Supporting Information

Vale 10.1073/pnas.1511912112

SI Methods

Scoring of Panels and Data. Panels were scored by counting the lettering (*A, B*, etc.) designations in figures. Data-containing tables and figure schematics were counted as panels. Videos in the supplemental material were not counted. Panels are an imprecise proxy for the experimental data contained within a paper, and therefore we estimated the number of distinct pieces of data in Fig. S2. For example, a single experiment may be displayed in multiple panels with separate letters, such as different views of a fluorescence micrograph. Conversely, a single labeled panel may contact multiple different types of experiments. Therefore, panels were scored as to whether they contained distinct pieces of data. To provide examples, if a representative image was in one panel and quantification of the same experiment was provided in another panel, then both panels would be counted as a single piece of data. Also, if the same experiment was quantified in multiple ways (e.g., analysis of different organelle sizes or multiple kinetic parameters from the same experiment) and presented in multiple panels, then it would still be counted as a single piece of data. Different views or slices of the same sample, views of the same crystal structure, and multiple probes (for DNA or protein) used for the same sample also were considered as a single piece of information. Identical experiments applied to two different cell lines were also considered as one piece of data. Sequence alignments were counted as a one piece of data as were tables. Differentiation of separate pieces of data was evaluated only between panels in a single figure and not between figures. Schematics and model figures were also not counted as "data" in this analysis. Two graduate students independently quantified the data presented in January and February 1984 articles in *Cell* to determine whether these criteria led to consistent scoring. The average number of distinct pieces of data per article were scored as 7.33 and 7.16, indicating good overall agreement between two independent scorers. The other months of January–June from 1984 and 2014 for *Cell* and *Nature* were scored by a single person.

Analysis of UCSF Graduate Student Publications. Several basic science graduate programs in the 1980s have disappeared or merged with other programs, and new graduate programs have formed more recently. To make a fair comparison between the 1980s and current times, we analyzed graduate student data from four basic science PhD degree-granting programs that have spanned both time periods: Biochemistry and Molecular Biology, Biophysics, Genetics, and Neuroscience. Because this study was focused on experimental science, students conducting exclusively theory or modeling studies were not counted in this analysis (five students in 2012–2014 in this category). Information on the time of entering graduate school and the time at which the degree was granted was obtained from the UCSF student registrar's office. Publication references and dates for the students were obtained by searching PubMed. Reviews or methods papers that were largely more detailed descriptions of previously published methods were not counted. "Shared authorship" represents a difficult issue, because this designation did not exist in the 1980s. While acknowledging potential drawbacks of doing so, we scored only the order of authorship; thus, a shared first author in the second position was counted as a second authorship in our analysis. The reason for doing so is to allow a more direct comparison with data from the 1980s, when cofirst or cosecond authorship was not used as a credit-sharing strategy. However, an exception was made for students that published only one cofirst, second-position authorship paper in their graduate work; in this case, this paper was counted as a first-author paper (six students in this category). A second complication was scoring papers that were published a year or more after a degree was awarded. We directly emailed faculty or students from the 1980s to inquire whether such late publications were a product of their thesis work or primarily from a subsequent postdoctoral period (which were not scored). With only a couple of exceptions, these late publications were from thesis work; in many cases, difficulties in communication after leaving the laboratory between student and PI in the "pre-Internet" era were cited as reasons for the delay in publication. However, papers published ~2 y beyond their graduation date were not scored in our analysis, unless it was their sole paper (one student). For the recent UCSF graduate students, we contacted the PIs of students who graduated between June 2013 and December 2014 to inquire whether the student was working on additional first- or second-author publications and whether the paper was in preparation, submission, revision, or in press. We added all anticipated publications to the student's data profile (17 students), estimating an approximate, best circumstance time of publication based upon the status described by the PI (~9 mo for in preparation, 6 mo for submitted, and 3 mo for revision). It is possible that some of these anticipated papers may not be published or published with a longer time frame. If a student did not produce a first- or a first/second-author publication, then a "0" was entered for that category of publications. In the 1979–1989 group, there were eight students without a first-author publication and four students for whom we could not find a record of any publication in PubMed although supporting evidence on the internet confirmed that they graduated. In the 2013–2014 group, there were nine students without an anticipated first-author publication and four students without an anticipated first/second-author publication. Students who did not publish a first-author paper were not included in the analysis of time to first-author publication.

