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.

Figure 4–source data 1. Published pre-2019 preprints by country.
supp_table05.csv.

Preprint publication patterns between countries and journals

After evaluating the country-level publication rates, we examined which journals were publishing these preprints and whether there were any meaningful country-level patterns (**Figure 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 63 significant links between journals and countries, of journal–country links with at least 15 preprints (**Figure 5a**). 11 countries had links to journals that published a disproportionate number of their preprints, but the United States had far more than any other country. 33 of the 63 significant links were between a journal and the United States: The U.S. is listed as the senior author on 41.7% of published preprints, but accounts for 74.5% of all bioRxiv preprints published in *Cell*, 72.7% of preprints published in *Science*, and 61.0% of those published in *Proceedings of the National Academy of Sciences* (*PNAS*) (**Figure 5b**).

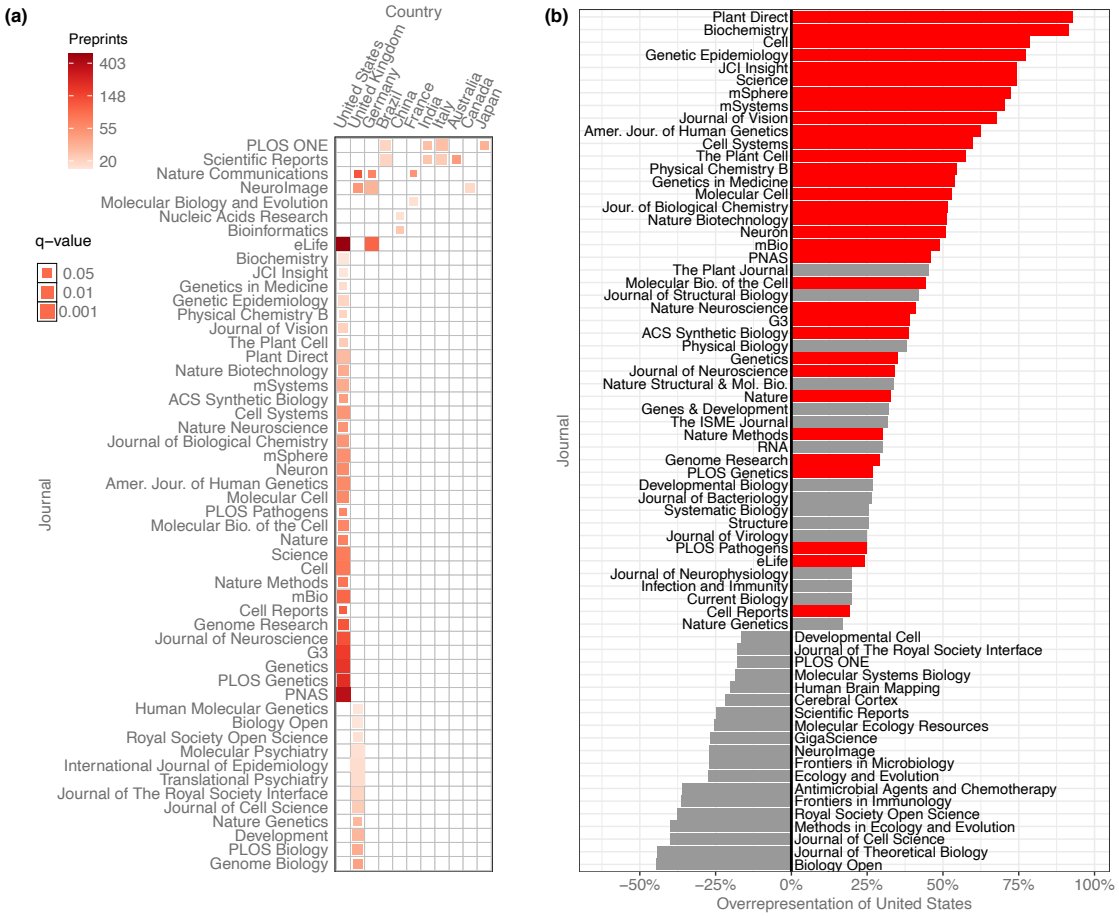

Figure 5. Overrepresentation of U.S. preprints. (a) A heat map indicating all disproportionately strong ($q < 0.05$) links between countries and journals, for journals that have published at least 15 preprints from that country. Columns each represent a single country, and rows each represent a single 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. See **Figure 5–source data 1** for the results of each statistical test. **(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 published at least 15 preprints with a U.S. senior author. 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. Journals for which the difference in representation was less than 15% in either direction are not displayed. The red bars indicate which of these relationships were significant using the Benjamini–Hochberg-adjusted results from χ^2 tests shown in panel A.

