



## OPEN ACCESS

## EDITED BY

Jiangen He,  
The University of Tennessee, Knoxville,  
United States

## REVIEWED BY

Petr Heneberg,  
Charles University, Czechia  
Antonio Perianes-Rodriguez,  
Universidad Carlos III de Madrid, Spain

## \*CORRESPONDENCE

Sean M. Cascarina  
[Sean.Cascarina@colostate.edu](mailto:Sean.Cascarina@colostate.edu)

RECEIVED 01 May 2023

ACCEPTED 24 August 2023

PUBLISHED 22 September 2023

## CITATION

Cascarina SM (2023) Self-referencing rates in biological disciplines. *Front. Res. Metr. Anal.* 8:1215401.  
doi: 10.3389/frma.2023.1215401

## COPYRIGHT

© 2023 Cascarina. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Self-referencing rates in biological disciplines

Sean M. Cascarina\*

Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, CO, United States

The use of citation counts (among other bibliometrics) as a facet of academic research evaluation can influence citation behavior in scientific publications. One possible unintended consequence of this bibliometric is excessive self-referencing, where an author favors referencing their own publications over related publications from different research groups. Peer reviewers are often prompted by journals to determine whether references listed in the manuscript under review are unbiased, but there is no consensus on what is considered "excessive" self-referencing. Here, self-referencing rates are examined across multiple journals in the fields of biology, genetics, computational biology, medicine, pathology, and cell biology. Median self-referencing rates are between 8–13% across a range of journals within these disciplines. However, self-referencing rates vary as a function of total number of references, number of authors, author status/rank, author position, and total number of publications for each author. Importantly, these relationships exhibit interdisciplinary and journal-dependent differences that are not captured by examining broader fields in aggregate (e.g., Biology, Chemistry, Physics, etc.). These results provide useful statistical guidelines for authors, editors, reviewers, and journals when considering referencing practices for individual publications, and highlight the effects of additional factors influencing self-referencing rates.

## KEYWORDS

self-referencing, self-citation, bibliometrics, research incentives, citation analysis, self-references

## Introduction

The use of bibliometrics (i.e., quantitative metrics applied to publication records) to assess research success has become commonplace. In principle, data-driven methods of evaluating individual researchers could contribute to more objective, consistent, comprehensive, and pragmatic assessments (Abramo and D'Angelo, 2011), but they are also prone to manipulation (Ioannidis, 2015), can inadvertently distort research incentives (Moher et al., 2018), and can be misused in the decision-making process (Hicks et al., 2015). For these reasons, recent guidelines on individual researcher assessment advocate integrating multiple, complementary quantitative measures (Ioannidis et al., 2016), as well as balancing these quantitative measures with qualitative, expert assessment of researcher contributions (Hicks et al., 2015; Moher et al., 2018).

Bibliometrics are often designed to approximate the "impact" of an individual within a scientific discipline (Wildgaard et al., 2014). One such statistic, the number of times a researcher's published work has been cited, is often an integral part of bibliometrics that attempt to estimate impact (Wildgaard et al., 2014; Waltman, 2016; Aksnes et al., 2019). While it is reasonable to expect influential

ideas and discoveries to be cited frequently, the potential motivations for citing a particular publication are numerous and are not always based purely on the scientific content of that publication (Bornmann and Daniel, 2008; Aksnes et al., 2019). Additionally, citations can be influenced by the authors themselves via self-referencing or “citation rings” (Fowler and Aksnes, 2007; Zhivotovsky and Krutovsky, 2008; Bartneck and Kokkelmans, 2011; Ioannidis et al., 2019; Van Noorden and Singh Chawla, 2019), by reviewers (Resnik et al., 2008; Thombs et al., 2015; Wren et al., 2019; Wren and Georgescu, 2022), by editors (Fong and Wilhite, 2017; Chaplain et al., 2020), and by journals (Wilhite and Fong, 2012; Van Noorden, 2013; Heneberg, 2014, 2016; Yu et al., 2014; Fong and Wilhite, 2017; Taşkin et al., 2021) (for additional discussion on all of these categories, see Ioannidis, 2015; Lockwood, 2020). Influence of citations is not always unethical, but requires a delicate (and often subjective) balance in appropriate cases. For example, peer reviewers can recommend adding references to their own work when evaluating a manuscript, but this should only occur in cases where the references are truly pertinent to the manuscript, and exclusion of suggested references generally should not be grounds for rejection (Thombs and Razykov, 2012).

Self-referencing can occur among publishers, individual journals, collections of authors, and individual authors, which can affect corresponding bibliometrics (Zhivotovsky and Krutovsky, 2008; Bartneck and Kokkelmans, 2011; Heneberg, 2014, 2016; Copiello, 2019; Ioannidis and Thombs, 2019; Taşkin et al., 2021). At each level, the motivations for self-referencing vary and can include both legitimate and illegitimate reasons. At the level of publishers and journals, self-referencing involves referencing an article previously published by journals from the same publisher or articles from the same journal, respectively. A number of publisher and journal self-references are to be expected: publishers often have multiple journals spanning similar subjects, and individual journals often specialize in particular topics, both of which increase the likelihood of publisher/journal self-referencing. Additionally, authors whose specialty fits a journal’s scope are likely to both publish in that journal and reference publications from the same journal. However, bibliometrics (particularly, the impact factor) for journals and publishers also incentivize self-referencing, potentially influencing the share of published content devoted to different types of articles for some journals (Heneberg, 2014; Ioannidis and Thombs, 2019).

For individual authors, there are many possible motivations for choosing a reference (Garfield, 1965; Brooks, 1985, 1986; Vinkler, 1987), and the motivations for self-referencing and non-self-referencing overlap (Bonzi and Snyder, 1991). Motivations for author self-referencing include (but are not limited to): (1) referencing prior publications that were foundational to the current work; (2) additional discussion or treatment of prior work (e.g., reviews, historical accounts, perspectives, etc.); (3) referencing a method previously developed; (4) refutations, corrections, retractions, or corrigenda pertaining to prior publications; and (5) in inappropriate cases, unrelated self-references intended to boost bibliometric indicators for the authors (all adapted from Garfield, 1965; Bonzi and Snyder, 1991). These motivations, among others (Garfield, 1965; Bonzi and Snyder, 1991), are applied in varying degrees to each instance of

self-referencing, are not mutually exclusive, and are challenging to measure.

Motivations for self-referencing strongly overlap with intended functions of the self-references. For example, Zhao et al. (2018) divided references into self and non-self, then assigned each reference to one of 5 functional categories: Applied, Contrastive, Supportive, Reviewed, and Perfunctory. These categories closely parallel the motivations described above, as well as categories previously used to classify citations generally (Cano, 1989). In addition, the number of times each reference was cited within a single publication and the section(s) in which each citation appeared within that publication were evaluated. Based on a combined assessment of these data, they concluded that self-references are often of greater importance than non-self-references (Zhao et al., 2018). However, it should be noted that the percentage of self-references categorized as Perfunctory (~23%) was second only to the Applied category (32%) and was also higher than the percentage of non-self-references classified as Perfunctory (~14%). Therefore, while the largest share of self-references appears to be of greatest importance within a publication, a distinct fraction may also be of least importance (at least, according to the criteria applied in that study). For a comprehensive review of motivations for referencing and functions of references, see Bornmann and Daniel (2008) and Aksnes et al. (2019).

A number of studies have reported self-referencing rates for a variety of broad fields such as Biology, Chemistry, and Physics (Snyder and Bonzi, 1998; Hyland, 2001, 2003; Aksnes, 2003; Kacem et al., 2020; Szomszor et al., 2020), as well as narrower disciplines such as microbiology (Aksnes, 2003) and soil science (Ma et al., 2021). Authorship and citation practices, including self-referencing behaviors, vary by field and discipline (e.g., Snyder and Bonzi, 1998; Kacem et al., 2020; Szomszor et al., 2020; Rosenblatt et al., 2023). As a striking example, self-referencing rates were significantly lower in neuroscience compared to neurology and psychiatry (Rosenblatt et al., 2023), highlighting the importance of characterizing self-referencing behavior in a discipline-specific manner. To date, self-referencing has been examined in only a small fraction of biological disciplines and/or subdisciplines [e.g., neuroscience (Rosenblatt et al., 2023), neurology (Rosenblatt et al., 2023), neurobiology (Tagliacozzo, 1977), psychiatry (Rosenblatt et al., 2023), plant physiology (Tagliacozzo, 1977), and a variety of physiology subdisciplines (Larcombe and Voss, 2011)]. Therefore, it is important to ascertain what a “normal” self-referencing rate is in each discipline.

Here, self-referencing rates for individual authors are evaluated at the per-article level for >94 k publications from a subset of selected journals, representing a cross-section of biological disciplines over the past 20 years. Shifts in self-referencing rate distributions are examined as a function of journal, publication date, and total number of works referenced. Factors associated with unusually high self-referencing rates such as author status and total number of publications are also explored. Furthermore, the separation of disciplines enables the quantification of interdisciplinary and journal-specific differences relating these variables to self-referencing rates, which are not apparent when these disciplines are combined and evaluated as a single, broad field. These observations provide a range of quantitative guidelines for

authors, reviewers, editors, and journals based on recent historical self-referencing practices, while also highlighting instances of situationally appropriate extreme self-referencing.

## Materials and methods

### Quantitative indicators

For each publication evaluated in this study, a self-referencing rate is calculated separately for each author. For a given publication, the self-referencing rate ( $r$ ) for an individual author listed at position ( $i$ ) in the author list is defined as the percentage of total publications ( $p_t$ ) in the reference list that are also authored or co-authored by the same person ( $p_a$ ):

$$r_i = \frac{p_a}{p_t} \times 100$$

For the purposes of this study, the self-referencing rate associated with each publication is defined as the maximum  $r_i$  among all authors of that publication.