SI Q&A Regarding Preprints

The following questions or comments (paraphrased here in italics) were raised by others in response to the initial posting of this Perspective on bioRxiv. My responses are presented below each question.

Reproducibility and Quality. *We already have a looming problem of irreproducibility. Preprints will just encourage more irreproducible results to be spread throughout the community.*

This issue is indeed important, because preprints open up the possibility of wide-spread science communication before peer review. Preprints might allow work to be disseminated before mistakes are caught by peer review and thus could lead researchers down wrong tracks. On the other hand, many peer-reviewed articles have proven to be inaccurate, and there is little data on the success rate of peer review in filtering out irreproducible, inaccurate or fraudulent data. It might be better to have many people see the work right away, allowing the possibility of inadvertent mistakes to be caught and helping peer reviewers and the authors themselves to produce a more accurate final product. Furthermore, a high profile result will likely be replicated right away and thus validated before it is published in a high profile journal. A good commenting system on preprints might help this process.

The immediate exposure of preprints also will likely be a motivating factor for accuracy. Many researchers intentionally do not complete all of their experiments in their first journal submission because the journals emphasize "impact" in their first round of screening. Thus, mistakes in an initial journal submission and peer

review are "invisible" and have no or minimal negative consequences for the author if the paper is rejected. This scenario contrasts with a preprint submission, in which all of the data are immediately transparent to the science community. This transparency will cause good scientists to be very cautious about their submission to a preprint server because that work will be seen and judged by their peers immediately. Having the scientist decide when his/her work is ready for dissemination will be empowering and accompanied by a sense of responsibility.

The subject of reproducibility, however, is very complex and should be taken into careful consideration. I would recommend developing a future plan for collecting data on how preprints impact scientific reproducibility but see no reason why preprint use cannot be encouraged in the near-term.

Journal filters are good. I don't have time to sort through work in a massive preprint server. I also am more assured of quality if I read work in top journals.

Preprints would not replace the journals but rather exist alongside them. You might prefer reading journals to learn about a new field, where the speed of accessing new information might be less important. However, preprints would allow you to access information faster in your own field, which might help to advance your research program. Thus, preprints and journal articles together can serve different needs in the scientific community. As discussed in the Perspective, it is also possible to experiment with filters that will allow users to sort through the content of preprint servers for benchmarks of quality (e.g., specific scientists, the funding source that supported the work, recommendations from user groups, etc.).

There are already many low interest, low quality papers being published in journals. Won't a preprint server just accentuate this problem and further plague our scientific community?

Most scientists seek to establish a good reputation and thus will want to showcase high quality work to their colleagues, regardless of whether it is through a preprint or journal. Some scientific material that is currently hard to publish, such as confirming a finding or reporting a negative result, might be posted on preprint servers, thus adding more scientific material than is currently being published in journals. However, as discussed above, the best solution will be to create better mechanisms of searching for relevant information that appears both in preprints and in journals.

What about medical sciences? If a preprint on a medical procedure or a drug is posted but is wrong, then it might have disastrous consequences on patient care.

This concern is reasonable, especially given existing attention on irreproducible work in medically related areas being published in peer review journals. The medical sciences community will have to confront this issue themselves and decide on the best path. Biology is not a single monolithic enterprise but is composed of many different disciplines and communities. These different communities can decide when or whether preprints represent a good mechanism for communicating their results. Following the history of arXiv, different communities (e.g., different branches of physics, mathematics, and computational sciences) embraced preprints at different times.