Figure 5–source data 1. Journal–country links. *supp_table06.csv*.

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 (**Figure 2**), which showed the U.S. and the U.K. are overrepresented while countries such as Turkey, Iran and Malaysia are underrepresented even after accounting for their comparatively low scientific output. Open science advocates have argued that there should not be a "one size fits all" approach to preprints and open access (Debat and Babini 2020; ALLEA 2018; Becerril-García 2019; Mukunth 2019), and further research is required to determine what drives certain countries to preprint servers—what incentives are present for biologists in Finland but not Greece, for example. Our 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 (**Figure 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 (**Figure 3**). When looking at international collaboration using pairwise combinations of these three measurements, countries fall along tidy gradients (**Figure 3–figure supplement 2**)—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. 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), what are the implications of the average Brazilian preprint receiving 37 percent fewer downloads than the average Dutch preprint? 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 Argentina and Taiwan? What role does language play in outcomes, and why do countries that get more downloads also tend to have higher publication rates? 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 the United States and United Kingdom. While it's possible this finding is coincidental, it 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; "New from eLife: Invitation to submit to Preprint Review" 2020). 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.

There were also 4,985 preprints (7.3%) for which we were not able to confidently assign a country of origin. An evaluation of these (see **Methods**) showed that the most prolific countries were also underrepresented in the "unknown" category, compared to the 148 other countries with at least one author. While it is impractical to draw country-specific conclusions from this, it suggests that the preprint counts for countries with comparatively low participation may be slightly higher than reported, an issue that may be exacerbated in more granular analyses, such as at the institutional level. Country-level differences in metrics such as downloads and publication rate may also be confounded with field-level differences: On average, genomics preprints are downloaded twice as many times as microbiology preprints (Abdill & Blekhman 2019b), for example, so countries with a disproportionate number of preprints in a particular field could receive more downloads due to choice of topic, rather than country of origin. Further study is required to determine whether these two factors are related and in which direction.

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.

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.

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. There are 45 countries that overlap between our analysis and that of Boudry and Chartron (2017). Of these, we found a modest correlation (Spearman's rho=0.295, p=0.0489) between a country's preprint publication rate and the rate at which publishers in that country assigned DOIs (**Figure 4–source data 2**). This indicates that countries with higher rates of DOI issuance (for publications dating back to 1955) also tend to have higher observed rates of preprint publication.

Attribution of preprints. Throughout the analysis, we define the "senior author" for each preprint as the author appearing last in the author list, a longstanding practice in biomedical literature (Riesenberg and Lundberg 1990; Buehring, Buehring, and Gerard 2007) corroborated by a 2003 study, which 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 59,562 preprints for which the country was known for the first and last author, 7,965 (13.4%) 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 position—is considered *straight counting* (Gauffriau et al. 2008). We repeated the process using *complete-normalized counting* (**Figure 1–source data 2**), 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=0.9971, $p=4.48\times10^{-197}$). 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 50 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. We determined significance by adjusting p-values using the Benjamini–Hochberg procedure.

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.), which counts articles, conference papers, and reviews in Scopus-indexed journals ("SJR – Help" n.d., "Scimago Journal & Country Rank" n.d.) We added the totals of these "citable documents" from 2014 through 2019 for each countries with at least 3,000 citable documents and 50 preprints. We used these totals to generate 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 2014–2019. While SCImago is not specific to life sciences research, it was chosen over alternatives because it had consistent data for all countries in our dataset. A 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 108,197 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). World maps were generated using the Equal Earth projection (Šavrič et al. 2019) and the rnaturalearth R package (South 2017), following the procedure described in Le et al. (2020). 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, and there are hundreds of preprints for which authors do not specify any affiliation information at all, including in the PDF

manuscript itself. 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 indexed 488,660 total author entries, one for each author of each preprint. These entries each included one of 136,456 distinct affiliation strings, which we processed using the ROR API before making manual corrections.

We first focused on assigning countries to preprints that were in the "unknown" category. We started by manually adding institutional assignments to "unknown" affiliation strings that were associated with 10 or more authors. We then used sub-strings within affiliation strings to find matches to existing institutions, and finally generated a list of individual words that appeared most frequently in "unknown" affiliation strings. We searched this list for words indicating an affiliation that was at least as specific as a country (e.g. "Italian," "Boston," "Guangdong") and associated any affiliation strings that included that word with an institution in the corresponding country. Finally, we evaluated any authors still in the "unknown" category by searching for the presence of a country-specific top-level domain in their email addresses—for example, uncategorized authors with an email address ending in ".nl" were assigned to the Netherlands. Generic domains such as ".com" were not categorized, with the exception of ".edu," which was assigned to the United States. While these corrections would have negatively impacted the institution-level accuracy, it was a more practical approach to generate country-level observations.

