nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-019-0585-7

Large-scale analysis of micro-level citation patterns reveals nuanced selection criteria

Julia Poncela-Casasnovas[1], Martin Gerlach[1], Nathan Aguirre[1] and Luís A. N. Amaral[1,2,3*]

The analysis of citations to scientific publications has become a tool that is used in the evaluation of a researcher's work; especially in the face of an ever-increasing production volume[1–6]. Despite the acknowledged shortcomings of citation analysis and the ongoing debate on the meaning of citations[7,8], citations are still primarily viewed as endorsements and as indicators of the influence of the cited reference, regardless of the context of the citation. However, only recently has attention[9,10] been given to the connection between contextual information and the success of citing and cited papers, primarily because of the lack of extensive databases that cover both types of metadata. Here we address this issue by studying the usage of citations throughout the full text of 156,558 articles published by the Public Library of Science (PLoS), and by tracing their bibliometric history from among 60 million records obtained from the Web of Science. We find universal patterns of variation in the usage of citations across paper sections[11]. Notably, we find differences in microlevel citation patterns that were dependent on the ultimate impact of the citing paper itself; publications from high-impact groups tend to cite younger references, as well as more very young and better-cited references. Our study provides a quantitative approach to addressing the long-standing issue that not all citations count the same.

The study of scientific enterprise has a long history, but has recently experienced a sharp increase in interest due to the availability of large digitized bibliometric databases. One of the current quantitative foci is on networks of citations[1]. Citations provide a simplified abstraction that allows for the systematic study of matters such as the organization of knowledge[4,12], the importance of mentoring and teams[13–16], or innovations[17] and their impact[18,19]. There is also a general understanding that citations are intended as recognition of peers' work and therefore may inform about the quality of the cited work[7,20–23]. Recent survey-based studies[8] have reported a correlation between perceived quality and number of citations, at least in the case of a researcher's own work. More broadly, a study of citations among US-produced films—which can be interpreted as the equivalent of scientific citations in the film industry—found that peer citations provided the most predictive proxy for the identification of culturally, historically or aesthetically significant films[24]. Despite the interest in citations, most researchers agree that citations are an imperfect measure[20,21,23], and are not free from biases[25], such as the rich-gets-richer mechanism[2], or confounding factors such as age dependencies[26,27], gender[16,28], field, institution, journal and author dependencies, and fads.

A different stream of research, based on text analysis of small-sized samples of papers and surveys of academics, has investigated the multiple ways in which citations are used by authors[29,30]. This research showed that homage, credit, methodology identification, providing background, and the correction or criticism of the work of others are common reasons for citing, as are non-scientific reasons, such as rhetorical construction. In addition to these individual reasons, the cited references also signal domain knowledge and membership of a specific scientific field[31].

As is the case for many human activities, researchers sometimes copy what others in their field do or cite, as a learning mechanism and as a way of building on previous knowledge. It has been shown that such copying gives rise to multiplicative growth[32], which in turn creates heavy-tail distributions, such as those found for the number of citations that papers accrue[33,34]. Interestingly, studies on the propagation and rates of citation errors[35–37] have shown that references are also literally copied from one paper to another. Thus, the references cited in a manuscript may be chosen by a combination of what the authors think is relevant, the consensus of what is important in the field at that time and literal copying of references from other papers.

If this is indeed how researchers choose references, then there should be micro-level citation patterns that can be empirically uncovered using data-driven approaches. The increasing availability of large, full-text datasets allows for the study of the use of citations in context[38] for a large number of papers. Therefore, we investigated whether there are attributes of the references themselves that signal the reasons they were chosen by the authors of a scientific paper. To address this question, we generated a dataset that characterizes both citing papers and references. We acquired the full text of 156,558 articles published in PLoS journals between 2005 and 2016 (we included only primary research articles, and excluded reviews, editorials and corrections), and combined these data with author-disambiguated records of the citation history of all publications, whether they were citing, cited or both (Supplementary Methods 1, Supplementary Fig. 1). This information was obtained from the Web of Science database, which contains over 60 million records on scholarly publications (as of January 2017).

We extracted a total of 2,320,774 unique references that were used a total of 5,787,630 times across all sections of all PLoS papers (see the Data section in the Methods). We then identified the exact location within the text where a given reference was cited. The names and section labels allowed us to align the content of each paper in our dataset according to the typical four-section structure: introduction, methods, results, and discussion (see the Data section in the Methods for statistics on other rarely used types of sections). The 5,787,630 data entries in our dataset are referred to here as 'records'; each record comprised the following four fields: reference ID, citing paper ID, section and reference age. We defined the age of a reference as the

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA. [2]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, USA. [3]Department of Physics and Astronomy, Northwestern University, Evanston, IL, USA.
*e-mail: amaral@northwestern.edu