Journals and Preprints. *With potential comments being posted on preprints, won't such comments endanger the subsequent journal-based peer review process?*

This possibility will have to be tested in practice. arXiv does not have a comment system, while bioRxiv does. One might argue that commenting could improve subsequent peer review if thoughtful people use the commenting system. For example, a particularly good comment on a preprint could help a journal referee in the review. Importantly, because the identity of the preprint commenter is known, the system will prevent competitors from making negative remarks behind a cloak of anonymity. Furthermore, through a preprint, authors can receive direct feedback on their work from the community. Such comments, some of which might not have arisen through journal peer review, can help authors to revise their work and publish a better paper in the end. Thus, preprints could facilitate direct feedback to authors and information for referees, both of which could lead to improved revisions of the work.

Someone posts a preprint with a quick and dirty experiment to make a claim. I worked much harder to establish proof with a more complete and convincing set of evidence. I am now forced to post my preprint a month later. Won't journals be reluctant to publish my paper because they will have seen the earlier posted work?

Quite the opposite may occur. Currently, journals want to publish stories first, but some of this drive may diminish if work routinely appears first as preprints. Journals then may be incentivized to look more toward quality than speed and seek to publish the definitive work that will stand the test of time and become the publication that is most cited. Also, the issue of speed versus quality of research already exists in the present journal system. For example, a researcher can quickly publish a study with minimal data in a lower journal; this publication can potentially color another journal's view of a more extensive manuscript being submitted later. Furthermore, if a scientist repeatedly has a pattern of reporting quick and dirty experiments to beat competitors rather than doing complete and thoughtful work, then this behavior will tarnish his/her reputation and will not be a path to long-term success. In addition, there is "version control" with preprints; if someone rushes out an incomplete paper and then subsequently wants to correct mistakes, they can upload a new version, but the original version remains on the site for all to see.

How are news and publicity handled if there is a preprint submission as well as a subsequent journal publication.

Historical examples from arXiv reveal various ways in which publicity has been handled. In some cases, a journal or press will "find" a preprint on arXiv and run a story on the work before journal publication. In some cases, the preprint will be posted on arXiv at the time of acceptance to a journal (but before publication), and the press will cite the arXiv preprint and name the journal in which it will be ultimately published. Even government agencies such as NSF have issued press releases on a preprint. In other cases, publicity arises only with the greater attention associated with the journal publication. A critical issue is that the authors need to follow the embargo policy of the journal to which they intend to submit, which usually prohibits the authors from speaking about their work directly with the press themselves before publication: See *Nature*'s guidelines on publicity at www.nature.com/authors/policies/confidentiality.html. In general, news and publicity have been managed successfully with arXiv and the journal system.

A bigger issue to me is open access.

Preprints are free for anyone in the world. Use of this system will therefore ensure that there is always a version of the manuscript that is freely available, regardless of the journal in which it is eventually published. However, for certain journals, the very final and accepted version of an article cannot be posted as preprint for up to 6 mo from the time of publication (e.g., see *Nature* guidelines cited above).

Having preprints listed on PubMed would be helpful as one-stop shopping to find science content.

Currently PubMed is only for peer-reviewed articles. To facilitate content discovery, one could imagine developing a new biologist-friendly search engine that will search for content on PubMed, bioRxiv, and arXiv. On the other hand, such functions could be integrated into PubMed. Both solutions are workable, and the community and NIH should decide on the best course of action.

Ethical and Practical Issues for Biology. *Experimental biology is moving so fast. I am worried that if I post on bioRxiv or arXiv then someone will scoop me by rushing a paper to a journal and perhaps be luckier in the publication process.*

The possibility that results/ideas might be "stolen" from a preprint, resulting in the loss of credit for a researcher, seems to be a prevalent concern in the biology community. This potential scenario is why some argue that preprints simply will not work in biology as they have in physics. Here is an excerpt from a reviewer's comment on this Perspective from PNAS:

"Should the author choose to continue to push the prepublication format, he might anticipate the following criticism of his logic. He poses that prepublication works for experimental physics so it can work for experimental biology. This analogy seems flawed. Physics today is like biology 40 years ago. The experimental systems needed to address a problem are unique: for example, a synchrotron to address a problem in subatomic physics (like a bicoid mutant that nobody but Ed Lewis had). Hence, a prepublication is safe. Nobody can quickly generate the data of the prepublication or has preliminary data similar to the prepublication. What makes current biology so exciting is the lightning fast connections that are made between very rapidly moving systems. These same connections generate problems for the prepublication concept. Here is the scenario that critics will bring forward. One has a very nice unpublished discovery and talks about it at a meeting or university. A member of the audience has some preliminary results in another system that, in the context of the talk, all of a sudden make sense. With much greater confidence, the member of the audience adds a few experiments and publishes these results and common conclusion in a prepublication. This minimal publication is much weaker than the lecture but nonetheless gets priority. This scenario can't or would rarely happen in physics but would be the fear of every biologist talking about unpublished results. Nobody would share unpublished results because of the speed at which unrefereed results could be published. In this light, the author might use the text to probe a little deeper why biology did not move to the prepublication format if, in fact, biology and physics are interchangeable. As it is, he will get much criticism for comparing apples to oranges."

These remarks are thoughtful and reflect many people's concerns. My response is that not all biology experiments are so lightning fast to repeat. Some are, but most papers are fairly complex and not trivial to repeat in a few weeks even by a well-established competing laboratory. However, talking about work in a lecture constitutes a problem for establishing priority, as the referee indicates. Physicists tend to acknowledge and credit information transmitted in public talks. Because not everyone can attend a given lecture, part of the motivation for establishing arXiv was to create a common access point where a discovery could be announced to the community. Because arXiv is so widely viewed by the community, it is very difficult for an individual to "steal and run" with an idea/experiment with the excuse that they never saw it on arXiv. If preprints are going to be successful, they must carry with them the gravitas of priority. Finally, I asked Paul Ginsparg, founder of arXiv, whether there were examples of transgressions where a result was posted on arXiv and then it was copied and published faster in a journal by someone else in an attempt to claim priority. He could not think of any such occurrence and also thought that the physics community would not tolerate such behavior. They also would not tolerate someone publishing a cheap paper on arXiv in response to hearing an outstanding work or idea in a public lecture. Furthermore, work appearing close in time as preprints (e.g., within a couple of months) will be compared based upon quality and acknowledged as codiscoveries if they deserve to be, just as is the case with journal publications. Perhaps physicists are not behind biologists (see the referee's comment) but rather are ahead of us in science communication and ethics.

Biologists develop specialized reagents and strains for their work; there is an obligation to release these reagents immediately to the community upon publication. Will this obligation apply to preprint postings?

Some investigators may be happy to release their reagents or share software at the preprint stage. Others may be reluctant to do so until after journal publication, especially given current concerns described above. However, the community may wish to develop a policy if preprints become more widely used. One could imagine a grass-roots agreement providing a grace period (for example, 1 y), by the end of which all reagents, strains, and all source data must be made publicly available after a preprint is posted. Any such recommendation policy could be reevaluated as the system matured.

The main problem in the life sciences is the lack of academic and industry jobs and excessive competition for those jobs. Getting my work out earlier with a preprint is not going to help me get a job, especially if everyone is posting preprints.

Agreed. Posting preprints will not help you get a job per se because that process is determined by competition with other applicants. However, a preprint might be useful in some circumstances because it will allow a potential employer to access your work if it has not yet been published (e.g., if held up in a prolonged review). Currently, a manuscript listed as "submitted" on a curriculum vitae counts for very little. In physics, recent preprints on arXiv play a crucial role in evaluating candidates for jobs.

The love of just a few elite journals is the biggest problem in life sciences these days. I don't see how the preprints are going to solve this issue.

Preprints will not solve this issue. However, they might represent the start of a longer term change in how scientific work is evaluated. Because of the necessity to stay informed, scientists will read preprints in their own field and make judgments of the quality before it has a journal name attached to it. Grant reviewers might also start to comment on the quality of work posted as a preprint, if it is presented as key evidence for a new research program. For such a vision to succeed, the best work in biology needs to be posted on a preprint server and not solely routed to the elite journals. Leaders in the biomedical community will have to post their best work as preprints to set an example.