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. After manual corrections, there were 20,099 unknown affiliation strings associated with 51,855 authors.

There were also corrections made to existing institutional assignments, which are used to make the country-level inferences about author location. It appears the ROR 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" file.

In total, 12,487 institutional assignments were corrected, affecting 52,037 author entries (10.6%).

Prior to the corrections, an evaluation of the ROR assignments in a random sample (n=488) found the country-level accuracy was 92.2% ± 2.4%, at a 95% confidence interval. After an initial round of corrections, the country-level accuracy improved to 96.5% ± 1.6%. (These samples were sized to evaluate errors in the assignment of institutions rather than countries, which, once institution-level analysis was removed from the study, became irrelevant.) After another round of corrections that assigned countries to 14,690 authors in the "unknown" category, we pulled another random sample of corrected affiliations—using a 95% confidence interval, the sample size required to detect 96.5% assignment accuracy with a 2% margin of error was calculated to be 325 (Naing, Winn & Rusli 2006). Manually evaluating the country assignments of this sample showed the country-level accuracy of the corrected affiliations was 95.7% ± 2.2%.

Though preprints that were assigned a country could be categorized with high accuracy, we also sought to characterize the preprints that remained in the "unknown" category after corrections, to evaluate whether there was a bias in which preprints were categorized at all. Among the successfully classified preprints, the distribution across countries is heavily skewed—the 27 most prolific countries (15%) account for 95.3% of categorized preprints. Accordingly, characterizing the prevalence of individual countries would require an impractically large sample made up of a large portion of all uncategorized preprints. Instead, we split the countries into two groups: the first contained the 27 most prolific countries. The second group contained the remaining 148 countries, which account for the remaining 2,960 preprints (4.7%). We used this as the prevalence in our sample size calculation. Using a 95% confidence interval and a precision of 0.00235 (half the prevalence), the sample size (with correction for a finite population of 4,985) was calculated to be 307 (Naing, Winn & Rusli 2006). Within this sample, we found that preprints with a senior author in the bottom 148 countries were present at a prevalence of 12.6% ± 3.9%.

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 was formerly 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.3762814
- Supplementary tables are available in CSV format in the same repository. Legends for the supplementary tables are below.

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.

ALLLEA. "Systemic Reforms and Further Consultation Needed to Make Plan S a Success." 2018. European Federation of Academies of Sciences and Humanities. December 12, 2018. https://allea.org/systemic-reforms-and-further-consultation-needed-to-make-plan-s-a-success/.

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.

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.

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

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.

Le, Trang T., Daniel S. Himmelstein, Ariel A. Hippen Anderson, Matthew R. Gazzara, and Casey S. Greene. 2020. "Analysis of ISCB honorees and keynotes reveals disparities." *bioRxiv*. https://doi.org/10.1101/2020.04.14.927251.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. "Bias in Peer Review." *Journal of the American Society for Information Science and Technology* 64 (1): 2–17.

Mammides, Christos, Uromi M. Goodale, Richard T. Corlett, Jin Chen, Kamaljit S. Bawa, Hetal Hariya, Frith Jarrad, et al. 2016. "Increasing Geographic Diversity in the International Conservation Literature: A Stalled Process?" *Biological Conservation* 198 (June): 78–83.

Mattsson, Pauline, Carl Johan Sundberg, and Patrice Laget. 2011. "Is Correspondence Reflected in the Author Position? A Bibliometric Study of the Relation between Corresponding Author and Byline Position." *Scientometrics* 87 (1): 99–105.

Mongeon, Philippe, and Adèle Paul-Hus. 2016. "The Journal Coverage of Web of Science and Scopus: A Comparative Analysis." *Scientometrics* 106 (1): 213–28.

Moya-Anegón, Félix de, Zaida Chinchilla-Rodríguez, Benjamín Vargas-Quesada, Elena Corera-Álvarez, Francisco José Muñoz-Fernández, Antonio González-Molina, and Victor Herrero-Solana. 2007. "Coverage Analysis of Scopus: A Journal Metric Approach." *Scientometrics* 73 (1): 53–78.

Mukunth, Vasudevan. 2019. "India Will Skip Plan S, Focus on National Efforts in Science Publishing." The Wire: Science. October 26, 2019. https://science.thewire.in/the-sciences/plan-s-open-access-scientific-publishing-article-processing-charge-insa-k-vijayraghavan/.