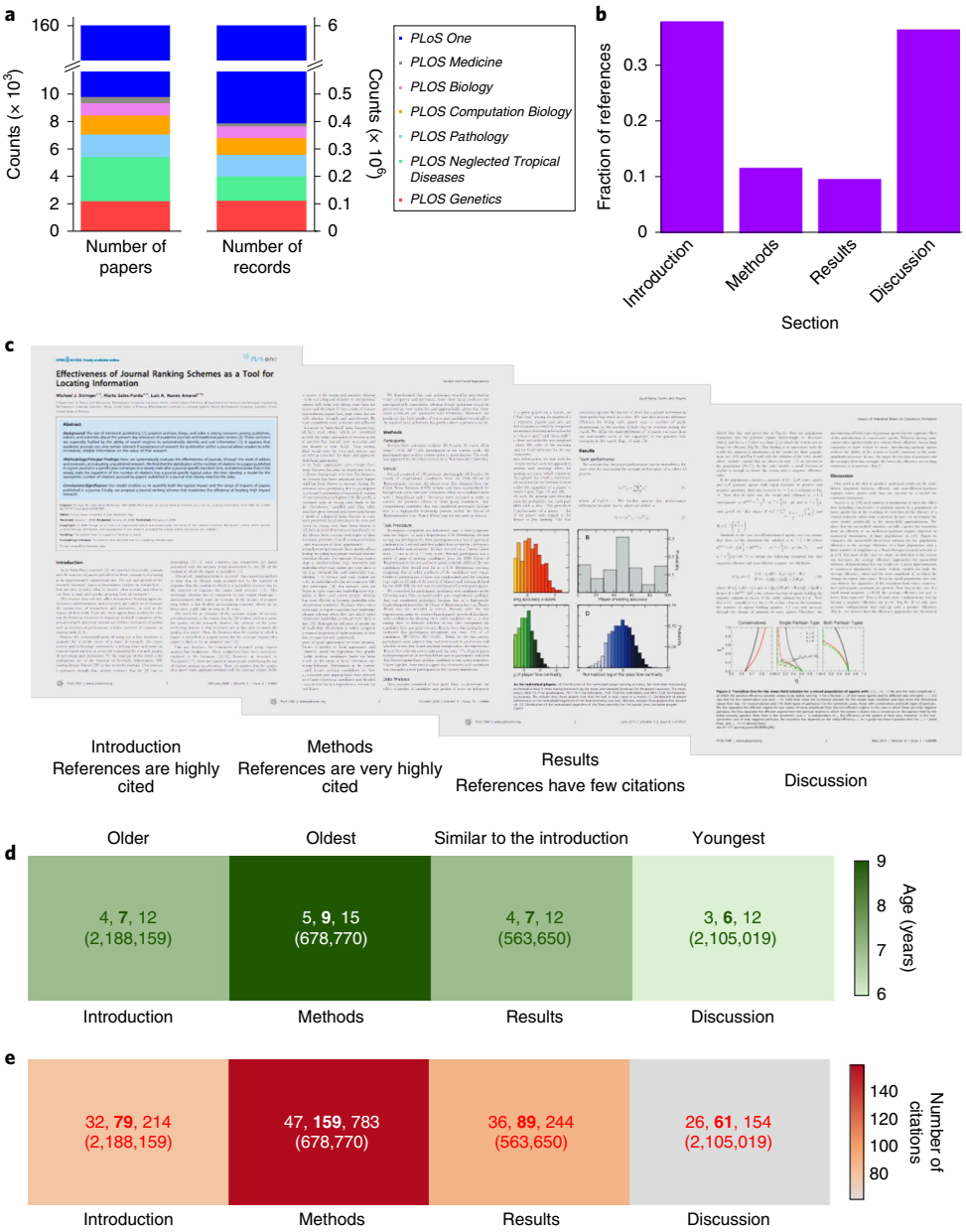

Fig. 1 | References vary in age and impact according to the section in which they are cited. a, The number of published papers (left) and records (right) in each of the PLoS journals that were included in our dataset. **b**, The fraction of references used in each article section from all of the papers in our dataset. **c**, Hypothesized characteristics of the references cited by a paper depend on the section in which they are cited. Credit: images in **c** reproduced from refs. [45,50–52] (left–right), PLoS. **d**,**e**, Age statistics (**d**) and numbers of citations received (**e**) for all references used in our dataset according to the section of the paper in which they are cited. Age and citation values are shown as the 25th (P_{25}), 50th (P_{50}) and 75th (P_{75}) percentiles (50th percentiles are bolded as they are the mean); values in parentheses indicate the number of records in each bin.

difference between the years of publication of the citing paper and of the reference. We excluded all self-citation records, because they may have been selected for different reasons than other references[39]. This resulted in the removal of 1,136,639 records, or 16%, a percentage consistent with previously reported values[40].

Around 74% of the references in our dataset are used in either the introduction or discussion (Fig. 1c). The references used in the methods are the most highly cited as well as the oldest, whereas the references in the discussion are younger than other references in the papers (Fig. 1d,e).

Interestingly, we also found significant differences in the citation practices of highly cited papers and poorly cited papers that were published in the same year and in the same journal. We find that the authors of highly cited papers tend to cite significantly younger references across all sections, and that papers from higher citation percentiles cite more highly cited references (see Supplementary Figs. 8–10 for P values of all possible pairwise comparisons using the Mann–Whitney U-test; Supplementary Methods 2). Previous research has suggested the notion that the authors of highly cited papers strategically select a successful combination of references[41]. We expand on this idea, and suggest that these authors of highly cited papers are more adept at identifying impactful research, and do so early on. One could argue that good scientists have good 'scientific taste'.

Although papers published in *PLoS One* constitute a substantial proportion of the dataset (Fig. 1a), we verified that we could still