All percentile and quartile values (including medians) were calculated using the “scoreatpercentile” module in the Scipy package with “fraction” interpolation (Virtanen et al., 2020). Briefly, when a calculated score corresponding to a percentile of interest lies between two real values in the observed dataset (i.e., a non-integer position), the score ( $S$ ) is estimated according to the fractional distance between the two neighboring values:

$$S = i_x + (j_x - i_x) \times (x \bmod 1)$$

where  $x$  represents the position of the desired percentile in the sorted array (lowest value to highest value),  $i_x$  represents the value to the left of position  $x$  in the sorted array, and  $j_x$  represents the value to the right of position  $x$  in the sorted array. Percentiles were calculated separately for each journal without stratifying by publication year (with two exceptions—Figures 1B, D—where medians were stratified by journal and publication year).

### Data acquisition and processing

A complete list of PubMed IDs (PMIDs) was obtained from the NCBI website (<https://ftp.ncbi.nlm.nih.gov/pub/pmc/>) on 1/4/2023. For all PMIDs corresponding to the pre-defined set of journals, author lists were retrieved for each publication by a separate PubMed query. Subsequent queries were performed for all references within each publication to retrieve author lists for each reference. All PubMed queries were performed between 11/3/2022–1/23/2023 and included all publications formally published between the years 2002–2022 (inclusive) for each journal. PubMed queries were performed using the Entrez module in the Biopython package (Cock et al., 2009). Failed PubMed queries for both primary publications and referenced publications were repeated a minimum of 10 times. Biological disciplines were initially chosen based on the availability of publication data among suitable

representative journals. These disciplines include general biology, genetics, pathology/pathogens, medicine, computational biology, and cell biology, with each discipline represented by a pair of journals that exemplify that discipline. Additionally, three so-called “high-impact” journals were included to examine potential correlations between impact factor and self-referencing behavior.

For each successfully retrieved primary publication, a self-referencing rate was calculated for each author using a sequential string matching procedure. First, author names were checked for an exact string match. If no match was found, author names were subsequently truncated to various forms of first/middle initial and full surname and attempted to match against authors of the referenced publication (e.g., “John M. Smith” was truncated to “J M Smith”, “JM Smith”, and “J Smith”). Note that this subsequent step only identifies matches if the author names in the referenced publication are in the form of initials in the official PubMed query results. Although PubMed provides a downloadable open-access subset of full-text articles, the references therein only contain names with first/middle initials and full surname. Therefore, the method of performing individual PubMed queries was preferred over the more extensive downloadable database, as individual queries typically retrieve full author names which improve author disambiguation.

Journal impact factors were obtained manually for each journal from <https://jcr.clarivate.com/jcr/home>, accessed on 1/22/2023.

The ranked-author database was downloaded from the external link in the corresponding publication (Ioannidis et al., 2020; database version 5; last updated 10/28/2022). For all analyses involving author rank and/or total publications derived from the ranked-author database, only data corresponding to authors of the publications in this study’s dataset were included. Author rank was derived from the composite score corresponding to career-long, non-self-citations only. Authors with a numerical rank >1 million were excluded from analyses, as author rank values become increasingly sparse.

### Calculation of sliding means

To estimate shifts in percent self-reference distributions as a function of a second variable (e.g., total references; Figure 2), the 25<sup>th</sup>, 50<sup>th</sup> (median), and 75<sup>th</sup> percentiles were calculated for the distribution of percent self-references at each point along the  $x$ -axis (independent variable). Curves were then smoothed by calculating a sliding mean. Different window sizes ( $W$ ) for sliding means were required to account for differences in the scale of the independent variable:  $W = 11$  for total references (Figure 2) and number of authors (Figure 3);  $W = 21$  for total publications (Figure 5A);  $W = 5$  for percent of all “self” publications referenced (Figure 5B) and percentage of authors appearing in the ranked-authors database (Figure 6B); and  $W = 101$  for author rank (Figure 5C). Quartile values for percent self-references were only calculated if 3 or more publications had the specified number of total references ( $x$ -axis). Linear interpolation was performed across gaps in the data, with the following maximum gap sizes ( $G$ ):  $G = 10$  for total references (Figure 2) and number of authors (Figure 3);  $G = 50$  for total publications (Figure 5A);  $G = 5$  for percent of all

“self” publications references (Figure 5B) and percentage of authors appearing in the ranked-authors database (Figure 6B); and  $G = 5,000$  for author rank (Figure 5C). For all independent variables, linear interpolation was only necessary when reaching high x-axis values, where data become more sparse. When applicable, sliding means were calculated until a gap larger than the maximum gap size was encountered. For non-integer independent variables (e.g., percent of all “self” publications referenced), raw data were grouped into whole-integer bins using the numpy digitize function with *right = True* (i.e., with each bin including values equal to the right edge of that bin) prior to calculating quartile values and sliding means.

## Results

### Self-referencing rates are similar across biology-related disciplines

PubMed is a popular literature database for the life sciences, composed of  $>30$  million publications and typically searched  $>2.5$  billion times per year ([https://www.nlm.nih.gov/bsd/medline\\_pubmed\\_production\\_stats.html](https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html)). Although PubMed provides a complete set of downloadable, open-access articles, publications in a small subset of journals were used instead because individual PubMed queries improve author disambiguation for purposes of this study (see Materials and Methods). Analyses were initially focused on the *PLOS* (Public Library of Science) family of journals for five main reasons: (1) the core *PLOS* journals were founded shortly after the inception of PubMed, ensuring that most publications would be available for retrieval in automated searches; (2) specialized *PLOS* journals span a variety of biology-related disciplines (e.g., general biology, genetics, pathology/pathogens, medicine, and computational biology); (3) the *PLOS* journals have been open access since their inception, potentially increasing the likelihood of retrieving a complete set of publications; (4) the *PLOS* journals strongly favor the publication of primary research; and (5) the volume of publications in the core *PLOS* journals and all referenced works in those publications was computationally feasible.

For the purposes of this work, the operational definitions of “self-referencing” and “self-citation” are adopted from Szomszor et al. (2020). Specifically, self-referencing refers to the inclusion of one’s own prior publication in the reference list of a later publication. The per-article “self-referencing rate” can be estimated for each author individually and is defined as the percentage of references in each publication that were also (co-)authored by that individual. To examine self-referencing rates, an author list was retrieved for each publication in the core *PLOS* journals (“primary publication”) and cross-referenced against the author lists for each publication cited (“referenced publication”) in that article. The self-referencing rate assigned to each primary publication was then simplified as the maximum self-referencing rate among authors of that primary publication, which was typically achieved by the last/corresponding author of each publication (evaluated further in a later section). Estimated self-referencing rates are provided in Supplementary Table S1. Computational estimates of self-referencing rates are extremely consistent with

manual calculations performed for a randomly selected subset of publications ( $R^2 = 0.976$ ; Supplementary Figure S1, Table S2).

Self-referencing rate distributions are similar across the *PLOS* journals (Figure 1A, Supplementary Table S3), with median self-referencing rates between 8.2–11.4% and long, positively skewed tails. *PLOS Pathology* and *PLOS Medicine* exhibit the greatest positive skew and average self-referencing rate (13.4 and 13.5%, respectively). *PLOS Computational Biology* exhibits both the least positive skew and lowest mean self-referencing rate (10.1%). To determine whether self-referencing rates are changing over time, publications were stratified by journal and publication year. No clear trend in median self-referencing rate is apparent for any of the *PLOS* journals (Figure 1B, Supplementary Table S3), suggesting that self-referencing rates have been relatively stable over the lifetime of these journals.

While the *PLOS* journals provided a suitable case study for the reasons outlined above, it is possible that publisher practices or policies could influence self-referencing rates during the peer review or editorial stages. For each specialized *PLOS* journal, a non-*PLOS* journal with similar discipline, impact factor, and number of publications in the same 20-year timeframe was selected for comparison (Supplementary Figures S2A, B). Two additional cell biology journals were included to account for this large life science discipline that does not have a dedicated *PLOS* journal. Furthermore, the journals *Cell*, *Nature*, and *Science* were included to explore potential effects correlated with high impact factors. Self-referencing distributions for the non-*PLOS* journals are largely similar to those of *PLOS* journals (Figure 1C, Supplementary Table S3), though important minor differences are observed. Like its corresponding *PLOS* journal, *Bioinformatics* has the lowest median self-referencing rate and less positive skew compared to most other journals. However, in contrast to its *PLOS* counterpart, *BMC Medicine* has a low median self-referencing rate and less positive skew compared to many journals, which may in part be attributed to a higher number of total references compared to *PLOS Medicine* (Supplementary Figure S2C). *Nature* and *Science* have the highest median self-referencing rate and greatest positive skew of all journals, potentially due to publication of research in non-life-science disciplines, which can have substantially larger author lists and higher self-referencing rates (e.g., physics; Van Noorden and Singh Chawla, 2019). Median self-referencing rates for these journals span a broader range in the early 2000’s but exhibit some degree of convergence into a smaller range in recent years (Figure 1D, Supplementary Table S3), which may suggest that publisher practices and/or publishing conventions are evolving over time and becoming more standardized.

Importantly, cumulative comparison of self-referencing rates across all journals evaluated indicates that median self-referencing rates fall within a narrow range (Figure 1E) with similar distributions (Figure 1F) in biology-related disciplines, though there are minor interdisciplinary differences. These estimates are largely consistent with average self-referencing rates in neurology, neuroscience, and psychiatry (Rosenblatt et al., 2023), albeit with mean self-referencing rates slightly higher in neurology and psychiatry compared to all journals evaluated here.