Naing, L., T. Winn, and B.N. Rusli. 2006. "Practical Issues in Calculating the Sample Size for Prevalence Studies." *Archives of Orofacial Sciences* 1: 9–14.

Narock, Tom, and Evan B. Goldstein. 2019. "Quantifying the Growth of Preprint Services Hosted by the Center for Open Science." *Publications* 7 (2): 44.

Neuwirth, Erich. 2014. "RcolorBrewer: ColorBrewer Palettes. R Package Version 1.1-2." *The R Foundation*. https://CRAN.R-project.org/package=RcolorBrewer.

"New from eLife: Invitation to submit to Preprint Review." *eLife*, 13 May 2020 (accessed 23 June 2020). https://elifesciences.org/inside-elife/d0c5d114/new-from-elife-invitation-to-submit-to-preprint-review.

Nuñez, Martin A., Jos Barlow, Marc Cadotte, Kirsty Lucas, Erika Newton, Nathalie Pettorelli, and Philip A. Stephens. 2019. "Assessing the uneven global distribution of readership, submissions and publications in applied ecology: Obvious problems without obvious solutions." *Journal of Applied Ecology* 56: 4–9.

Okike, Kanu, Mininder S. Kocher, Charles T. Mehlman, James D. Heckman, and Mohit Bhandari. 2008. "Nonscientific Factors Associated with Acceptance for Publication in The Journal of Bone and Joint Surgery (American Volume)." *The Journal of Bone and Joint Surgery. American Volume* 90 (11): 2432–37.

Penfold, Naomi C., and Jessica K. Polka. 2020. "Technical and Social Issues Influencing the Adoption of Preprints in the Life Sciences." *PloS Genetics* 16 (4): e1008565.

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

R Core Team. 2019. *R: A Language and Environment for Statistical Computing* (version 3.6.2). http://r-project.org.

Research Organization Registry. 2019. *ROR API* (version a3b153c). Github. https://github.com/ror-community/ror-api.

Riesenberg, D., and G. D. Lundberg. 1990. "The Order of Authorship: Who's on First?" *JAMA: The Journal of the American Medical Association* 264 (14): 1857.

Ross, Joseph S., Cary P. Gross, Mayur M. Desai, Yuling Hong, Augustus O. Grant, Stephen R. Daniels, Vladimir C. Hachinski, Raymond J. Gibbons, Timothy J. Gardner, and Harlan M. Krumholz. 2006. "Effect of Blinded Peer Review on Abstract Acceptance." *Journal of the American Medical Association* 295 (14): 1675–80.

Saposnik, Gustavo, Bruce Ovbiagele, Stavroula Raptis, Marc Fisher, and S. Claiborne Johnston. 2014. "Effect of English Proficiency and Research Funding on Acceptance of Submitted Articles to Stroke Journal." *Stroke* 45 (6): 1862–68.

Sarabipour, Sarvenaz, Humberto J. Debat, Edward Emmott, Steven J. Burgess, Benjamin Schwessinger, and Zach Hensel. 2019. "On the Value of Preprints: An Early Career Researcher Perspective." *PLOS Biology* 17 (2): e3000151.

Šavrič, Bojan, Tom Patterson, and Bernhard Jenny. 2019. "The Equal Earth map projection." *International Journal of Geographical Information Science* 33 (3): 454–465. doi: 10.1080/13658816.2018.1504949.

Schwarz, Greg J., and Robert C. Kennicutt Jr. 2004. "Demographic and Citation Trends in Astrophysical Journal Papers and Preprints." *arXiv [astro-Ph]*. arXiv. http://arxiv.org/abs/astro-ph/0411275.

"Scimago Journal & Country Rank." N.d. Accessed February 15, 2020. https://www.scimagojr.com.

Sever, Richard, Ted Roeder, Samantha Hindle, Linda Sussman, Kevin-John Black, Janet Argentine, Wayne Manos, and John R. Inglis. 2019. "bioRxiv: The Preprint Server for Biology." *bioRxiv*. https://doi.org/10.1101/833400.

"SJR - Help." N.d. Accessed April 7, 2020. https://www.scimagojr.com/help.php.

South, Andy. 2017. "rnaturalearth: World Map Data from Natural Earth. R Package Version 0.1.0." *The R Foundation*. https://CRAN.R-project.org/package=rnaturalearth.

"Statistics." N.d. Global Research Identifier Database. Accessed February 2, 2020. https://www.grid.ac/stats.