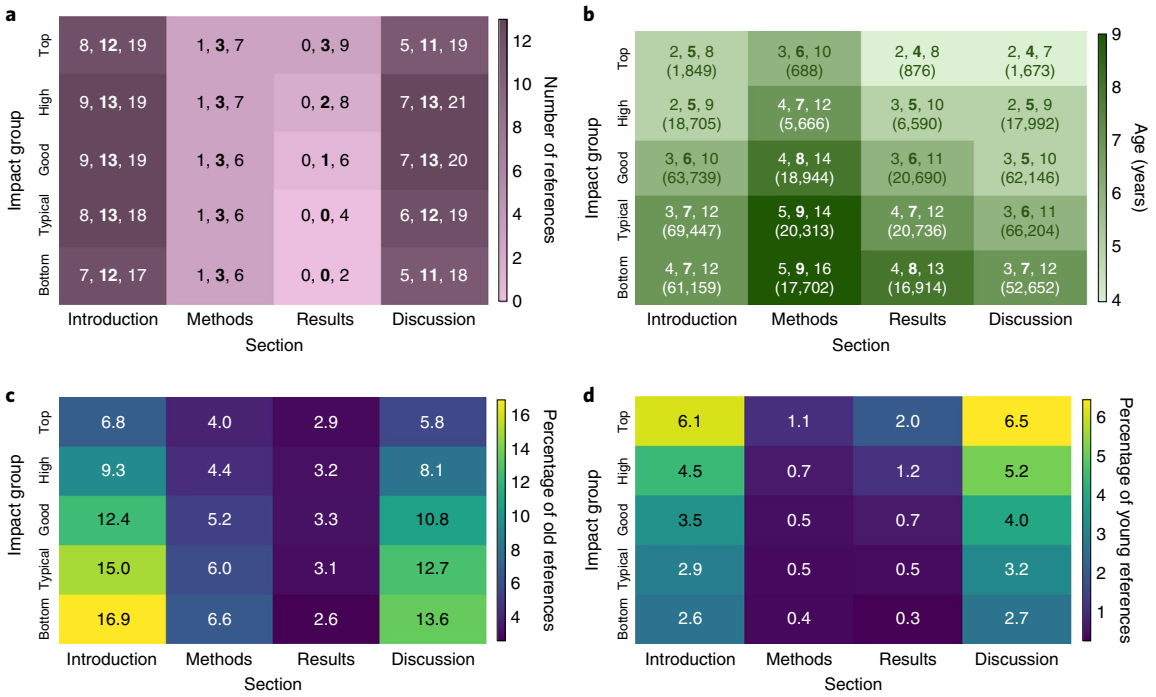

Fig. 2 | The number of references used in each section is mostly independent of the paper's impact group, although authors of highly cited PLoS papers cite younger references and use a higher percentage of young references. a, The number of references cited, given as P_{25}, P_{50}, P_{75}, by section and impact group for all papers in our dataset ($n=156{,}558$). **b**, Age in years of the references, given as P_{25}, P_{50}, P_{75}, cited by papers published in PLoS journals in 2011 ($n=14{,}351$ papers citing 357,866 unique references, yielding 564,251 records). Values in parentheses indicate the number of records in each bin. The differences are statistically significant for most pairwise comparisons using a Mann–Whitney *U*-test[53] (Supplementary Fig. 8). **c**,**d**, Average percentage of old references (**c**; $n=300{,}035$) and young references (**d**; $n=953{,}394$) cited by the 156,558 papers in our dataset, by section and by impact group. Darker colours indicate higher values.

obtain qualitatively similar results when conditioning on the basis of particular journals or scientific fields (Supplementary Figs. 2 and 3). The papers in our dataset cite an average of 48 references ($\sigma=26$; interquartile interval (R_{25-75}) = 34–58). As expected, these citations are not distributed uniformly across the different sections of a paper. We found that 38% of the references are cited in the introduction, 36% are cited in the discussion, and 12% and 10% are cited in the methods and results, respectively (Fig. 1b). These values are consistent with previous reports on the location of citations[11]. We also found that cited references are an average of 9.7 years old ($\sigma=9.0$, $R_{25-75}=3$–13), and that the average number of citations for all cited references is 83 ($\sigma=391$, $R_{25-75}=17$–82).

The average number of citations that we calculated is much larger than the expected average value for all papers recorded in the Web of Science. First, by focusing on the citation patterns of papers, we ignored those that did not receive any citations. It has been estimated that over one-quarter of papers never receive a single citation[34]. Second, this large average is not surprising because, by construction, we consider here a biased subset of papers that includes only those that have been cited at least once in a PLoS publication, which suggests that they may have been cited by other publications. Third, when constructing the set of references that were cited in a given section, we included them multiple times if they were cited by multiple papers. This resulted in a larger median number of citations than if we had included cited papers only once.

Going beyond a purely descriptive picture, we next investigated whether the location where a reference is cited informs about the reason it was selected. We hypothesized that scholars select the references to cite in different sections of a manuscript for section-specific reasons. For example, there tends to be a broad, long-lasting consensus on what methods are most appropriate for the study of a given problem. Thus, we expect references cited in the methods to be older and more highly cited than references used in other sections of the manuscript (Fig. 1c). By contrast, in the discussion, the authors may compare their findings with those reported in other studies, and they may also delineate future research directions, so it is likely that the references will tend to be younger. The references cited in the introduction may be used to display the authors' knowledge of the set of papers that define a field of research, and provide the background for the current work. Thus, we expect those references to trend towards being older and more highly cited. In the results section, because authors may connect their findings to work cited in the introduction, we expect references to overlap with those in that section. More importantly, 'better' researchers may have a better grasp of current work in their field, and may also be better at identifying promising new methods, ideas and unanswered questions, therefore selecting better references, and maybe even younger references than their peers.

We found substantial variation in the number of citations of references used in different sections of papers (Fig. 1e, Supplementary Figs. 6 and 7), ranging from an average of 1,952 in the methods ($P_{50}=159$, where P_n is the nth percentile) to an average of 188 in the discussion ($P_{50}=61$). These findings are consistent with previous anecdotal evidence that methods papers are the most highly cited[42]. Conditioning on the basis of section, we found that the average reference age (Fig. 1d, Supplementary Figs. 4 and 5) ranges from 11.8 years in the methods ($P_{50}=9$) to 8.6 years in the discussion ($P_{50}=6$).

We thus found a robust micro-level pattern for the selection of references across the four typical sections of a paper. The introduction and discussion sections include the highest fraction of references. The references cited in the introduction are older and highly

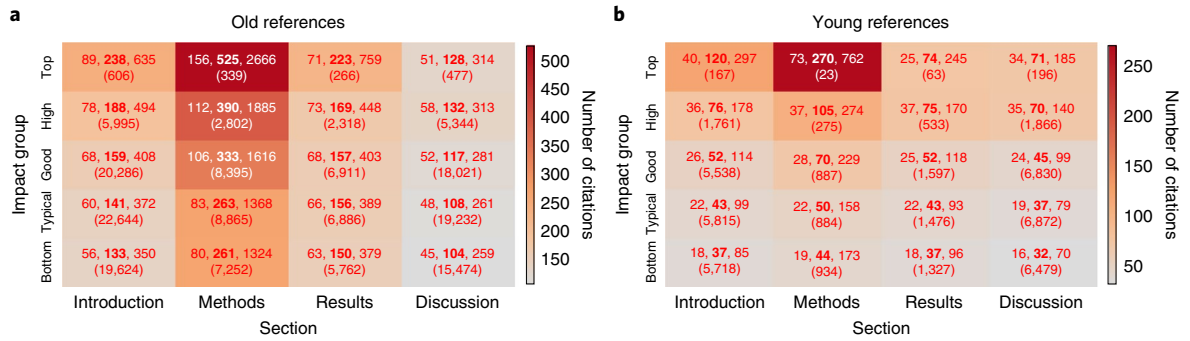

Fig. 3 | Authors of highly cited PLoS papers cite more highly cited references, especially in the case of young references. a, The number of citations of old references that were cited by 14,028 PLoS papers published in 2011. These papers cited 123,212 unique old references and yielded 183,640 records. **b**, The number of citations of young references that were cited by 11,282 PLoS papers published in 2011. These papers cited 34,436 unique young references and yielded 51,289 records. The difference between the median values are statistically significant for most pairwise comparisons using a Mann–Whitney *U*-test (Supplementary Figs. 9 and 10). The number of citations are given as P_{25}, P_{50}, P_{75}; the numbers of records in each bin are shown in parentheses. Darker colours indicate higher numbers.