Feasibility of Preprints and the Potential of Other Mechanisms. *Scientists are set in their ways. No one is going to use preprints.*

Scientists are indeed conservative with regard to their habits. They are also unlikely to change their habits simply based on altruism. However, they will use preprints if they provide practical benefits for their research and careers. Preprints could benefit scientists if they (*i*) allow them to establish priority for a discovery in a more predictable way than navigating an unpredictable journal review process, (*ii*) allow them to use preprints as evidence of productivity in grant applications, particularly in cases where a new research direction is being pursued, (*iii*) enable graduate students to provide evidence of scholarly work for graduation or postdoc applications, thereby potentially decreasing their training time by many months, and (*iv*) allow them to obtain feedback on their work earlier than is currently possible through the journal system. Although it is unlikely that everyone will switch to preprints, its use might increase significantly if people try it and have good experiences (as has occurred with arXiv). Also, younger scientists may be "less set in their ways" than more senior scientists. They have grown up socializing in an Internet world so the notion of sharing information through a preprint server will not seem so foreign to them.

I like the idea of preprints, but I hesitate to advise junior faculty in my department to submit preprints as it might not be good for their career.

Here is a recent experience from James Fraser, a junior faculty member at UCSF:

"We submitted the paper (www.ncbi.nlm.nih.gov/pubmed/26280328) and the preprint (biorxiv.org/content/early/2015/02/03/014738) in February. In the intervening months before the paper was published online in August (publication went smoothly, with a supportive editor and constructive reviews), the following events happened based upon the information made available through the preprint:

i) Our software was downloaded by multiple groups around the world and used locally at UCSF to improve other EM structures.
ii) I was invited to an EM validation meeting to discuss the work (even though we hadn't published in that area before).
iii) My student was invited to speak at the local bay area EM meeting.
iv) My student got a fellowship (ARCS). He probably would have gotten it anyway—but having a bioRxiv doi to point to for his "in review" paper may have helped.
v) I talked to people about the method at multiple meetings and was able to point them to the preprint to judge for themselves."

bioRxiv has been around since 2013 and it has a small following. Hasn't the experiment been done already and the answer is in hand?

I would argue that the experiment has not been done properly. Currently there are several major disincentives for preprints, which include (*i*) an inability to cite a bioRxiv preprint on NIH grants, (*ii*) possible restrictions in subsequently publishing the work in certain journals, and (*iii*) the potential of being scooped because it is unclear as to whether a preprint constitutes "priority" among scientific peers for a discovery. Note that priority among peers is a culture issue of assigning credit within the profession and differs from the legal term of "disclosure" (e.g., for a patent), which encompasses any public presentation. Given these restrictions, it is difficult to strongly recommend preprints in their current state. These deterrents need to be removed to give preprints a fair chance.

There are better ways of transmitting scientific information than a preprint plus journal system. Why stall the inevitable by encouraging preprints? Shouldn't we build a completely new system that will replace both journals and preprints?

Scientific communication will likely evolve in new ways over the coming years and decades. The question is how to get from where we are now to where we want to be. Replacing the current journal system now with something new is likely to meet considerable resistance and thus fail. Preprints, on the other hand, represent a viable evolutionary intermediate. Preprints can coexist with the journal system and thus do not represent an either/or choice for scientists. Also, supporting preprints should not prevent other desirable changes in the science communication system that our community would like to establish later on (e.g., changes in pre- or postpublication review and evaluation). Indeed, a short-term success with preprints would convey a message to our community that we are not locked into the status quo and that other changes are possible over time.

F1000Research *has a complete publishing platform that communicates the preprint but also initiates transparent peer review and then indexes successfully peer reviewed papers on PubMed. What about such a publication system?*

F1000Research has an interesting and new publishing process. However, submitting a work to *F1000Research* precludes submitting the same work in another journal (unlike bioRxiv or arXiv). Individual scientists will have to decide on a publishing mechanism that makes sense for them—submission to *F1000Research*, or through a preprint server (bioRxiv/arXiv) plus a subsequent journal of their choice, or through other mechanisms.