"Trends in Preprints." 2019. PLOS. October 8, 2019. https://plos.org/blog/announcement/trends-in-preprints/.

Vence, Tracy. 2017. "Journals Seek out Preprints." *The Scientist*, January 18, 2017. https://www.the-scientist.com/news-opinion/journals-seek-out-preprints-32183.

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.

Supplementary figures

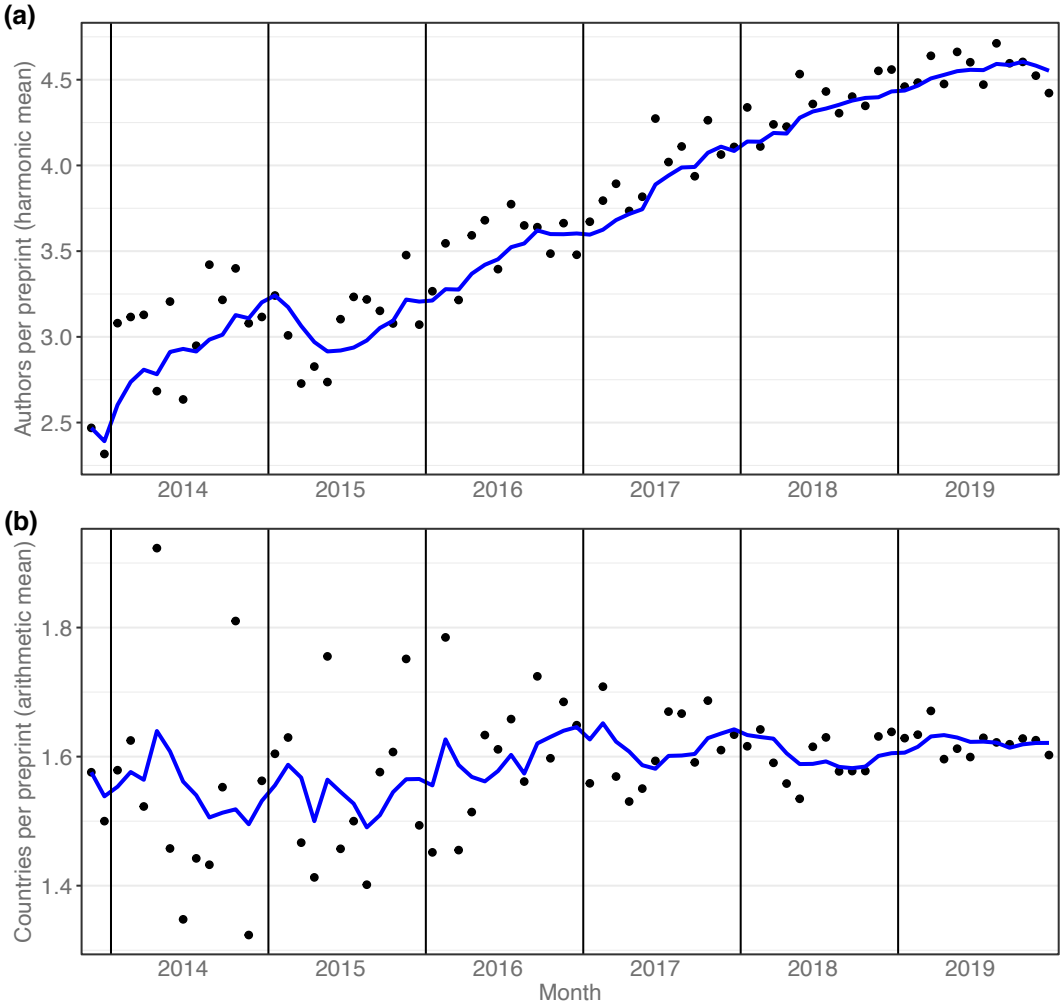

Figure 1–figure supplement 1. Preprint-level collaboration. (a) shows the average number of authors per paper over time. The x-axis indicates the year; the y-axis indicates the harmonic mean authors per preprint. Each point indicates the average of papers posted in a single month; the blue line indicates the six-month moving average. (b) illustrates the number of countries per preprint, over time. The x-axis indicates time; the y-axis indicates the arithmetic mean countries per preprint. Each point indicates the average unique countries found in all preprints posted in a single month. The blue line indicates the six-month moving average.