cited whereas the references in the discussion tend to be the youngest. The methods contains few references, but they have the highest impact and age.

Next, we investigated whether authors of highly cited PLoS papers select their references differently from those of poorly cited PLoS papers. We grouped papers according to the number of citations that they had accumulated as of 31 December 2017 into five exclusive 'impact' groups as follows: bottom 30% (bottom), 31% to 60% (typical), 61% to 90% (good), 90% to 99% (high) and the top 1% (top). Owing to the time necessary to accrue citations—and thus the inherent time dependence of classification into impact groups—we restricted the comparison to PLoS papers that were published in the same year.

First, we analysed the number of references that were cited in each section. It was visually apparent that there were no relevant differences by impact group (Fig. 2a)—most references were cited in the introduction (the median number of references was consistently between 12 and 13), followed by the discussion (the median number of references was consistently between 11 and 13), regardless of the impact group that the paper belonged to. The only noticeable difference is the near absence of citations in the results section of low-impact papers; the median number of references used in the results by papers in the bottom impact group was 0, versus 3 for papers in the top impact group and 2 for papers in the high impact group. Next, we analysed the age of the references in PLoS papers for the different sections. We find that, independently of the citation group, the references cited in the methods are consistently the oldest, whereas references cited in the discussion are consistently the youngest (Fig. 2b). Interestingly, we also found a clear trend in the age of the references when we considered higher impact groups. Whereas references cited by authors of papers in the bottom impact group had a median age of 9 years old (in the methods section) and 7 years old (in the discussion section), the median age of the references cited by the papers in the top group were only 6 years old and 4 years old for methods and discussion sections, respectively. This difference is particularly striking given that we excluded all self-citations—we expect that self-citations are generally younger than other citations. We also reproduced these age patterns for different publication years, journals and fields (Supplementary Figs. 15, 18 and 21), as well as for alternative definitions of impact groups (Supplementary Fig. 25).

Motivated by the differences observed in the ages of the cited references between the different impact groups, we analysed the percentages of old (age $\geq$ 10 yr) and young (age $\leq$ 1 yr) references that were cited by papers. Although we found no differences for the methods and results sections, we found for the introduction and discussion sections that top impact papers use a lower percentage of old references and a higher percentage of young references than bottom impact papers (Fig. 2c,d).

These systematic differences in the selection of references by authors of papers from different impact groups raises the question of whether top PLoS papers cite not only younger references, but also references with an ultimately higher impact. We therefore analysed the number of citations of the references cited in PLoS papers for the different paper sections and impact groups. To control for the differences in the age of references across groups, we focused separately on old and young references.

It is visually apparent that the most-cited old references in a PLoS paper—regardless of impact group—are those found in the methods section (Fig. 3a). Interestingly, however, we found that the more highly cited papers cite more highly cited references across all sections. For example, the median numbers of citations of old and young references cited in the introduction of papers in the top impact group are 238 and 120, respectively, compared to 133 and 37 for old and young references cited in the introduction of papers in the bottom impact group, respectively (Fig. 3b). This finding is particularly important because—at the time of publication of the citing paper—the number of citations of a young reference will be quite low, as that reference was probably published around the time at which the citing papers were being submitted (see Supplementary Fig. 11 for a heat map showing early citations of young references). Thus, this finding suggests that authors of highly cited PLoS papers select higher-impact references, both in the case of references that were presumably well-known at the time of the publication of the PLoS paper, and in the case of references that were 'fresh off the press'. We reproduced this pattern for different definitions of old and young references (Supplementary Fig. 14) and subsets of the data conditioned on the basis of year of publication (Supplementary Figs. 16 and 17), PLoS journals (Supplementary Figs. 19 and 20), scientific disciplines of PLoS papers as provided by PLoS (Supplementary Figs. 22 and 23) or for references as inferred from the text of abstracts using topic models[43] (Supplementary Figs. 34 and 35, Supplementary Methods 3). We considered individual and group references (Supplementary Fig. 24), and different choices for binning into impact groups (Supplementary Figs. 26 and 27). Moreover, we found that for the young references cited in PLoS publications during a given year—whether computing their number of citations as of 2016, after 8 years or at 1 year old—the citation patterns described are qualitatively the same (Supplementary Figs. 11–13).