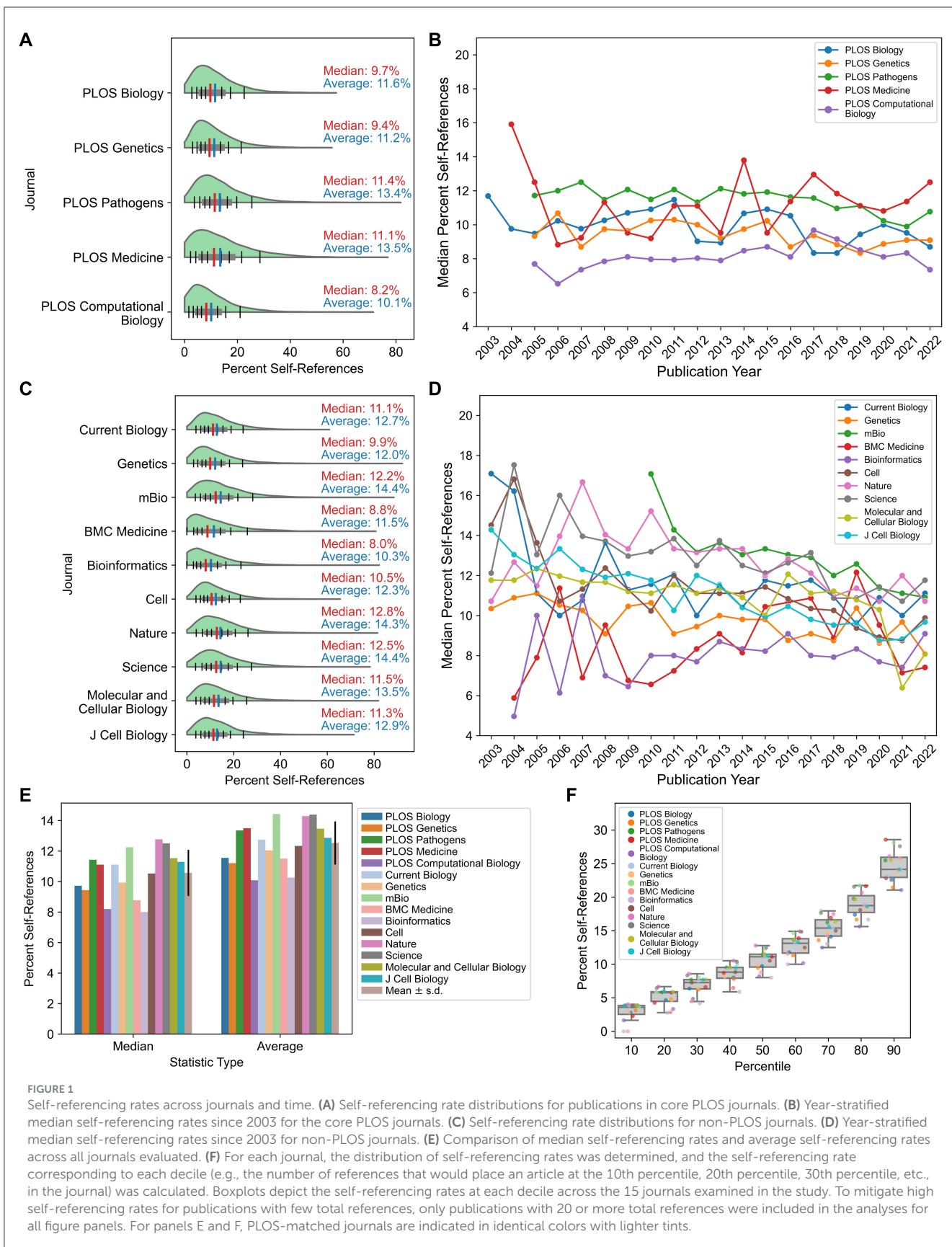


FIGURE 1

Self-referencing rates across journals and time. **(A)** Self-referencing rate distributions for publications in core PLOS journals. **(B)** Year-stratified median self-referencing rates since 2003 for the core PLOS journals. **(C)** Self-referencing rate distributions for non-PLOS journals. **(D)** Year-stratified median self-referencing rates since 2003 for non-PLOS journals. **(E)** Comparison of median self-referencing rates and average self-referencing rates across all journals evaluated. **(F)** For each journal, the distribution of self-referencing rates was determined, and the self-referencing rate corresponding to each decile (e.g., the number of references that would place an article at the 10th percentile, 20th percentile, 30th percentile, etc., in the journal) was calculated. Boxplots depict the self-referencing rates at each decile across the 15 journals examined in the study. To mitigate high self-referencing rates for publications with few total references, only publications with 20 or more total references were included in the analyses for all figure panels. For panels E and F, PLOS-matched journals are indicated in identical colors with lighter tints.

## Self-referencing rates are dependent on the total number of references

Publications in *PLOS Medicine*, *Nature*, and *Science* have some of the highest median and average self-referencing rates (Figure 1). However, these publications also tend to have notably fewer total references, on average, compared to publications in the other journals evaluated (Supplementary Figure S2C). Higher self-referencing rates might be expected for publications with fewer total references since, in those instances, relatively few citations to prior work could constitute a high percentage of total references.

To examine this directly, trends in the median (50<sup>th</sup> percentile) and interquartile range (from the 25<sup>th</sup> percentile to the 75<sup>th</sup> percentile) were studied as a function of the total number of references. *A priori*, quartile values were expected to decrease for publications with a larger number of total references since high percent self-reference values would be more difficult to achieve as total references increase. Indeed, for all journals, quartile values typically decrease steadily as the total number of references increases (Figure 2, Supplementary Table S4), indicating that the distributions of self-referencing rates shift toward lower percent self-reference values. Total references have a relatively strong effect on self-referencing rates, with most journals exhibiting a 5 to 10 percentage-point decrease in median self-referencing rate as total references increase. However, not all journals exhibit identical behavior. For example, *PLOS Biology* starts with a relatively low median self-referencing rate (~11.9%) but is less sensitive to total references than most journals, with the lowest self-referencing rate reaching ~6.5% and corresponding a range of only ~5.4% (the second-lowest range of all journals). *Molecular and Cellular Biology* (*MCB*) starts with the highest median self-referencing rate (~17.1%) for publications with a low number of total references but exhibits a steep decline, eventually resulting in one of the lowest self-referencing rates (~5.5%; range ~11.6%) for publications with a high number of total references (Figure 2). Similarly, *PLOS Medicine* exhibits a steep decline in self-referencing rates as total references increase; however, a large number of *PLOS Medicine* publications have relatively few references (Supplementary Figure S2C), suggesting that its high overall median self-referencing rate is driven predominantly by the practice of citing fewer publications. In contrast, publications in *Nature* and *Science* simultaneously tend to have few references (Supplementary Figure S2C), high initial self-referencing rates at a low number of total references (Figure 2), and early plateaus (and sometimes even an increase) in quartile values as a function of total references; all three factors likely contribute to their unusually high median self-referencing rates. Publications in *Cell* also exhibit relatively high initial self-referencing rates at a low number of total references (Figure 2), but the majority of publications have a larger number of total references (Supplementary Figure S2C). Consequently, the majority of *Cell* publications are found within a region of the curve with a lower self-referencing rate, driving down the overall mean and median self-referencing rates relative to *Nature* and *Science* (Figures 1C, E).

In summary, these observations suggest that self-referencing rates across many journals are strongly dependent on the total number of references. The degree and shape of the dependence

differ by journal and are likely influenced by journal policies related to total allowable references or total article length restrictions, among other factors (see Discussion). Therefore, variation in total references results from both individual author habits and systemic factors that affect total references.

## Total number of authors has only a minor effect on maximum self-referencing rates

Differences in the total number of authors on each publication are observed across journals (Supplementary Figure S2D). Previous bibliometric studies indicate a correlation between the total number of authors on a publication and the self-referencing/self-citation rate in some fields (Snyder and Bonzi, 1998; Aksnes, 2003; Leimu and Koricheva, 2005). However, these studies use a different metric: the *collective* self-referencing rate, which is defined as the percentage of references for which any author on the primary publication was also an author of the referenced publication. This is in contrast to the *individual* self-referencing rate (e.g., Fowler and Aksnes, 2007; Szomszor et al., 2020), which is calculated independently for each author and is here used to determine the maximum individual self-referencing rate. Compared to the maximum individual self-referencing rate, the collective self-referencing rate is likely more strongly correlated with the total number of authors. Nevertheless, the total number of authors could still affect the maximum individual self-referencing rate by increasing the probability that a higher maximum self-referencing rate is achieved, since each additional author may have unique self-referencing habits and a unique pool of prior publications that could be referenced.

Self-referencing rates exhibit a weaker dependence on the total number of authors and more variability across journals (Figure 3, Supplementary Table S5) when compared to the dependence on the total number of references. Nearly all journals exhibit at least a slight increase in typical self-referencing rates as the number of authors increases. However, for many journals such as *PLOS Biology*, *Current Biology*, *mBio*, *PLOS Computational Biology*, *Cell*, *MCB*, and *J Cell Biology* this increase does not exceed more than ~3 total percentage points gained. For the journals *PLOS Genetics*, *Genetics*, *PLOS Pathogens*, *Bioinformatics*, and *Nature*, moderate increases of ~3–5 percentage points are observed. Notably, both medical journals (*PLOS Medicine* and *BMC Medicine*) exhibit large increases of ~7–10 percentage points as the total number of authors increases, potentially indicating a domain-specific practice. Publications in *Science* also exhibit a large increase in self-referencing rate, but only starting at publications with ~30 or more authors.