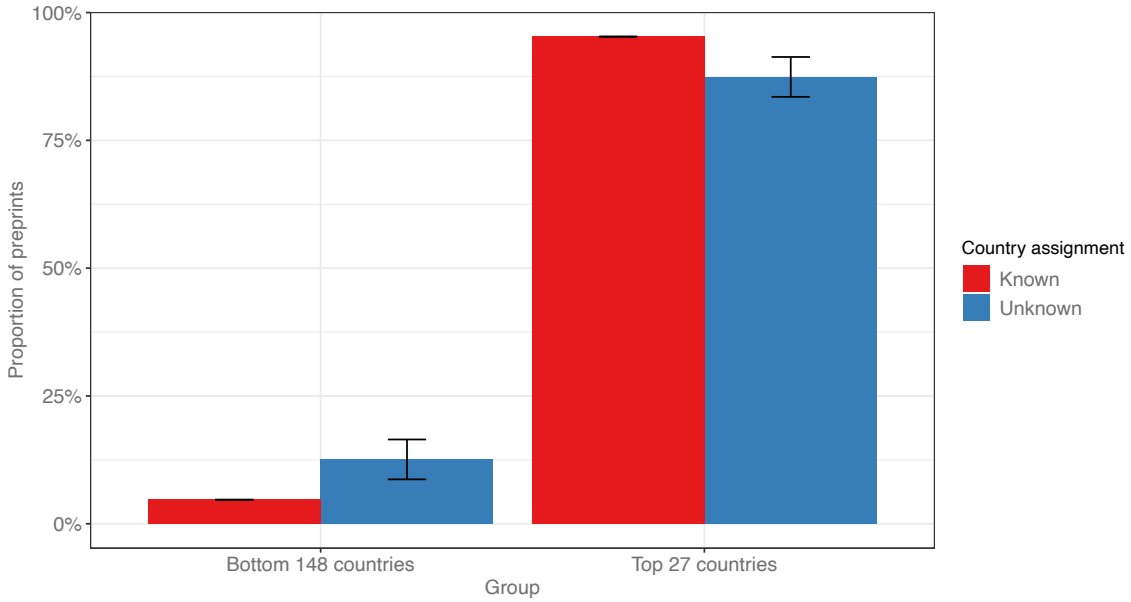

Figure 1–figure supplement 2. Preprints with no country assignment. This bar plot compares the observed prevalence of preprints from countries split in two groups: the 27 most prolific countries, and the remaining 148 countries for which at least one bioRxiv author was observed. The red bars indicate the proportion of preprints from countries in each group, out of all preprints with a country assignment. The blue bars indicate the proportion of preprints from countries in each group, out of a random sample of 325 preprints with no country assignment. The error bars indicate the margin of error at a 95% confidence interval.

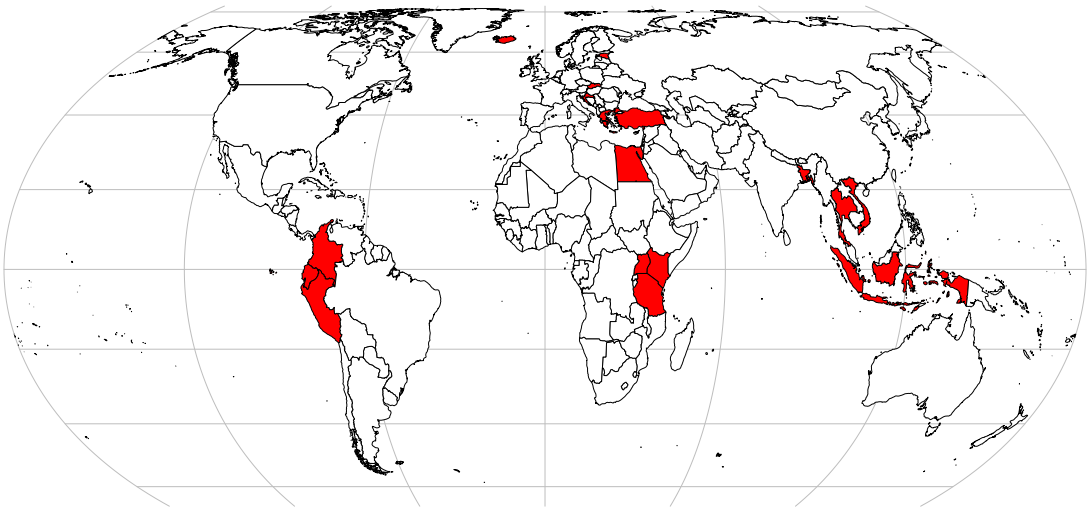

Figure 3–figure supplement 1. Map of contributor countries. World map indicating (in red) the location of contributor countries, defined as all countries listed on at least 50 international preprints, but as senior author on less than 20% of them.