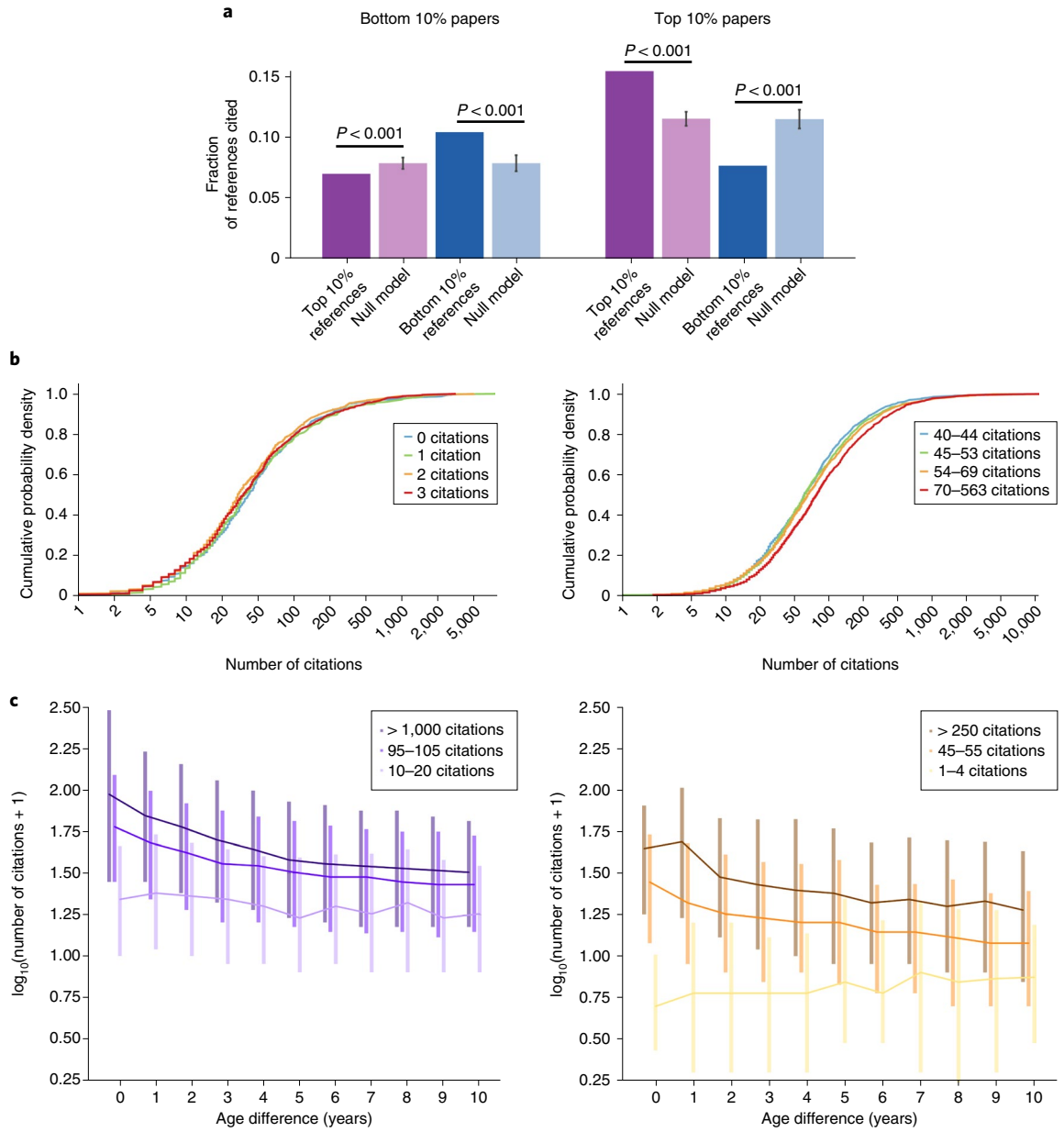

Fig. 4 | Highly cited papers have a higher-than-expected probability of citing highly cited references, and a lower-than-expected probability of citing poorly cited references. a, Comparison of the utilization of highly cited (top 10%; R_{25-75} = 50–1,173 citations) and poorly cited (bottom 10%; R_{25-75} = 0–5 citations) references by highly cited and poorly cited biological sciences PLoS papers that were published in 2008 versus the expected value from a null model. Error bars indicate the 95% confidence interval obtained from 1,000 replications of the randomization procedure. **b**, Cumulative distribution of the number of citations accrued by young references of PLoS papers with different numbers of citations. We separated the bottom 10% (left) and top 10% (right) PLoS papers from 2011 into four mutually exclusive groups each (see Supplementary Figs. 31 and 32 for different years). **c**, Dependence on age difference between citing and cited papers of the logarithm of number of citations (+1) of the papers that cite subsets of papers that were published in the period 1990–1995 in *PNAS* (left) and *Physical Review E* (right). The box plots show R_{25-75}. See Supplementary Fig. 33 for four other journals (*Physical Review Letters*, *Nature*, *Science* and *Journal of Theoretical Biology*).

The differences in the future impact of young references that were selected in papers from different impact groups is striking. To establish the robustness of these differences, we compared our results with an appropriate null model. Specifically, we checked whether the utilization rates of both top 10% and bottom 10% impact group references by either top 10% or bottom 10% impact PLoS papers in a given year and field were different from what would be expected if selection was random—that is, if authors of papers simply pulled references from the pool of available references in their field at the time of publication without any thought to quality or timeliness (see the Randomization section in the Methods; we also test other top percentages for top and bottom definitions, with quantitatively the same results). To ensure the validity of our findings, we preserved the structure of how references are grouped in the citing paper. Figure 4a shows our results for 2,008 PLoS papers in the field of biological sciences.

We counted the fraction of 'top 10%' and 'bottom 10%' references cited, we defined these groups as references that are in the top

and bottom 10% citation percentile, respectively. In comparison to the random null model, utilization of highly cited references is significantly increased in the top 10% group of papers (15.4% versus 11.5%, respectively, $P < 0.0001$) whereas utilization of poorly cited references is significantly decreased (7.6% versus 11.5%, respectively, $P < 0.0001$). For the bottom 10% group of PLoS papers, we found the opposite pattern: utilization of papers in the top 10% and bottom 10% groups of references is significantly lower, or higher, than the in the null model, respectively (for both comparisons, $P < 0.0001$). These results are robust with respect to the paper's publication year, different percentage thresholds for the definition of top and bottom publications (Supplementary Fig. 28), and hold for each subject field separately and for different publication years (Supplementary Figs. 29 and 30). We observed that the fraction of top references used by bottom papers is significantly lower—but only slightly lower than—expected by chance; we can say that the bottom papers do not avoid citing highly cited references. Notably, the null model expectations for the fraction of references in the bottom 10% category is lower than the corresponding one for papers in the top 10% category. This is due to the fact that papers in higher impact groups cite more references—papers in the bottom 10% category that were published in 2011 cited an average of 43 references, whereas papers in the top 10% category cited an average of 54 references.