One possible explanation for the weak relationship between percent self-references and the total number of authors is that authors are unlikely to contribute equally to the self-referencing rate. For example, “senior authors” (i.e., principal investigators or lab directors) are presumably the most likely to achieve the highest self-referencing rate since they typically have the longest history of closely related publications, but senior authors usually constitute a small fraction of the author list. In contrast, “junior authors” (e.g.,

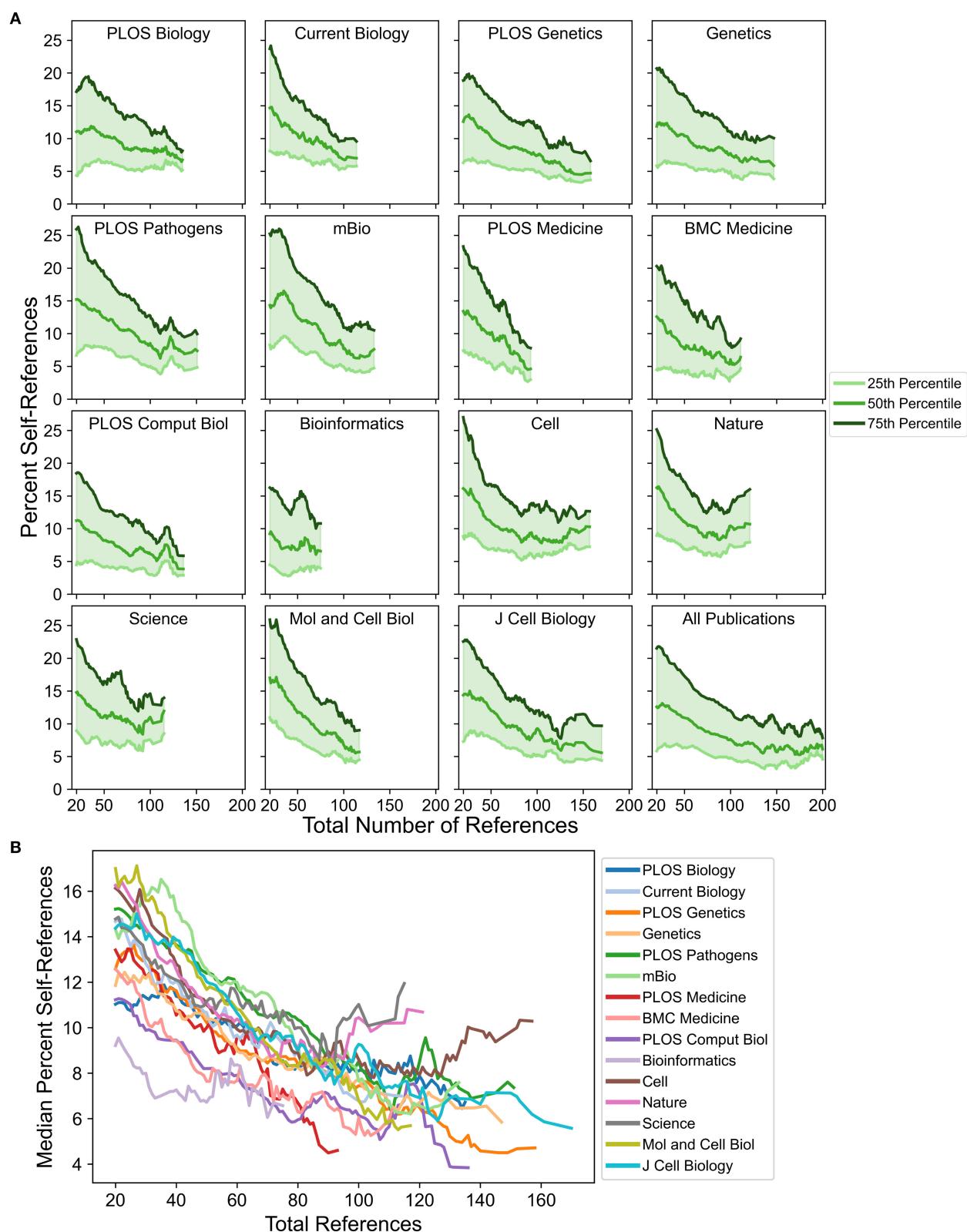


FIGURE 2

Changes in self-referencing rate distributions as a function of total references. **(A)** Distributions of percent self-references (as indicated by values corresponding to the 25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup> percentiles) for each journal shift as a function of total references. **(B)** Direct comparison of the median percent self-reference value across journals as a function of total references. For all analyses, curves were smoothed by calculating a sliding mean with a window size of 11 (see Materials and Methods).

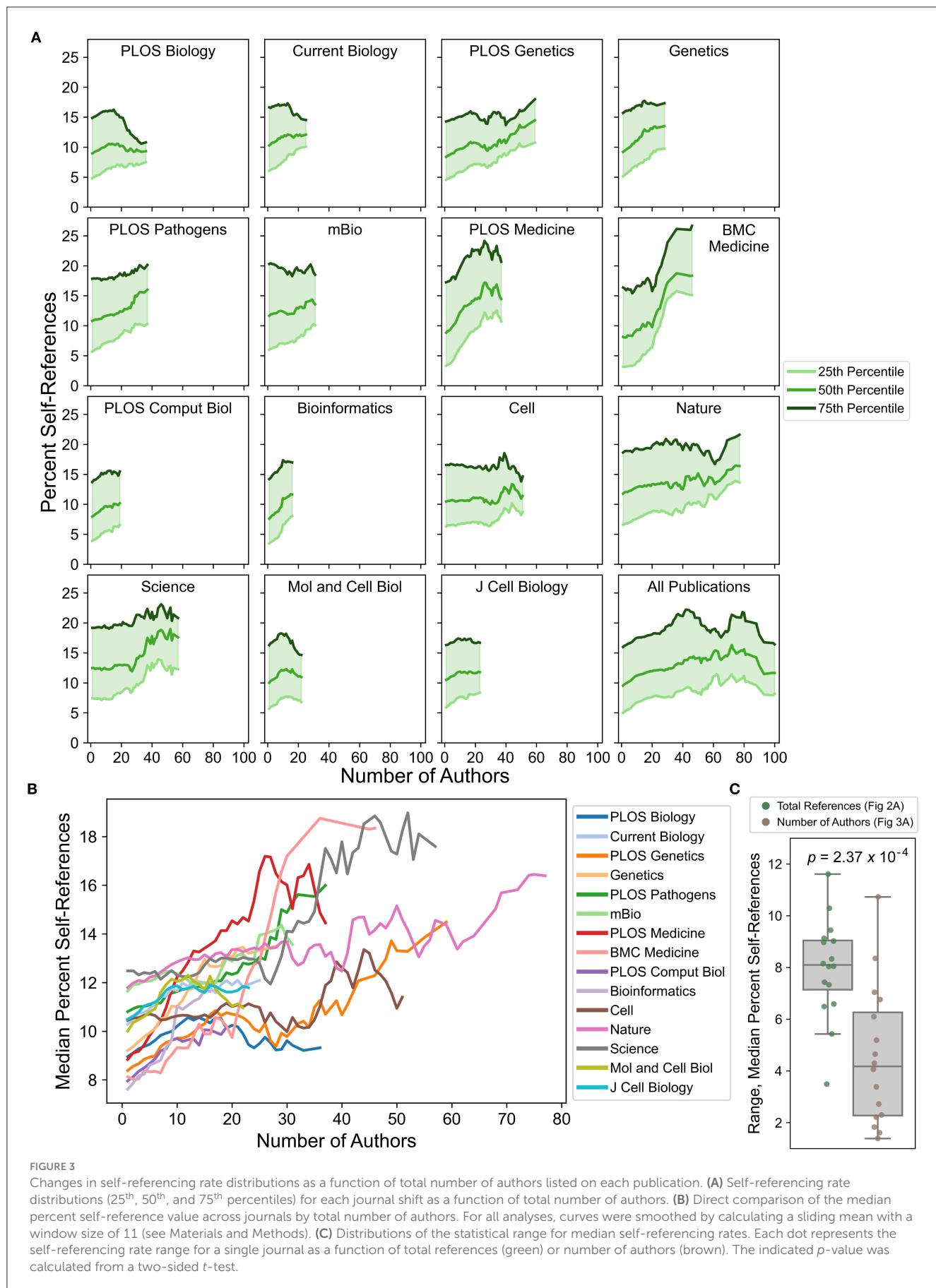


FIGURE 3

Changes in self-referencing rate distributions as a function of total number of authors listed on each publication. **(A)** Self-referencing rate distributions (25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup> percentiles) for each journal shift as a function of total number of authors. **(B)** Direct comparison of the median percent self-reference value across journals by total number of authors. For all analyses, curves were smoothed by calculating a sliding mean with a window size of 11 (see Materials and Methods). **(C)** Distributions of the statistical range for median self-referencing rates. Each dot represents the self-referencing rate range for a single journal as a function of total references (green) or number of authors (brown). The indicated  $p$ -value was calculated from a two-sided  $t$ -test.

undergraduates, graduate students, post-docs, and research faculty) often have fewer related publications compared to the senior author but constitute the majority of the author list. From this perspective, self-referencing rates would be most strongly influenced by the addition of senior authors, as occurs for collaborative publications. Therefore, the total number of authors is likely an imperfect proxy for the number of senior authors included in the collaboration.

To examine this more directly, publications were binned according to the number of authors. In a conventional research group architecture, at least one junior scientist participates in the research project for every senior scientist involved. Therefore, publications with three or fewer authors were generally assumed to originate from a single laboratory and have a single senior author. Indeed, the maximum self-referencing rate is achieved by the last author for ~85% of publications with three or fewer total authors ([Figure 4A](#)), suggesting both that the large majority of publications in this group likely originate from a single laboratory and that the senior author is nearly always the highest self-referencing author. In the remaining 15% of publications, the last author was sometimes not the corresponding author, suggesting that the senior author is still the maximum self-referencing author in a portion of these instances. Publications with four or more total authors are considered potentially collaborative, with the probability of collaboration expected to increase as the total number of authors increases. For all total-author bins examined, the last author achieves the maximum self-referencing rate for most of the publications. The share of publications progressively decreases for all total-author bins as a function of author position when moving inward from the last author: this confirms that the most-senior authors are indeed the most likely to achieve the highest self-referencing rate, and that less-senior authors are progressively less likely to achieve the maximum self-referencing rate.

Some senior authors frequently collaborate with the same senior co-author(s), which could lead to situations where the sets of self-references are highly overlapping. In these instances, even a single additional self-reference could shift the maximum self-referencing rate away from the last author to another senior author. For example, although the last author may not achieve the maximum self-referencing rate among collaborative publications (e.g., 30–39 authors), they may still be responsible for the majority of the self-references, and the share of unique additional self-references associated with non-last authors may be minimal.