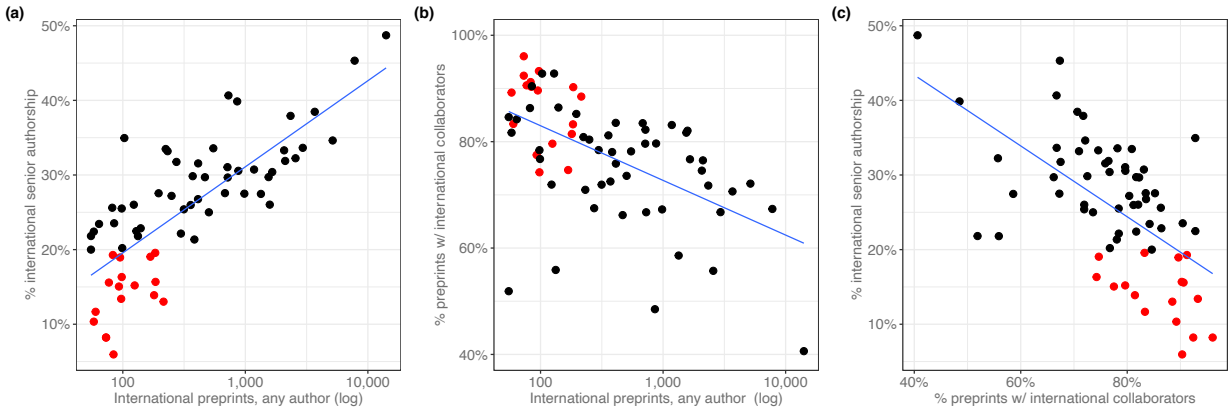

Figure 3–figure supplement 2. International collaboration correlations. Each point represents a country; the red points indicate those in the "contributor country" category. Blue lines indicate lines of best fit for each plot, though they are unrelated to the Spearman correlations reported for these relationships. **(a)** A scatter plot showing the relationship (Spearman's ρ=0.781, p=1.09×10^{-14}) between a country's total international preprints (x-axis; log scale) and the proportion of those preprints for which they are the senior author (y-axis). **(b)** A scatter plot showing the relationship (Spearman's ρ=-0.578, p=3.68×10^{-7}) between a country's total international preprints (x-axis; log scale) and the proportion of preprints with a contributor from that country that also include at least one contributor from another country (y-axis). **(c)** A scatter plot showing the relationship (Spearman's ρ=-0.572, p=5.32×10^{-7}) between the proportion of preprints with a contributor from that country that also include at least one contributor from another country (x-axis) and the proportion of those preprints for which that country appears in the senior author position (y-axis).

Supplementary tables

All supplementary tables are available as CSV files to facilitate machine readability. Legends are below, with a description of each table column:

Figure 1–source data 1. Preprints per country. *supp_table01.csv.* Each row represents a single country, sorted in descending order by the "senior_author" and "any_author" columns. The "alpha2" column indicates the two-letter country code defined in ISO 3166-1. "country" indicates the country name as recorded in the ROR dataset. "senior_author" lists the number of bioRxiv preprints for which the final author in the author list specified an affiliation to an institution in that country. "any_author" lists the number of bioRxiv preprints for which at least one author (in any position) specified an affiliation to an institution in that country.

Figure 1–source data 2. Preprint counting methods at the country level. *supp_table07.csv.* Each row represents a country, sorted in descending order using the "cn_total" and "straight_count" columns. The "country" column is the country name as recorded in the ROR

dataset. "cn_total" lists the number of preprints attributed to that country using the complete-normalizing counting technique. "straight_count" lists the number of preprints attributed to that country using the straight-counting technique.

Figure 2–source data 1. Country productivity and bioRxiv adoption. *supp_table02.csv*. Each row represents a single country, sorted in descending order by the "citable_total" and "senior_author_preprints" columns. The "alpha2" column indicates the two-letter country code defined in ISO 3166-1. "country" indicates the country name as recorded in the SCImago dataset. The "y2014" through "y2018" columns list the total number of citable documents attributed to that country in the SCImago dataset for the year specified. "citable_total" indicates the sum of all citable documents from that country from 2014 through 2018. "senior_author_preprints" lists the number of senior-author preprints attributed to that country from 2013 through 2019.

Figure 3–source data 1. Combinations of senior authors with collaborator countries. *supp_table03.csv*. Each row represents a combination of two countries, sorted alphabetically by the "contributor" and "senior" columns. The "contributor" column indicates the name of the contributor country. "senior" indicates the name of the country that appears as a senior author. "count" lists the number of preprints that include at least one author listing an affiliation from the country in the "contributor" column *and* a senior author listing an affiliation from the country in the "senior" column.

Figure 3–source data 2. Links between contributor countries and the senior-author countries they write with. *supp_table04.csv.* Each row represents a combination of two countries. The "contributor" column indicates the name of the contributor country. "senior" indicates the name of the country that appears as a senior author. "p" lists the p-value of a Fisher's exact test, as described in the "Methods" section. "with" lists the number of preprints that include an author from the "contributor" country *and* a senior author from the "senior" country. "without" lists the number of preprints that include an author from the "contributor" country but do *not* list a senior author from the "senior" country. "seniortotal" lists the total number of senior-author preprints attributed to the country in the "senior" column. "padj" lists the p-value from the "p" column, adjusted to control the false-discovery rate using the Benjamini–Hochberg procedure.

Figure 3–source data 3. International collaboration. *supp_table08.csv.* Each row represents a country, in alphabetical order by the "country" column. The "country" column indicates the country name as recorded in the ROR dataset. "alpha2" indicates the two-letter country code defined in ISO 3166-1. "intl_senior_author" lists, of bioRxiv preprints that include authors from at least two countries, the number for which the final author in the author list specified an affiliation to an institution in the specified country. "intl_any_author" lists, of bioRxiv preprints that include authors from at least two countries, the number for which at least one author (in any position) specified an affiliation to an institution in the specified country. "all_any_author" lists the number

of bioRxiv preprints—international or not—for which at least one author (in any position) specified an affiliation to an institution in that country. "intl_senior_rate" is the "intl_senior_author" column divided by the "intl_any_author" column. "intl_collab_rate" is the "intl_any_author" column divided by "all_any_author" column. The "contributor" column lists a boolean value indicating whether the specified country meets the criteria of being a "contributor country": a value of at least 50 in the "intl_any_author" column and a value of less than 0.2 in the "intl_senior_rate" column.

Figure 4–source data 1. Published pre-2019 preprints by country. *supp_table05.csv*. Each row represents a country, sorted in descending order by the "published" and "total" columns. The "country" column indicates the country name as recorded in the ROR dataset. "total" lists the number of preprints last updated prior to 2019 that list a senior author who declared an affiliation in the specified country. "published" lists, of the preprints counted in the "total" column, the number that are listed as published on the bioRxiv website.

Figure 4–source data 2. Publication rates and DOI usage. *supp_table09.csv*. Each row represents a country, sorted alphabetically. The "doi_rate" field lists the percentage of published papers from that country issued a Digital Object Identifier (DOI), according to Boudry and Chartron (2017). The "pub_rate" field lists the proportion of preprints from that country posted before 2019 that have been published.

Figure 5–source data 1. Journal–country links. *supp_table06.csv*. Each row represents a combination of country and journal, sorted in ascending order using the "padj" column, then descending order using the "preprints" column. "country" indicates the name of a country as recorded in the ROR dataset. "journal" indicates the name of a journal that has published preprints from the specified country. "preprints" indicates the number of preprints last updated prior to 2019 that were published by the specified journal that list a senior author affiliated with the specified country. "expected" indicates the number of preprints we would expect the specified journal to have published from the specified country, if the country *and* journal both published the same number of papers, but the journal's publications mirrored the country-level proportions observed in published bioRxiv preprints overall. "p" indicates the p-value of a chi-squared test as described in the "Methods" section. "padj" lists the p-value from the "p" column, adjusted to control the false-discovery rate using the Benjamini–Hochberg procedure. "journaltotal" lists the total preprints published by the specified journal that were last updated on bioRxiv prior to 2019. "countrytotal" lists the total preprints posted to bioRxiv prior to 2019 that list a senior author affiliated with the specified country.

Table 2–source data 1. Country assignment accuracy measurements. *supp_table08.csv.* Each row represents a country, sorted in alphabetical order using the "country" column. The "country" column lists the two-letter country code defined in ISO 3166-1. "uncorrected_total" indicates the

number of institutional affiliations attributed to the specified country before manual correction. "unchanged" indicates the number of institutional affiliations for which the country attribution did not change after correction. "removed" indicates the number of institutional affiliations that were attributed to a different country after correction. "added" indicates the number of institutional affiliations that were attributed to a different country *before* correction but then attributed to the specified country. "precision" lists the proportion of the specified country's institutional affiliations that were detected before correction. ("unchanged" column divided by "uncorrected_total" column.) "recall" lists the proportion of the institutional affiliations attributed to the specified country before correction that remained attributed to that country after correction. ("unchanged" column is the numerator; denominator is "unchanged" column plus "added" column.)