To uncover the mechanism by which more highly cited references are selected by high-impact publications, we next investigated the possibility of an 'endorsement effect'[44] from papers to their references. That is, could it be that the success of high-impact PLoS papers is driving the success of the young references they cite? To answer this question, we first analysed the relationship between the number of citations accrued by young references and the number of citations accrued by the citing papers (Fig. 4b). In all cases, we observed that the median number of citations of the references is much larger than the number of citations of the citing PLoS papers of a given year. First, even for the most highly cited PLoS papers (the top 10%), the median number of citations of their references is about two times larger. Second, focusing on the number of citations that young references accrue during the first year after they were published revealed the same pattern (see Supplementary Fig. 11 for complete results). These results are not consistent with the hypothesis of an 'endorsement effect'.

Subsequently, we tested whether our finding that the authors of high-impact PLoS papers tend to select more highly cited references—and tend to do so earlier than unsuccessful ones—holds for other journals as well. In particular, we compared general, high-profile journals, such as *Proceedings of the National Academy of Sciences of the USA* (*PNAS*), with specialized, medium-impact journals, such as *Physical Review E*. For each studied journal, we considered papers that were published between 1990 and 1995, to allow papers citing them to have accrued their ultimate number of citations[45]. We then partitioned these considered papers into subsets according to their impact. Figure 4c shows the distributions of citations of citing papers as a function of the age of the citing paper. It is visually apparent, and statistically significant (see Supplementary Tables 1–6 for the P values from the Kolmogorov–Smirnov test for all pairwise comparisons) that highly cited papers published in *PNAS* are cited by their most-cited papers shortly after their publication, and that over time citing papers have lower impact. This downward trend is also apparent, and statistically significant, for the subset of medium-impact papers published in *PNAS*, but is not present for the low-impact papers published in *PNAS*. In the case of *Physical Review E*, we find qualitatively similar but less pronounced trends (Fig. 4c, right). These findings further support our hypothesis that the authors of high-impact papers seem to be more selective and more timely when choosing the references for their papers.

Taken together, the results from our study show substantial and statistically significant differences in the selection of references by papers in the top impact groups versus the bottom impact groups. Higher impact papers tend to cite younger references, and a higher fraction of very young (less than 1 year old) references. Higher impact papers also select references that are already more highly cited or that will ultimately become more highly cited. Using the full text of a cohort of PLoS papers to trace the usage of references within the text, these findings hold across all sections of a paper. Considering the strong variations in the usage of references across sections, our much more in-depth analysis offers a resolution to the differing views on whether citation networks show such an assortative property when aggregating all citations[46,47].

One limitation of this study is the relatively short time span available for using PLoS papers to trace the usage of references in the full text (2005–2016; the annual number of publications exceeds 1,000 only in 2007). This yields a low resolution when tracing the usage of references across sections over time. The problem is exacerbated by the fact that one needs to wait approximately 10 yr to obtain a stationary distribution for a reliable estimate of the number of citations a reference receives[45]. We hope that the increasing availability of the full text for all papers in large scale bibliometric databases such as the Web of Science will allow improved temporal resolution.

If citations are considered as endorsements, then our results offer a quantitative approach to addressing the long-standing issue of whether all citations should be 'counted' equally. Moreover, when predicting the future impact of scientific work[21], our analysis suggests that it might be beneficial to take into account information about where in the text a reference occurred. Finally, our study expands on the previous hypothesis of a 'winning combination' of references that is strategically used by successful papers[41], suggesting that the authors of impactful papers have better 'scientific taste' when selecting the scientific foundation on which to build their own work.

Methods

Data. We collected the entire corpus of papers published by PLoS throughout the period 2005–2016 (see Supplementary Methods 1 for more details on data collecting, parsing and processing). PLoS is a non-profit, open-access publisher that covers a wide variety of subject areas. PLoS includes seven different peer-reviewed academic journals: the specialty journals *PLoS Medicine*, *PLoS Biology*, *PLoS Computational Biology*, *PLoS Genetics*, *PLoS Pathology* and *PLoS Neglected Tropical Diseases*, and the multidisciplinary journal *PLoS One*.

We obtained the full text of each article from PLoS through the PLoS text and data mining research API (http://api.PLoS.org). We made use of the fact that each article is classified into at least one of the nine top-level categories used by PLoS (biology and life Sciences, computational sciences, engineering, medicine, physical sciences, research and analysis methods, Earth sciences and ecology, social sciences, political sciences, and people and places). This yielded 156,558 papers (146,772 of them from the largest journal, *PLoS One*). The standardized XML format allowed us to unambiguously identify the structure (sections, paragraphs and others) and the location of every reference cited in the text of each paper. Using the XML tags provided by PLoS, we divided each paper into the following sections: introduction, methods, results and discussion (IMRaD), whenever necessary, we reordered the sections of a paper to conform with this structure.

We recorded how often each reference occurs in each of the sections of a paper. We considered only the first appearance of a given reference in a given section of a paper. A small fraction of references is placed in unlabelled parts of the text, domain-specific sections or in miscellaneous sections; we excluded these records from our analysis. Specifically, the fraction of references in each of the sections labelled with non-IMRaD names are: results–discussion, 3.45%; conclusions, 0.26%; mixed, 0.005%; and NA, 0.65%.

We matched citing papers and references with records from Web of Science (Clarivate Analytics) to obtain the number of citations (as of 2017), year of publication, journal and author list. Web of Science is still the most complete and accurate database for assessing the number of citations, it contains 59,679,483 records in total. First, we matched the citing papers across the two databases using the digital object identifier (DOI) provided by PLoS. This allowed us to unambiguously match 98% of the citing papers. Identifying the corresponding record of each citing paper yielded the list of references as indexed by Web of Science.

We matched the reference papers by comparing the title recorded in the PLoS data with the title provided in the Web of Science database. Specifically, we removed all punctuation, ignored capitalization and then used regular expressions to match the records. This procedure allowed us to match 85% of all references

across the two databases. We further identified and removed self-citations, which we defined as a citation to a paper that was authored by any of the authors of the citing PLoS paper. For this, we took advantage of unique author IDs in the Web of Science data[48] provided to us by the Distinct Author Identification System (Clarivate Analytics; see ref. [49] for details). We also excluded any PLoS papers that were not categorized as an article (such as reviews, editorials or corrections). In total, this procedure yielded 156,558 papers that cited 2,320,777 unique references that occurred 5,787,634 times across all sections.