Are you sure that preprints will work in biology?

No. I am sure only of death and taxes. However, we have to try experiments in scientific communication. Preprints seem to be a relatively easy one to try because the cost, infrastructure, and potential harm are minimal and the current barriers are not so difficult to overcome. If preprints are tried but do not succeed, then the answer will be clear and new ideas can be investigated.

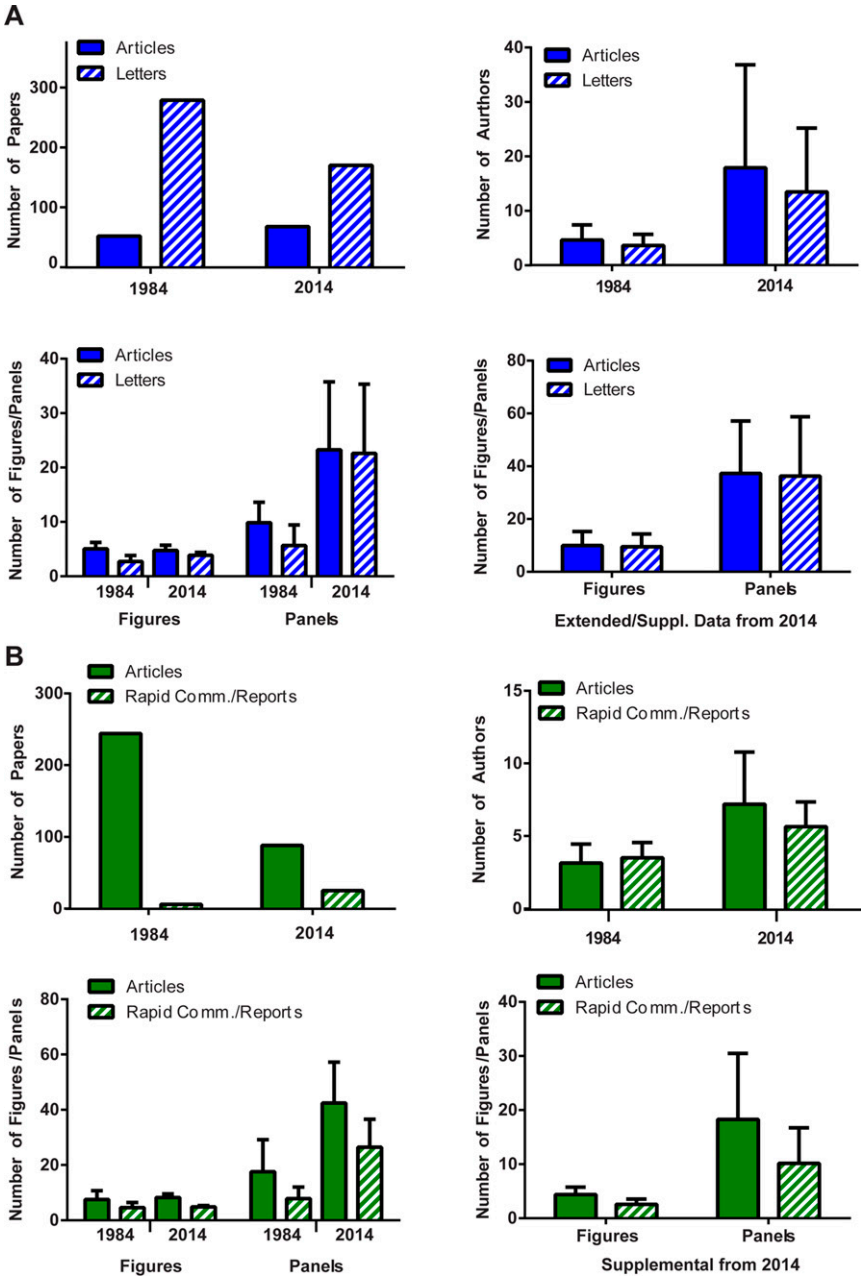

Fig. S1. Breakdown of information for long and short format papers. (*A*) Data for *Nature*: long format (Articles) and short format (Letters). (*B*) Data for *The Journal of Cell Biology* (JCB): long format (Articles) and short format [Rapid Communications (1984 name) or Reports (2014 name)]. The means and SDs are shown. These data from long and short format papers were combined together in the analysis in Fig. 1.