For each publication, percent self-references were sequentially calculated starting with the last author. For each subsequent author (moving inward from the last author), a self-reference was only included in the calculation if it had not already been included as a self-reference for the preceding authors, effectively exploring the unique contribution of each author to a collective self-referencing rate. The last author tends to contribute the overwhelming majority of self-references for all total-author bins ([Figure 4B](#), [Supplementary Table S6](#)). Surprisingly, this is largely independent of the total number of authors, with only a marginal rightward shift in the percent self-reference distribution for publications with 40–50 authors. This suggests that self-referencing habits of the most-senior author tend to be insensitive to the number of additional co-authors (including senior authors), and may reflect a certain baseline self-referencing that authors deem appropriate for

describing prior work. However, for subsequent authors (i.e., 2nd-to-last, 3rd-to-last, etc.), the unique contribution to percentage of self-references is dependent on both the total number of authors and their proximity to the end of the author list.

Collectively, these results corroborate prior observations that self-referencing is dependent on the total number of authors. However, the majority of unique self-references for most publications are attributable to the last author in a manner that is insensitive to the total number of authors (with the possible exception of certain disciplines such as medicine). Minor additional self-references are subsequently contributed by senior co-authors, and this contribution scales with the total number of authors and proximity to the end of the author list.

## Self-referencing rates are related to author rank and total number of publications

One possible explanation for high self-referencing rates is an unusually significant contribution to a particular field. While “contribution significance” is at least partially subjective and difficult to measure, two important components of contribution significance are breadth (e.g., the total number of publications associated with an individual) and depth (e.g., the degree of impact, on average, associated with an individual) ([Wildgaard et al., 2014](#)). A recently published database of highly ranked researchers ([Ioannidis et al., 2019, 2020](#)) contained both the total number of publications by each listed researcher and a composite score (based on six distinct citation measures; [Ioannidis et al., 2016](#)) used to rank each researcher, which are suitable estimates of breadth and depth.

For all authors present in the ranked-authors database, the relationship between self-referencing rates for the maximum self-referencing author and the total number of publications for that author was explored. Very high self-referencing rates (>30%) typically occur at the low end of the total-publications spectrum, with a dense cluster between ~50–250 total publications ([Supplementary Figures S3A, B](#)). However, a notable portion of the scatterplot extends toward higher total publications ([Supplementary Figure S3A](#)). To better understand how the distribution of self-referencing rates changes as a function of total publications by an author, quartile values were calculated for self-referencing rate distributions as a function of total publications. Quartile values for self-referencing rate distributions rise sharply until ~200 total publications ([Figure 5A](#)). Beyond 200 publications, quartile values continue to rise on average, but at a much slower rate. This is consistent with previous observations ([Mishra et al., 2018](#)) that higher self-referencing rates tend to occur for authors with more total publications, but suggests a more nuanced relationship between total publications and self-referencing rates: beyond a certain point, the effect of each additional publication on self-referencing rate diminishes.

Hypothetically, authors could attempt to reference most or all of their prior work in their own publications, which would likely result in both a high self-referencing rate and a high percentage of the author’s total existing publications referenced. The self-referencing rate quartile values exhibit an initial increase

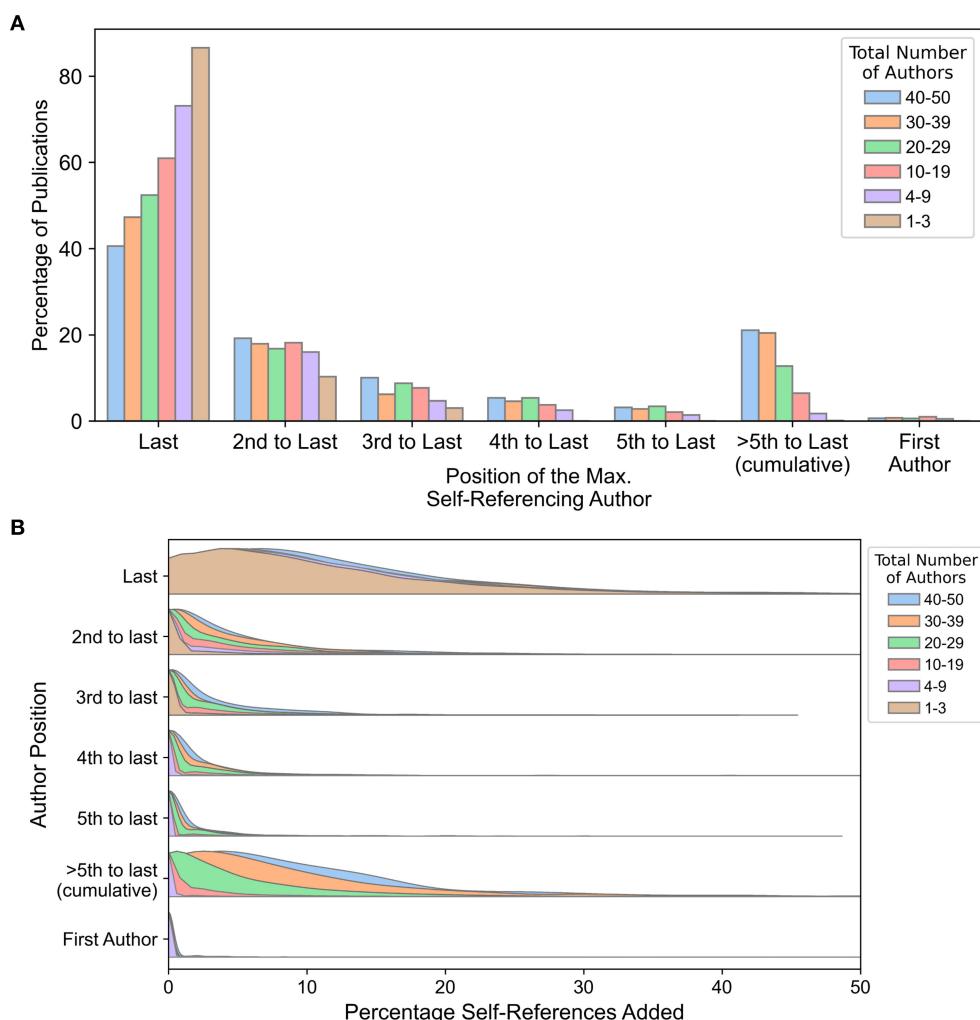


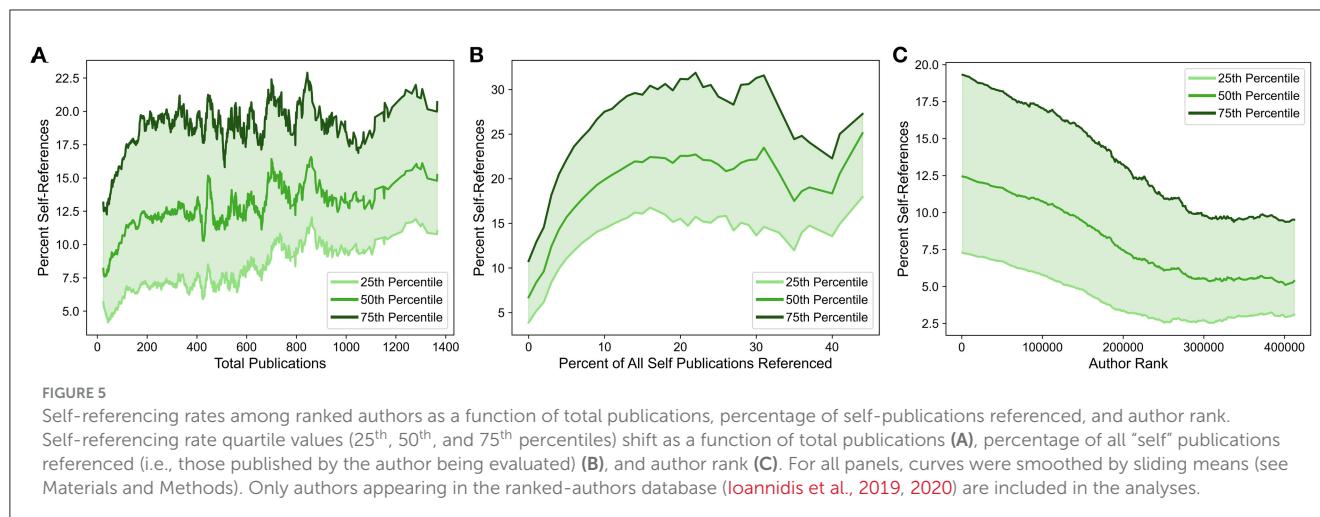
FIGURE 4

Self-referencing statistics by author list position. (A) The share of publications (as a percentage) for which the maximum self-referencing rate was assigned to each author for the final five authors and the first author, stratified into bins by total number of authors. (B) Distributions of the percentage of self-references added to the collective self-referencing rate by each author position. The percentage of self-references added is calculated sequentially by determining the individual self-referencing rate for each author while excluding self-referenced publications that were already accounted for in a preceding author, starting with the last author and moving inward (i.e., last author, 2nd-to-last author, 3rd-to-last author, etc.). For both panels, the author on single-author publications is classified as the last author. In cases of ties between authors, the maximum self-referencing rate was attributed to the most-senior author (i.e., the author nearest to the end of the author list), as this reflects the number of unique references contributed by each additional author.

with an increasing percentage of total publications referenced but plateaus at ~15% of total publications referenced (Figure 5B), at which point the median self-referencing rate stabilizes at ~22–23%. The magnitude of the increase in self-referencing rate is larger than that observed as a function of total publications. However, instances with both a high self-referencing rate and a high percentage of total publications referenced represent a small minority of publications (Supplementary Figures S3C, D), suggesting that most authors are not attempting to cite a large proportion of their work at once. Since a simultaneously high self-referencing rate and high percentage of publications referenced would be difficult to achieve for authors with a large number of total publications, this suggests that high self-referencing rates may also be achieved by citing a larger

fraction of prior work, particularly for authors with fewer total publications.

The ranked-authors database includes a ranking that is independent of self-citation. This aids in mitigating downstream effects of authors that boost their apparent impact by heavily referencing their own work (though non-self-references are not entirely independent of self-references; Van Raan, 1998; Fowler and Aksnes, 2007). A steady shift from higher to lower self-referencing rates is observed when traversing from high-ranking authors to lower-ranking authors (Figure 5C). Very high self-referencing rates are also preferentially clustered toward authors with the highest rank (Supplementary Figures S3E, F). It is worth noting that high-ranking authors also tend to have a larger number of total publications and that a higher number of total publications is



likely to garner more total citations (Supplementary Figures S4A–C). However, the decrease in self-referencing rate as author rank decreases is still observed when authors are stratified based on the total number of publications, although the magnitude of the effect progressively decreases for author groups with larger total publications (Supplementary Figures S4D–F). This suggests that the correlation between self-referencing rate and author rank is at least partially independent of the total number of publications. Therefore, high self-referencing rates often correspond to works from unusually influential authors, as judged both by total publications and non-self, citation-based author rank.

Although clear trends are observed within the ranked-authors database, these authors still constitute a biased subset of researchers, since all authors included in the database are considered highly influential. Therefore, self-referencing rates for authors included in the ranked-authors database were compared to self-referencing rates for authors excluded from the ranked-authors database. The distribution of self-referencing rates is significantly shifted toward higher self-referencing percentages for ranked authors compared to non-ranked authors (Figure 6A). Additionally, the percentage of publications for which the maximum self-referencing author appears in the ranked-authors database steadily increases among publications with increasingly high self-referencing rates (Figure 6B). For example, only ~10% of publications with 0% self-references correspond to publications where the maximum self-referencing author appears in the ranked author database. However, this number jumps to nearly 60% among publications with ~20% self-references, and further increases to ~70–80% for publications with ≥40% self-references. Thus, highly influential authors, regardless of their rank in the database, exhibit higher self-referencing rates than non-ranked authors.

## Discussion

### Self-referencing in biology-related disciplines

Self-referencing in science often serves a legitimate purpose: since most scientific endeavors build upon previous work from the

same research group, appropriate self-referencing can help readers contextualize a publication by highlighting prior foundational work (Gläzel et al., 2006; Zhao et al., 2018; Szomszor et al., 2020). From this perspective, strict, one-size-fits-all rules limiting self-referencing could lead to referencing practices that misrepresent existing literature. Based on a limited poll, most researchers are against journals “policing” self-referencing (Van Noorden and Singh Chawla, 2019). On the other hand, the use of citations as a simple summary statistic influencing professional outcomes (e.g., grant funding, hiring/promotion decisions, recognition, etc.) incentivizes researchers to disproportionately cite their own work, which also misrepresents existing literature (Gläzel et al., 2006; Szomszor et al., 2020). Many of the cited studies acknowledge the potential existence of “excessive” self-referencing, though opinions on precise definitions and criteria for “excessive” vary. Therefore, self-referencing practices require a delicate balance that must be navigated by research authors and journals alike.

The statistical analyses presented here of self-referencing rates over the past 20 years reveal modern, unwritten conventions among researchers in biology-related fields, and how these conventions continue to evolve. In principle, self-referencing rates could be automatically detected by journals upon manuscript submission using computational methods similar to the approach utilized here, and estimated percentile values for self-referencing rates could be provided to peer reviewers to inform their assessment. In that sense, journals would not be “policing” self-referencing: rather, they would be providing contextual information to the peer review experts primarily responsible for evaluating a manuscript—a task that they are often prompted to do anyway, without this contextual information. Even in the absence of an automated system, peer reviewers can estimate the self-referencing rate for each manuscript and compare it to these statistical benchmarks. Importantly, self-referencing rates and their corresponding percentile benchmarks could be normalized by the total number of references: this would prevent authors from artificially lowering their percentage of self-references by including extraneous, non-self-references to inflate the total number of references. Journals, reviewers, and authors could also consider the total number of publications by the author(s), which could justifiably increase self-referencing rates

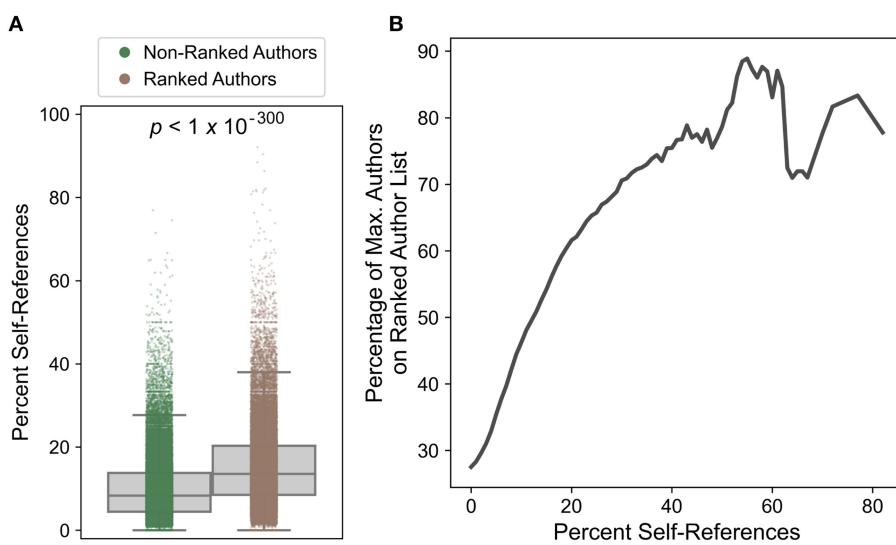


FIGURE 6

Comparison of self-referencing rates for ranked and non-ranked authors. **(A)** Self-referencing rate distributions for ranked and non-ranked authors. The indicated  $p$ -value was calculated from a two-sided t-test and was below the precision limit of the software. In cases of ties, the most-senior author involved in the tie was evaluated for inclusion/exclusion in the ranked-authors database. **(B)** Percentage of the maximum self-referencing authors that appear on the ranked-author list as a function of minimum percent self-references. Publications were parsed into bins (at increments of 1%) based on their corresponding self-referencing rates. At each self-referencing rate percentage point (x-axis), all publications with a self-referencing rate greater than or equal to that percentage point were included in the evaluated subset. For each publication subset, the percentage of the maximum self-referencing authors included in the ranked-authors database (y-axis) is shown (smoothed by sliding means; see Materials and Methods).

if the authors have a relatively large number of citable works; but even this should be considered thoughtfully as it can be gamed by authors (Ioannidis, 2015).

Although these statistical benchmarks are indicative of modern self-referencing conventions, they do not preclude the possibility of publications with exceptionally high self-referencing rates. Such situations could include researchers working on emerging or underexplored topics, researchers with exceptional influence in a particular field (e.g., “founders” of a field of study), or publications with relatively few references, which may be short, narrowly focused publications. For example, two publications were identified with an estimated self-referencing rate  $>90\%$ : however, upon manual inspection, both publications appear to be invited papers accompanying an award to an individual researcher and were intended to be an autobiographical account of their discoveries in a particular field. Even with these exceptions, self-referencing rates rarely exceed  $\sim 20\%$  (surpassed only by the 90<sup>th</sup> percentile across journals in Figure 1F), indicating that situationally high self-referencing rates are uncommon.

To the best of my knowledge, none of the journals evaluated provide concrete guidelines or restrictions on self-referencing which, again, appears to actually be preferred among researchers (Van Noorden and Singh Chawla, 2019). A lack of explicit self-referencing policies is also advantageous for the present study, since *a priori* restrictions to self-referencing could affect the observed distributions. However, some journals have policies that indirectly affect self-referencing rates. For example, *Nature* and *Science* recommend 50 or fewer total references for their main article types which, in light of the relationship between total references and self-referencing rates (Figure 2), may also partially explain the higher

mean and median self-referencing rates and more positive skew in the distribution of self-referencing rates. *Bioinformatics* does not explicitly limit the total number of references but imposes length restrictions that include the references, which indirectly affects self-referencing rates. *Cell*, *J Cell Biology*, *MCB*, *mBio* and *Current Biology* impose article length restrictions but exclude the references from these restrictions. The remaining journals (*PLOS* journals, *Genetics*, and *BMC Medicine*) do not restrict article length or references. The only journal that explicitly addressed self-referencing was *BMC Medicine*, whose editorial policies included a statement that excessive self-referencing and citation manipulation are strongly discouraged and grounds for rejection, but they do not define what constitutes “excessive” self-referencing. This lone example amounts to only  $\sim 6.7\%$  of the journals evaluated here having an explicit self-referencing policy, which is lower than estimates in other fields such as critical-care medicine (Sanfilippo et al., 2021a) and anesthesiology (Sanfilippo et al., 2021b), albeit with a relatively small sample size in the present study.

Some journals explicitly restrict self-referencing to a certain percentage, although this does not currently seem to be a common practice. For example, in their author guidelines *Current Opinion in Cell Biology* expresses a hard upper limit of 20% self-references, while *Langmuir* (a chemistry journal) states that a 25% self-reference upper limit is imposed but that exceptions can be made by consulting with the editors (both sets of guidelines accessed on 1/31/2023). Neither journal specifies whether these upper limits apply to *individual* or *collective* self-referencing rates. Assuming these limits refer to maximum individual self-referencing rates, they would roughly correspond to the 85<sup>th</sup>-95<sup>th</sup> percentile for the biological journals evaluated in this study. If the limits correspond

instead to collective self-referencing rates they are potentially more restrictive, particularly for studies with many authors or multiple senior authors. Regardless, such limits—based on self-reference rate cutoffs rather than percentiles—have the potential to change self-referencing distributions over time, eventually resulting in new “normal” ranges. Furthermore, these limits do not account for the relationship between self-referencing percentages and the total number of references in the manuscript (Figure 2), the total number of authors on the publication (Figures 3, 4), or the total number of citable publications associated with the self-referencing author (Figure 5). This effectively penalizes publications with shorter reference lists, more total authors, or a large number of citable prior publications, respectively. These factors, in conjunction with statistical benchmarks, enable a more informed and contextualized assessment of self-referencing rates by authors, peer reviewers, editors, and journals.

Finally, I wish to make clear that I am not advocating for or against any particular policy regarding self-referencing; these data are provided merely to aid in the development of informed decisions by relevant experts, which will depend on each situation. Discussion of individual researchers was also intentionally omitted to avoid shaming or “criminalizing” (Van Noorden and Singh Chawla, 2019; Kacem et al., 2020) self-referencing, and because a focus on specific individuals is here considered less important and substantive than the broader statistics of self-referencing.

## Limitations of the study

There are a number of limitations worth discussing. First, although failed PubMed queries were repeated multiple times (see Materials and Methods), a small subset of PubMed queries for primary publications and referenced publications were still not successful. Failed queries most often occurred for non-research article types (e.g., editorials, author corrections, interviews, blog-like articles, news articles, conference proceedings, citable databases, etc.) or for journals not indexed on PubMed. Nevertheless, self-referencing rates (as well as the total number of publications by journal in Supplementary Figure S1A) should only be viewed as estimates rather than exact figures. Despite this limitation, estimates of self-referencing rates from automated PubMed queries are quite consistent with values derived from manual calculation of self-referencing rates (Supplementary Figure S1). Second, although the main focus has been on journals that predominantly publish experimental research, the full dataset of publications does include reviews, author corrections, and other article types. These alternative article types may exhibit distinct self-referencing rate distributions but are not distinguishable using the employed methodology. However, the effects of these alternative article types are mitigated by two main factors: (1) the criterium that publications required a minimum of 20 total references for inclusion in all analyses performed, which disproportionately eliminated publications such as author corrections and “letter to the editor” publications since they typically had few total references; and (2) failed PubMed queries, which disproportionately corresponded to non-research

article/reference types. Third, although the individual PubMed query method improves author disambiguation compared to the use of databases with author name initials, author disambiguation is an ongoing challenge among bibliometric studies and may affect both self-referencing rates (potentially overestimating rates for authors with unusually common names) and matching of authors to the ranked-authors database. Authors could not be disambiguated by author affiliation due to inconsistent listing of author affiliations across PubMed publications. Fourth, comparison of self-referencing rates to total publications and fraction of publications self-referenced is imperfect, since publication dates were not accessible in the corresponding ranked-authors database. Ideally, for each publication with a calculated self-referencing rate, only the total publications prior to that publication would be included. However, the total number of career publications is considered a reasonable proxy for this purpose. Fifth, authors who frequently and/or selectively publish in the journals evaluated in this study are more strongly weighted than others. However, given the scope of this study (>94 k publications evaluated), any single author is unlikely to have a large effect, particularly since most authors publish their work across a wide variety of journals. Finally, this study is clearly limited in scope to a small subset of journals. The data indicate that, while key indicators of self-referencing rates tend to fall within a relatively narrow range, there are also minor differences between journals which may reflect differences in publishing practices or content matter. These differences do not preclude the use of self-referencing guidelines, but rather indicate that thoughtful consideration (including aforementioned situational exceptions) should be exercised when authors, peer reviewers, and journals apply these statistical guidelines to individual cases. On the other hand, these data also reveal an advantage of studying self-referencing at the journal level: studies that combine publications across journals according to a field of study are inevitably combining sets of publications that were subject to distinct editorial regulations (e.g., article length, reference number, etc.), which have a demonstrable effect on self-referencing rates. Since publishers vary in the number of publications produced per year, publishers with high publishing rates will have the greatest weight in calculated self-referencing rates for field-grouped publication sets.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/[Supplementary material](#). All code required to reproduce data in this publication are available at <https://github.com/RossLabCSU/SelfReferencingRates2023>.

## Author contributions

SC completed all aspects of this study including conceptualization, data acquisition, data visualization, statistical analyses, manuscript composition, and manuscript revision.

## Acknowledgments

I thank Dr. Eric D. Ross and Dr. Jennifer G. DeLuca for fruitful discussion, and Lindsey Bush for editing suggestions.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frma.2023.1215401/full#supplementary-material>

### SUPPLEMENTARY FIGURE S1

Comparison of manual vs. automated calculation of self-referencing rates. A total of 100 publications were randomly selected from all publications evaluated. Self-referencing rates were calculated manually by examining the reference list in each publication, then compared to the self-referencing rate calculated via the automated PubMed query pipeline.

### SUPPLEMENTARY FIGURE S2

Comparison of number of publications, impact factor, number of references, and number of authors across journals. (A) Total number of publications evaluated in this study for each journal. (B) Impact factor of each journal. (C) Distributions of total number of references within the primary publications evaluated for each journal. (D) Distributions of the number of authors for the primary publications evaluated for each journal.

### SUPPLEMENTARY FIGURE S3

Self-referencing rates as a function of total publications, fraction of total publications referenced, and author rank. (A) Scatterplot depicting the percentage of self-references vs. total publications. (B) Scatterplot focusing on the percentage of self-references vs. total publications only for publications with very high self-referencing rates ( $\geq 30\%$ ) and authors with fewer than 2,000 total publications. (C, D) Similar to (A, B), but with the percentage of total publications self-referenced serving as the x-axis variable. (E, F) Similar to (A, B), but with author rank (Ioannidis et al., 2019, 2020) serving as the x-axis variable. For all panels, only the subset of authors appearing in the ranked-author database were included in analyses. Author rank in (E, F) was limited to 1 million and 400 k, respectively, for visual clarity only. Point density was estimated using a kernel density estimate and colored accordingly, with dark red indicating high-density regions and dark blue indicating low-density regions.

### SUPPLEMENTARY FIGURE S4

Author rank vs. total publications. (A) Plot depicting shifts in author-rank quartile values as a function of total publications. (B) Scatterplot depicting author rank as a function of total publications. (C) Scatterplot of author rank vs. total publications, focusing on authors with  $\geq 500$  total publications. Authors with an unusually large number of total publications are concentrated among top-ranking authors. For (B, C), point density was estimated using a kernel density estimate and colored accordingly, with dark red indicating high-density regions and dark blue indicating low-density regions. (D–F) Shifts in percent self-reference distributions as a function of author rank, grouped by total publications as indicated in the title of each panel.

### SUPPLEMENTARY TABLE S1

Estimated self-referencing rates for each publication. Publications are listed by journal and PubMed ID. Authors are listed in the order that they appear in the original article, with each author separated by a semicolon. For each author, the number of self-references and percentage of self-references are estimated separately and appear in the same format as the listed authors (i.e., separated by semicolons, in the same order as the authors). Note that the number of self-references and the total number of references only reflect references that were successfully retrieved from PubMed queries: references not indexed on PubMed were excluded from analyses.

### SUPPLEMENTARY TABLE S2

Comparison of automated vs. manually calculated self-referencing estimates. A total of 100 publications were randomly selected from all publications analyzed in this study. Columns are identical to those in **Supplementary Table S1**, with the addition of three new columns containing manually calculated values for the number of self-references, the total number of references in each publication, and the percentage of self-references.

### SUPPLEMENTARY TABLE S3

Self-referencing rate summary by journal. For each journal, the overall self-referencing rate at each indicated percentile is shown (columns 4–15). Additionally, the overall median self-referencing rate is indicated for each publication year from 2003–2022 (inclusive; columns 16–35).

### SUPPLEMENTARY TABLE S4

Self-referencing rate values at percentiles as a function of total number of references. For each journal, publications were binned based on the total number of references in each publication. Self-referencing rates were then calculated at each percentile (from 1–100) separately for each total-references bin. For columns 3–234, the column header indicates the total number of references, with values corresponding to the self-referencing rate estimated for the percentile indicated in column 2. Key percentile markers (25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup> percentiles) are those shown in **Figure 2**. The 100<sup>th</sup> percentile represents the maximum self-referencing rate. Self-referencing rates within each total-references column were smoothed using a sliding mean (see Materials and Methods). "nd" represents "no data".

### SUPPLEMENTARY TABLE S5

Self-referencing rate values at percentiles as a function of number of authors. For each journal, publications were binned based on the number of authors on each publication. Self-referencing rates were then calculated at each percentile (from 1–100) separately for each number-of-authors bin. For columns 3–103, the column header indicates the number of authors, with values corresponding to the self-referencing rate estimated for the percentile indicated in column 2. Key percentile markers (25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup> percentiles) are those shown in **Figure 3**. The 100<sup>th</sup> percentile represents the maximum self-referencing rate. Self-referencing rates within each number-of-authors column were smoothed using a sliding mean (see Materials and Methods). "nd" represents "no data".

### SUPPLEMENTARY TABLE S6

Progressive contributions to self-referencing rates. For each publication, progressive self-referencing rates—defined as the percentage of unique self-references contributed by each author, starting with the last author—were assigned to each author.

## References

- Abramo, G., and D'Angelo, C. A. (2011). Evaluating research: from informed peer review to bibliometrics. *Scientometrics* 87, 499–514. doi: 10.1007/s11192-011-0352-7
- Aksnes, D. W. (2003). A macro study of self-citation. *Scientometrics* 56, 235–246. doi: 10.1023/A:1021919228368
- Aksnes, D. W., Langfeldt, L., and Wouters, P. (2019). Citations, citation indicators, and research quality: an overview of basic concepts and theories. *SAGE Open* 9, 1. doi: 10.1177/2158244019829575
- Bartneck, C., and Kokkelmans, S. (2011). Detecting h-index manipulation through self-citation analysis. *Scientometrics* 87, 85–98. doi: 10.1007/s11192-010-0306-5
- Bonzi, S., and Snyder, H. W. (1991). Motivations for citation: A comparison of self citation and citation to others. *Scientometrics* 21, 245–254. doi: 10.1007/BF02017571
- Bornmann, L., and Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *J. Doc.* 64, 45–80. doi: 10.1108/00220410810844150
- Brooks, T. A. (1985). Private acts and public objects: an investigation of citer motivations. *J. Am. Soc. Inf. Sci.* 36, 223–229. doi: 10.1002/asi.4630360402
- Brooks, T. A. (1986). Evidence of complex citer motivations. *J. Am. Soc. Inf. Sci.* 37, 34–36. doi: 10.1002/asi.4630370106
- Cano, V. (1989). Citation behavior: classification, utility, and location. *J. Am. Soc. Inf. Sci.* 40, 284–290. doi: 10.1002/(SICI)1097-4571(198907)40:4<284::AID-ASII10>3.0.CO;2-Z
- Chaplain, M., Kirschner, D., and Iwasa, Y. J. T. B. (2020). Editorial malpractice: a case report. *J. Theor. Biol.* 488, 110171. doi: 10.1016/j.jtbi.2020.110171
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163
- Copello, S. (2019). On the skewness of journal self-citations and publisher self-citations: Cues for discussion from a case study. *Learn. Publ.* 32, 249–258. doi: 10.1002/leap.1235
- Fong, E. A., and Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLOS ONE* 12, e0187394. doi: 10.1371/journal.pone.0187394
- Fowler, J. H., and Aksnes, D. W. (2007). Does self-citation pay? *Scientometrics* 72, 427–437. doi: 10.1007/s11192-007-1777-2
- Garfield, E. (1965). "Can citation indexing be automated?", in *Statistical Assoc. Methods for Mechanized Documentation, Symposium Proceedings, Washington 1964*, eds. M. E. Stevens, V. E., Giuliano, L. B., Heilprin (National Bureau of Standards (NBS) Miscellaneous Publication) 189–192.
- Glänzel, W., Debackere, K., Thijss, B., and Schubert, A. (2006). A concise review on the role of author self-citations in information science, bibliometrics and science policy. *Scientometrics* 67, 263–277. doi: 10.1007/s11192-006-0098-9
- Heneberg, P. (2014). Parallel worlds of citable documents and others: Inflated commissioned opinion articles enhance scientometric indicators. *J. Assoc. Inf. Sci. Technol.* 65, 635–643. doi: 10.1002/asi.22997
- Heneberg, P. (2016). From excessive journal self-cites to citation stacking: analysis of journal self-citation kinetics in search for journals, which boost their scientometric indicators. *PLOS ONE* 11, e0153730. doi: 10.1371/journal.pone.0153730
- Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., and Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature* 520, 429–431. doi: 10.1038/520429a
- Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English Specif. Purp.* 20, 207–226. doi: 10.1016/S0889-4906(00)00012-0
- Hyland, K. (2003). Self-citation and self-reference: Credibility and promotion in academic publication. *J. Am. Soc. Inf. Sci. Technol.* 54, 251–259. doi: 10.1002/asi.10204
- Ioannidis, J. P. A. (2015). A generalized view of self-citation: Direct, co-author, collaborative, and coercive induced self-citation. *J. Psychosom. Res.* 78, 7–11. doi: 10.1016/j.jpsychores.2014.11.008
- Ioannidis, J. P. A., Baas, J., Klavans, R., and Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLOS Biol.* 17, e3000384. doi: 10.1371/journal.pbio.3000384
- Ioannidis, J. P. A., Boyack, K. W., and Baas, J. (2020). Updated science-wide author databases of standardized citation indicators. *PLOS Biol.* 18, e3000918. doi: 10.1371/journal.pbio.3000918
- Ioannidis, J. P. A., Klavans, R., and Boyack, K. W. (2016). Multiple citation indicators and their composite across scientific disciplines. *PLOS Biol.* 14, e1002501. doi: 10.1371/journal.pbio.1002501
- Ioannidis, J. P. A., and Thombs, B. D. (2019). A user's guide to inflated and manipulated impact factors. *Eur. J. Clin. Invest.* 49, e13151. doi: 10.1111/eci.13151
- Kacem, A., Flatt, J. W., and Mayr, P. (2020). Tracking self-citations in academic publishing. *Scientometrics* 123, 1157–1165. doi: 10.1007/s11192-020-03413-9
- Larcombe, A. N., and Voss, S. C. (2011). Self-citation in American Physiological Society journals. *Scientometrics* 87, 347–356. doi: 10.1007/s11192-010-0328-z
- Leimu, R., and Koricheva, J. (2005). Does scientific collaboration increase the impact of ecological articles? *Bioscience* 55, 438–443. doi: 10.1641/0006-3568(2005)055-0438:DSCITI-2.0.CO;2
- Lockwood, M. (2020). Editorial: citation malpractice. *Proc. R Soc. A* 476, 20200746. doi: 10.1098/rspa.2020.0746
- Ma, Y., McBratney, A., and Minasny, B. (2021). Paper self-citation rates of leading soil science journals. *CATENA* 202, 105232. doi: 10.1016/j.catena.2021.105232
- Mishra, S., Fegley, B. D., Diesner, J., and Torvik, V. I. (2018). Self-citation is the hallmark of productive authors, of any gender. *PLOS ONE* 13, e0195773. doi: 10.1371/journal.pone.0195773
- Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P. A., Goodman, S. N., et al. (2018). Assessing scientists for hiring, promotion, and tenure. *PLOS Biol.* 16, e2004089. doi: 10.1371/journal.pbio.2004089
- Resnik, D. B., Gutierrez-Ford, C., and Peddada, S. (2008). Perceptions of ethical problems with scientific journal peer review: an exploratory study. *Sci. Eng. Ethics.* 14, 305–310. doi: 10.1007/s11948-008-9059-4
- Rosenblatt, M., Mehta, S., Peterson, H., Dadashkarimi, J., Rodriguez, R. X., Foster, M. L., et al. (2023). Trends in self-citation rates in Neuroscience literature. *eLife* 12, RP88540. doi: 10.7554/eLife.88540.1
- Sanfilippo, F., Tigano, S., Morgana, A., Murabito, P., and Astuto, M. (2021a). Self-citation policies and journal self-citation rate among Critical Care Medicine journals. *J. Intens. Care* 9, 1–5. doi: 10.1186/s40560-021-00530-2
- Sanfilippo, F., Tigano, S., Morgana, A., Murabito, P., and Astuto, M. (2021b). Self-citation policies in anaesthesiology journals. *Br. J. Anaesth.* 126, e21–e25. doi: 10.1016/j.bja.2020.10.002
- Snyder, H., and Bonzi, S. (1998). Patterns of self-citation across disciplines (1980–1989). *J. Inf. Sci.* 24, 431–435. doi: 10.1177/016555159802400606
- Szomszor, M., Pendlebury, D. A., and Adams, J. (2020). How much is too much? The difference between research influence and self-citation excess. *Scientometrics* 123, 1119–1147. doi: 10.1007/s11192-020-03417-5
- Tagliacozzo, R. (1977). Self-citations in scientific literature. *J. Doc.* 33, 251–265. doi: 10.1108/eb026644
- Taşkin, Z., Dogan, G., Kulczycki, E., and Zuccala, A. A. (2021). Self-citation patterns of journals indexed in the journal citation reports. *J. Informetr.* 15, 101221. doi: 10.1016/j.joi.2021.101221
- Thombs, B. D., Levis, A. W., Razykov, I., Syamchandra, A., Leentjens, A. F. G., Levenson, J. L., et al. (2015). Potentially coercive self-citation by peer reviewers: A cross-sectional study. *J. Psychosom. Res.* 78, 1–6. doi: 10.1016/j.jpsychores.2014.09.015
- Thombs, B. D., and Razykov, I. (2012). A solution to inappropriate self-citation via peer review. *Can. Med. Assoc. J.* 184, 1864. doi: 10.1503/cmaj.120597
- Van Noorden, R. (2013). Brazilian citation scheme outed. *Nature* 500, 510–511. doi: 10.1038/500510a
- Van Noorden, R., and Singh Chawla, D. (2019). Hundreds of extreme self-citing scientists revealed in new database. *Nature* 572, 578–579. doi: 10.1038/d41586-019-02479-7
- Van Raan, A. F. J. (1998). The influence of international collaboration on the impact of research results. *Scientometrics* 42, 423–428. doi: 10.1007/BF02458380
- Vinkler, P. (1987). A quasi-quantitative citation model. *Scientometrics* 12, 47–72. doi: 10.1007/BF02016689
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272. doi: 10.1038/s41592-019-0686-2
- Waltman, L. (2016). A review of the literature on citation impact indicators. *J. Informetr.* 10, 365–391. doi: 10.1016/j.joi.2016.02.007
- Wildgaard, L., Schneider, J. W., and Larsen, B. (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics* 101, 125–158. doi: 10.1007/s11192-014-1423-3
- Wilhite, A. W., and Fong, E. A. (2012). Coercive citation in academic publishing. *Science* 335, 542–543. doi: 10.1126/science.1212540
- Wren, J. D., and Georgescu, C. (2022). Detecting anomalous referencing patterns in PubMed papers suggestive of author-centric reference list manipulation. *Scientometrics* 127, 5753–5771. doi: 10.1007/s11192-022-04503-6

- Wren, J. D., Valencia, A., and Kelso, J. (2019). Reviewer-coerced citation: case report, update on journal policy and suggestions for future prevention. *Bioinformatics* 35, 3217–3218. doi: 10.1093/bioinformatics/btz071
- Yu, T., Yu, G., and Wang, M. Y. (2014). Classification method for detecting coercive self-citation in journals. *J. Informetr.* 8, 123–135. doi: 10.1016/j.joi.2013.11.001
- Zhao, D., Strotmann, A., and Cappello, A. (2018). In-text function of author self-citations: Implications for research evaluation practice. *J. Assoc. Inf. Sci. Technol.* 69, 949–952. doi: 10.1002/asi.24046
- Zhivotovsky, L. A., and Krutovsky, K. V. (2008). Self-citation can inflate h-index. *Scientometrics* 77, 373–375. doi: 10.1007/s11192-006-1716-2