We then used the Web of Science data to obtain the age and the number of citations (as of December 2017) for all papers and references.

Randomization. The randomization scheme that was used to establish the robustness of the differences between citation patterns of high and low-impact papers was as follows: we selected all records of references used by PLoS papers that were published in a given year from our data. Then, preserving the length of the reference list for each PLoS paper, we randomized the actual references included in each paper. We repeated this procedure 1,000 times for a given year, and we obtained the expected fraction of utilization of highly cited references and poorly cited references both by highly cited PLoS papers and by poorly cited PLoS papers.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data from PLoS are publicly available through its API (api.PLoS.org), the data from the Web of Science are available from Clarivate Analytics. We provide the conversion tables to link the DOIs of the PLoS papers used in this study, and the Web of Science unique IDs (of both the PLoS papers and the references they cite) here: https://doi.org/10.21985/N21X9J.

Code availability

Code for replication of all of our results is available via GitHub: https://github.com/juliettapc/my_In_text_citations.

Received: 5 September 2018; Accepted: 6 March 2019;
Published online: 15 April 2019

References

1. de Solla Price, D. J. Networks of scientific papers. *Science* **149**, 510–515 (1965).
2. Merton, R. K. The Matthew effect in science: the reward and communication systems of science are considered. *Science* **159**, 56–63 (1968).
3. Cronin, B. & Barsky Atkins, H. (eds) *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (ASIS, 2000).
4. Boyack, K. W., Klavans, R. & Börner, K. Mapping the backbone of science. *Scientometrics* **64**, 351–374 (2005).
5. Evans, J. A. & Foster, J. G. Metaknowledge. *Science* **331**, 721–725 (2011).
6. Zeng, A. et al. The science of science: from the perspective of complex systems. *Phys. Rep.* **714–715**, 1–73 (2017).
7. Bornmann, L. & Daniel, H. D. Selecting manuscripts for a high-impact journal through peer review: a citation analysis of communications that were accepted by *Angewandte Chemie International Edition*, or rejected but published elsewhere. *J. Am. Soc. Inf. Sci. Technol.* **59**, 1841–1852 (2008).
8. Radicchi, F., Weissman, A. & Bollen, J. Quantifying perceived impact of scientific publications. *J. Informetr.* **11**, 704–712 (2017).
9. Yegros-yegros, A., Lamers, W. S., Eck, N. J. V., Waltman, L. & Hoos, H. Patterns in citation context: the case of the field of scientometrics. In *Proc. 23rd International Conference on Science and Technology Indicators* 1115–1122 (2018).
10. Boyack, K. W., van Eck, N. J., Colavizza, G. & Waltman, L. Characterizing in-text citations in scientific articles: a large-scale analysis. *J. Informetr.* **12**, 59–73 (2018).
11. Bertin, M., Atanassova, I., Gingras, Y. & Lariviere, V. The invariant distribution of references in scientific articles. *J. Assoc. Inf. Sci. Technol.* **67**, 164–177 (2016).
12. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl Acad. Sci. USA* **105**, 1118–1123 (2008).
13. Guimerà, R., Uzzi, B., Spiro, J. & Amaral, L. A. N. Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308**, 697–702 (2005).
14. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
15. Malmgren, R. D., Ottino, J. M. & Amaral, L. A. N. The role of mentorship in protégé performance. *Nature* **465**, 622–627 (2010).
16. Zeng, X. H. T. et al. Differences in collaboration patterns across discipline, career stage, and gender. *PLoS Biol.* **14**, e1002573 (2016).
17. Iacopini, I., Milojević, S. & Latora, V. Network dynamics of innovation processes. *Phys. Rev. Lett.* **120**, 048301 (2018).
18. Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B. Atypical combinations and scientific impact. *Science* **342**, 468–472 (2013).
19. Ahmadpoor, M. & Jones, B. F. The dual frontier: patented inventions and prior scientific advance. *Science* **357**, 583–587 (2017).
20. Acuna, D. E., Allesina, S. & Kording, K. P. Predicting scientific success. *Nature* **489**, 201–202 (2012).
21. Wang, D., Song, C. & Barabási, A.-L. Quantifying long-term scientific impact. *Science* **342**, 127–132 (2013).
22. Petersen, A. M. et al. Reputation and impact in academic careers. *Proc. Natl Acad. Sci. USA* **111**, 15316–15321 (2014).
23. Moreira, J. A. G., Zeng, X. H. T. & Amaral, L. A. N. The distribution of the asymptotic number of citations to sets of publications by a researcher or from an academic department are consistent with a discrete lognormal model. *PLoS One* **10**, e0143108 (2015).
24. Wasserman, M., Zeng, X. H. T. & Amaral, L. A. N. Cross-evaluation of metrics to estimate the significance of creative works. *Proc. Natl Acad. Sci. USA* **112**, 1281–1286 (2015).
25. Tahamtan, I., Safipour Afshar, A. & Ahamdzadeh, K. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics* **107**, 1195–1225 (2016).
26. Milojević, S. How are academic age, productivity and collaboration related to citing behavior of researchers? *PLoS One* **7**, e49176 (2012).
27. Gingras, Y., Larivière, V., Macaluso, B. & Robitaille, J. P. The effects of aging on researchers' publication and citation patterns. *PLoS One* **3**, e4048 (2008).
28. West, J. D., Jacquet, J., King, M. M., Correll, S. J. & Bergstrom, C. T. The role of gender in scholarly authorship. *PLoS One* **8**, e66212 (2013).
29. Bornmann, L. & Daniel, H. What do citation counts measure? A review of studies on citing behavior. *J. Doc.* **64**, 45–80 (2008).
30. Valenzuela, M., Ha, V. & Etzioni, O. Identifying meaningful citations. In *AAAI Workshops* https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/viewPaper/10185 (2015)..
31. Popper, K. R. The nature of philosophical problems and their roots in science. *Br. J. Philos. Sci.* **3**, 124–156 (1952).
32. Krapivsky, P. L. & Redner, S. Network growth by copying. *Phys. Rev. E* **71**, 036118 (2005).
33. Redner, S. Citation statistics from more than a century of physical review. *Phys. Today* **58**, 49 (2005).
34. Stringer, M. J., Sales-Pardo, M. & Nunes Amaral, L. A. Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in 26 a scientific journal. *J. Am. Soc. Inf. Sci. Technol.* **61**, 1377–1385 (2010).
35. Liang, L., Zhong, Z. & Rousseau, R. Scientists' referencing (mis)behavior revealed by the dissemination network of referencing errors. *Scientometrics* **101**, 1973–1986 (2014).
36. Roach, V. J., Lau, T. K., Kee, W. D. N. & Kong, H. The quality of citations in major international obstetrics and gynecology journals. *Am. J. Obstet. Gynecol.* **177**, 973–975 (1997).
37. Davies, K. Reference accuracy in library and information science journals. *Aslib Proc.* **64**, 373–387 (2012).
38. Dias, L., Gerlach, M., Scharloth, J. & Altmann, E. G. Using text analysis to quantify the similarity and evolution of scientific disciplines. *R. Soc. Open Sci.* **5**, 171545 (2018).
39. Aksnes, D. W. A macro study of self-citation. *Scientometrics* **56**, 235–246 (2003).
40. Ioannidis, J. P. A. A generalized view of self-citation: direct, co-author, collaborative, and coercive induced self-citation. *J. Psychosom. Res.* **78**, 7–11 (2015).
41. Mukherjee, S., Romero, D. M., Jones, B. & Uzzi, B. The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: the hotspot. *Sci. Adv.* **3**, e1601315 (2017).
42. Van Noorden, R., Maher, B. & Nuzzo, R. The top 100 papers. *Nature* **514**, 550–553 (2014).
43. Gerlach, M., Peixoto, T. P. & Altmann, E. G. A network approach to topic models. *Sci. Adv.* **4**, eaaq1360 (2018).
44. Garfield E. The use of journal impact factors and citation analysis for evaluation of science. In *Proc. Cell Separation, Hematology and Journal Citation Analysis, Mini Symposium in Tribute to Arne Bøyum* (1998).
45. Stringer, M. M. J., Sales-Pardo, M. & Amaral, L. A. N. Effectiveness of journal ranking schemes as a tool for locating information. *PLoS One* **3**, e1683 (2008).
46. Bornmann, L., de Moya Anegón, F. & Leydesdorff, L. Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. *PLoS One* **5**, e13327 (2010).
47. Šubelj, L. & Bajec, M. Model of complex networks based on citation dynamics. In *Proc. 22nd International Conference on World Wide Web* 527–530 (ACM, 2013).
48. Zhao, Z., Rollins, J., Bai, L. & Rosen, G. Incremental author name disambiguation for scientific citation data. In *Proc. 2017 International Conference on Data Science and Advanced Analytics* 175–183 (2018).
49. Clarivate Analytics. *Web of Science Raw Data (XML): User Guide for Web of Science Raw Data* Clarivate.com https://clarivate.libguides.com/c.php?g=593069&p=4220414 (2016).