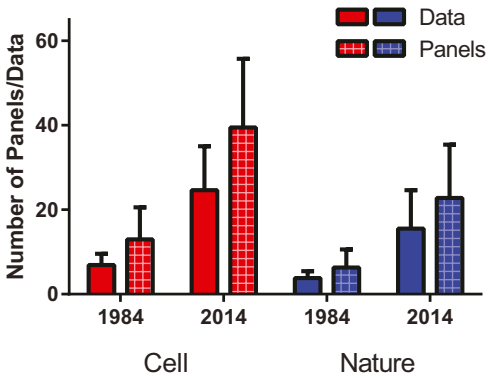

Fig. S2. Analysis of the number of panels (assigned as a letter in the figure) and distinct pieces of experimental data in the print versions of *Cell* and *Nature*. Each piece of "data" is defined as being derived from a distinct experiment or a significant type of new analysis (see *SI Methods*); as an example, two panels that show two views of a micrograph would be considered as a single datum in this analysis. The means and SDs are shown. Although the scoring of "distinct data" is admittedly subjective, the analysis shows an approximately similar ratio of data to panels in the two journals and between the two different time periods.

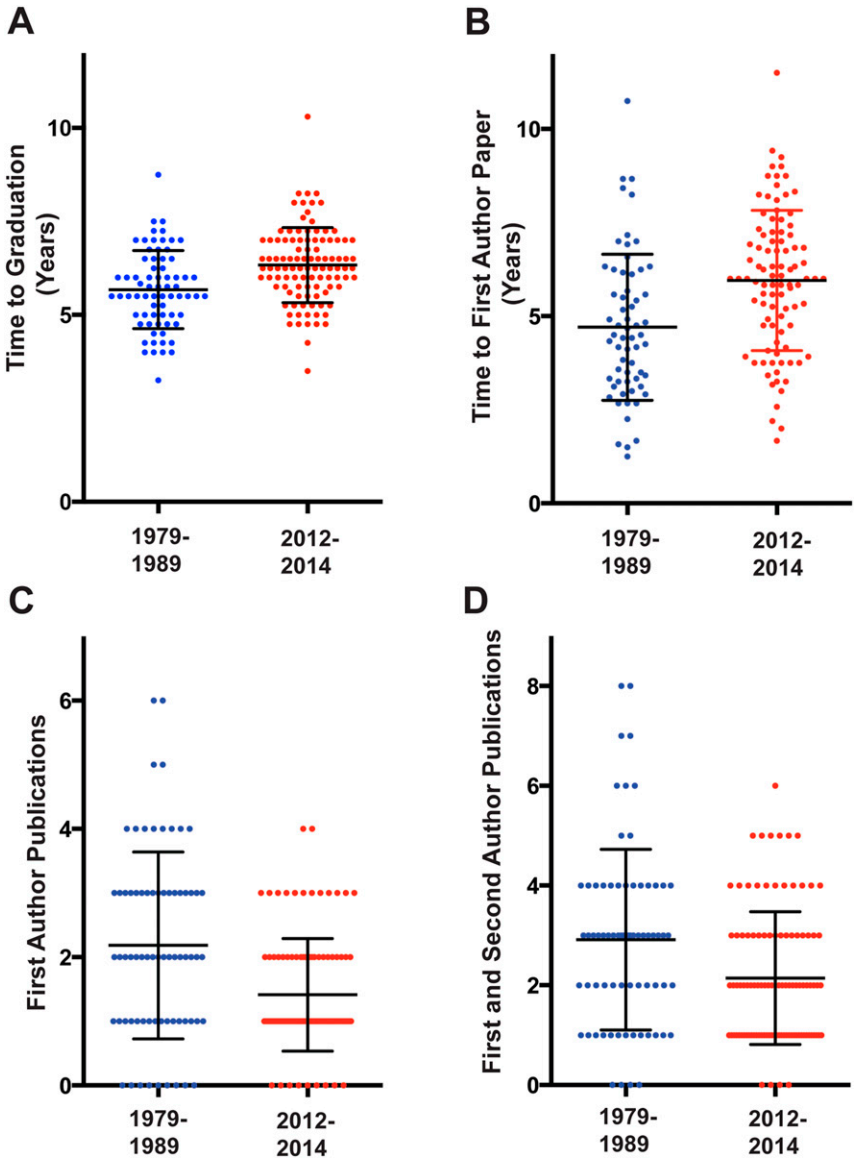

Fig. S3. Scatter plot of data on (*A*) time to graduation, (*B*) time to the first first-author publication from entering graduate school, (*C*) number of first-author publications, and (*D*) number of first- and second-author publications from UCSF graduate students corresponding to Table 1. The time periods of graduation are indicated on the *x* axis (n =71 for 1979–1989 graduates; n = 104 for 2012–2014 graduates). The middle black lines indicate the mean, and the error bars show SDs. Data for graduation and publication times were rounded to the nearest quarter of a year in this graph. The *P* value differences (Kolmogorov–Smirnov test) between the two time periods for time to graduation, time to the first first-author publication, number of first-author publications, and number of first- and second-author publications are 0.0007, 0.0002, 0.0009, and 0.0083.

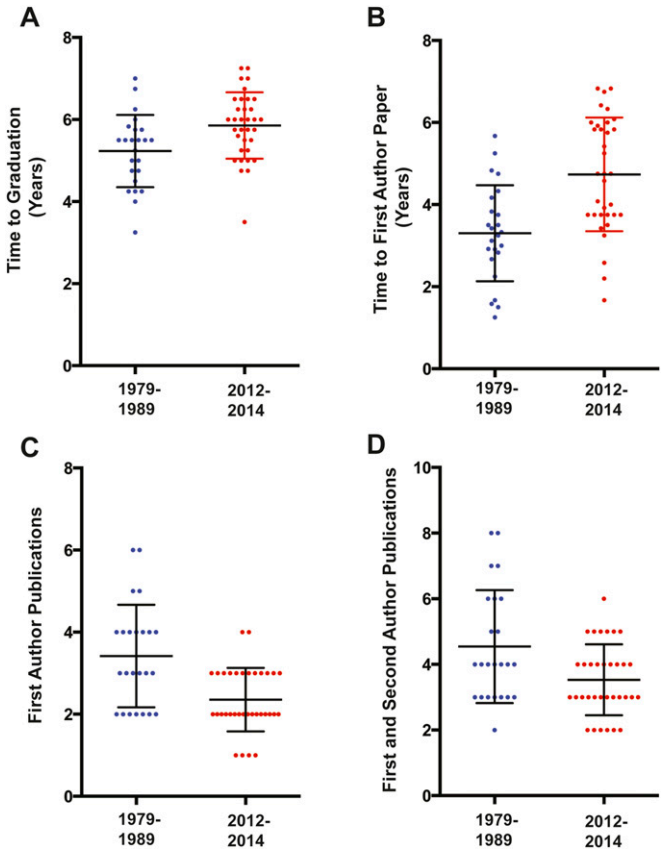

Fig. S4. Scatter plot of data of (*A*) time to graduation, (*B*) time to the first first-author publication from entering graduate school, (*C*) number of first-author publications, and (*D*) number of first- and second-author publications from the top one-third UCSF graduate student group with the best publication record corresponding to Table 1. The time periods of graduation are indicated on the *x* axis (n =24 for 1979–1989 graduates; n = 34 for 2012–2014 graduates). The middle black lines indicate the mean, and the error bars show SDs. Data for graduation and publication times were rounded to the nearest quarter of a year in this graph. The *P* value differences (Kolmogorov–Smirnov test) between the two time periods for time to graduation, time to first first-author publication, number of first-author publications, and number of first- and second-author publications are 0.03, 0.002, 0.022, and 0.289.