50. Chiao, J. Y., Bowman, N. E. & Gill, H. The political gender gap: gender bias in facial inferences that predict voting behavior. *PLoS One* **3**, e3666 (2008).
51. Duch, J., Waitzman, J. S. & Amaral, L. A. N. Quantifying the performance of individual players in a team activity. *PLoS One* **5**, e10937 (2010).
52. Sales-Pardo, M., Diermeier, D. & Amaral, L. A. N. The impact of individual biases on consensus formation. *PLoS One* **8**, e58989 (2013).
53. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).

Acknowledgements

L.A.N.A. thanks the John and Leslie McQuown Gift and support from the Department of Defense Army Research Office under grant number W911NF-14-1-0259. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

J.P.-C. contributed to the data preparation, wrote the codes for data analysis, statistical testing and figure plotting, contributed to the interpretation of the results and drafted the manuscript. N.A. collected, cleaned and prepared the data and performed preliminary analysis. M.G. contributed to the collection and the analysis of the data, contributed to the interpretation of the results and drafted the manuscript. L.A.N.A. conceived and designed the study, contributed to the interpretation of the results and drafted the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0585-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.A.N.A.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Luis A. N. Amaral

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	We collect the entire corpus of papers published by the Public Library of Science (PLOS) in the period 2005--2016 via the PLOS Text and Data Mining research API (http://api.plos.org). We match citing papers and references with records from Clarivate Analytics' Web of Science in order to obtain number of citations (as of 2017), year of publication, journal, and author list.
Data analysis	Custom Python code using standard packages including numpy, scipy, and pandas.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data from PLOS is publicly available via its API (\href{http://api.plos.org/}{api.plos.org}), the data from Web of Science is available from Clarivate Analytics. Summary tables for replication of all of our results will be made publicly available upon publication.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We study the usage of citations in scientific articles in PLOS and trace their bibliometric history from among 60 million records from the Web of Science.
Research sample	We acquire the full text of 156,558 articles published in PLOS journals between 2005 and 2016. We extract a total of 2,320,774 unique references, that are used a total of 5,787,630 times across all sections of all papers.
Sampling strategy	We consider all articles published in PLOS between 2005 and 2016 including only primary research articles, i.e. excluding reviews, editorials and corrections. In those articles, we consider all references that can be matched to an entry in the Web of Science database removing self-citations.
Data collection	Scientific articles from PLOS were downloaded in xml-format via the PLOS API. The cross-referencing with the Web of Science database was performed with custom Python code.
Timing	Data from PLOS was downloaded in January 2017.
Data exclusions	We consider all articles published in PLOS between 2005 and 2016 including only primary research articles, i.e. excluding reviews, editorials and corrections. In those articles, we consider all references that can be matched to an entry in the Web of Science database removing self-citations.
Non-participation	n/a
Randomization	n/a

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging