Journal of Informetrics 12 (2018) 461–480

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty

Henry Small*

SciTech Strategies, Inc., 105 Rolling Road, Bala Cynwyd, PA 19004, USA

ARTICLE INFO

Article history:
Received 14 February 2018
Received in revised form 21 March 2018
Accepted 22 March 2018

Keywords:
Biomedicine
Highly cited papers
Citation contexts
Method papers
Uncertainty
Hedging
Machine learning
Corpus linguistics
Logistic regression
Empiricism
Lowry's method

ABSTRACT

The top 1000 biomedical papers by number of citations are classified by method, type of method and non-methods by examination of citation contexts. Supervised machine learning is applied to the context data for a training sample of papers which is then used to classify the full list, revealing that words indicating utility are most important for the classification of methods. Further word analysis is carried out using corpus linguistics to uncover context words that characterize non-methods. Hedging words are found to play an important role for non-methods, and several are selected for further analysis with logistic regression. Other variables in the regression are a consensus variable based on the similarity of contexts for a paper and another variable based on whether citations come from "methods" sections of citing papers. Accuracy of predictions from logistic regression is comparable to machine learning. The results are interpreted in terms of the perceived certainty or uncertainty of the underlying knowledge, that is, methods and their outputs have higher certainty, and non-methods higher uncertainty. Evidence is found that hedging is inversely related to citation frequency. Implications of this work for the study of the development of science and the role of methods and tools in biomedical research are discussed.

1. Introduction

For many years scientists and bibliometricians have been puzzled by lists of the most cited papers in science. Why do these lists not conform to our expectation that key discoveries in science such as the theory of relativity, the genetic code, or quantum mechanics should appear near the top of the citation count rankings? Instead, we find that methodologies dominate. Even Eugene Garfield, as he was creating the first citation index in 1961, was somewhat dismayed that a paper by Oliver H. Lowry on protein determination was so heavily cited (Wouters, 1999, 72), so much so that for a moment he had doubts about the usefulness of his index. With stacks of printouts of citations to Lowry on the floor of his office, he wrote to Joshua Lederberg, the Nobel laureate who encouraged him to undertake the project: "I have a sort of panic about this sample and wonder whether this can be useful to anyone." (Wouters, 1999) Lederberg, however, told him not to worry, that the paper was the most frequently quoted paper in biochemistry because it had become the standard method for the protein determination. The Lowry paper turned out to only be the tip of the iceberg and many other highly cited method papers would be highlighted in subsequent years.

* Corresponding author.
E-mail address: hsmall@mapofscience.com

https://doi.org/10.1016/j.joi.2018.03.007
1751-1577/

Why are our expectations so far off the mark? Are we working under the false assumption that citations are a pure reflection of what is important in science, and that discoveries must carry the greatest importance? Garfield's eventual explanation was that breakthrough papers such as the Watson-Crick discovery of the DNA double-helix can be quickly superseded and replaced by improved formulations, or obliterated by becoming standard usage (Garfield, 1977). Of course, it has been found that some discoveries do achieve high rates of citation in a relatively short period of time, and appear on highly cited lists. Using data from a recent study, it was estimated that at least seven percent of papers in the top 1000 papers ranked by total citations were discoveries (Small, Tseng & Patek, 2017). An alternative hypothesis is that papers containing the methods and tools that scientists use to arrive at their findings should be expected to be the most heavily cited. This might be termed the utilitarian hypothesis, and begs the question, what makes Lowry's paper so useful and compelling to scientists? Is the frequent use of some methods simply the reflection of scientists wanting to obtain credible data to support their hypotheses?

The goal of this paper is to study the phenomenon of highly cited papers from the standpoint of what authors say when they cite them within the so-called citation contexts or citing passages. We will analyze citation contexts for linguistic markers that are associated with methods, and explore the hypothesis that high citation rates are associated with the certainty of the knowledge that is generated. Citation contexts for non-method papers such as discoveries will also be examined for possible linguistic cues that differentiate them from method papers and reveal their role in the knowledge system.

2. Background

The important role of methods in the advancement of biomedical knowledge has often been commented on. For example, Olby in his history of the double helix describes the crucial role that methods played in the elucidation of the structure of DNA (Olby, 1974, 435). Generally, methods and tools in science are seen as providing relatively firm points of reference against which theories can be tested or constructed. For example, Pierre Duhen asserted "Agreement with experiment is the sole criterion of truth for a physical theory." (Duhem, 1962, 21). And John Ziman commented ". . experimental evidence is public knowledge, *par excellence*, with the power of carrying complete conviction." (Ziman, 1968, 32). A successful theory can be seen as consisting of a mix of assumptions and empirical findings which fit together like the pieces of a puzzle. In Kuhn's theory, the paradigm provides a framework of high certainty for experimental and theoretical findings (Kuhn, 1970). In times of crisis, however, when experiment disagrees with theory, the weak link in the chain of reasoning must be found. According to Duhem, this is often more a matter of intuition than of logic (Duhem, 1962, 216). In the history of science, it is most often the theoretical constructs that must give way rather than the experimental findings obtained from the application of methods and tools. Thus, the development of science is critically dependent on the perceived uncertainty of theoretical constructs and the relative certainty of experimental methods.

Garfield was the first to draw attention to the prevalence of method papers in highly cited lists. Over the years, he published numerous essays in *Current Contents* that presented lists of most cited papers from various time periods and journal subsets that highlighted the prominence of methods (Garfield, 1977; Garfield, 1990; Garfield, 1991). In addition, the Citation Classics Commentary series published in *Current Contents*, where authors discussed their highly cited papers, often featured method papers. Lowry, himself, provided material for one such commentary in which he stated that his method, though widely cited, was not a great scientific accomplishment, but merely a more reliable version of earlier methods (Lowry, 1977). Garfield commented, "Is any reasonable person going to claim that the intellectual achievement represented by Einstein's Unified Field Theory is less significant than a convenient method of protein determination simply because Einstein is cited less frequently?" (Garfield, 1973). He goes on to suggest that perhaps it has to do with the relative number of investigators doing protein determination versus field theory.

Method papers also emerged as an issue in early clustering experiments with co-citation (Small & Griffith, 1974). It was found that very highly cited method papers had to be removed or normalized prior to clustering to break up large macro-cluster or giant components that joined together the various specialty clusters. Method papers were like diffuse clouds hovering over the specialties. This work illustrated the trans-specialty and sometimes trans-disciplinary nature of methods.

Studies that attempt to classify the reasons papers are cited usually come up with substantial numbers of citations that fall into a "methods or tools" category (Bornmann & Daniel, 2008), and many of the citer motivation classification schemes have explicit categories for the citation of methods. However, these studies usually are focused on samples of citing papers and do not look at the nature of the cited work. More recent studies that attempt to automate the recognition of citer motivation use a combination of the linguistic analysis of the citation context and location within the IMRaD structure of the scientific paper, but the focus is on the citing instance and not the cited work (Bertin, Atanassova, Sugimoto, 2016; Teufel, Siddharthan, & Tidhar, 2006). Recent studies of the distribution of references across the IMRaD structure of citing papers have found that method sections contain fewer and older references than other sections (Bertin, Atanassova, Gingras, 2016). Another similar study not explicitly looking at IMRaD sections found a consistent text location (measured in character centiles) for highly cited papers which the authors inferred was the location of the methods section but otherwise did not examine the nature of the cited work (Boyack, van Eck, Colavizza & Waltman, 2018).

Recently the journal *Nature* published a study of the most cited 100 papers with data obtained from the Web of Science (Van Noorden, Maher & Nuzzo, 2014). The authors begin by pointing out that some of the landmark discoveries of the 20th century do not appear in the top 100 papers, and on the contrary ". . . the vast majority describe experimental methods or software that have become essential in their fields." They claimed: "To make exciting advances, researchers rely on relatively

unsung papers to describe experimental methods, databases and software." As expected, the Lowry paper was in second place on the list.

Another issue beginning to be discussed is the relative certainty or uncertainty of scientific findings, and whether cue words or hedging terms can be used as indicators. Hyland, a pioneer in the study of hedging, states that such terms convey ". . both epistemic and affective meaning – that is, they not only carry the writer's degree of confidence in the truth of a proposition, but also an attitude to the audience." (Hyland, 2004, 87). Recently Chen and Song have analyzed the various ways that uncertainty is expressed in scientific papers by use of a variety of hedging terms and phrases. They argue that uncertainty detected in this way is a form of meta-knowledge (Chen & Song, 2018). DiMarco, Kroon and Mercer found that hedging cues are strongly correlated with citation contexts in scientific texts (DiMarco, Kroon & Mercer, 2006), and the same research group also observed that hedging terms appear less frequently in methods sections than other sections, attributing this to stylistic factors such as a more formulaic presentation in the methods section (Mercer & DiMarco, 2003). Earlier Latour and Woolgar had introduced the notion of linguistic modalities used by authors when discussing the work of others (Latour & Woolgar, 1979). In their view, words such as "reported", "first", "convincing", "difficult", "support" and "suggested" are used to convey or modify the degree of certainty or uncertainty in the underlying knowledge. Only when such modalities are omitted can the statement be considered to be a scientific fact. Latour also proposed a scale of what he called "facticity" on which the evolution of an idea could be plotted from initial uncertainty to its status as a scientific fact (Latour, 1987, 44).

The present paper is an outgrowth of a previous study that used machine learning and citation contexts to identify discovery papers in biomedicine (Small, Tseng & Patek, 2017). One of the findings of that study is that citation contexts citing discovery papers had the highest concentration of citation contexts containing the word "confirmed" in the first few years after publication compared to citations made in subsequent years. About one-third of the occurrences of "confirmed" in the citing contexts appeared in the first five years after the discovery publication. Examination of these citing sentences revealed that many included an indication of the basis of the confirmation and this basis was most often the application of a specific methodology. Rather than restricting the analysis to cases of confirmation, however, it was decided to look at methods in biomedical science generally. Like the previous study of discoveries, this study assumes that the content and meaning of a cited paper is defined by its usage.

3. Data

The goal of this paper is to explore the citation contexts for method and non-method papers to find linguistic and other cues that reflect how the papers were perceived by the citing authors. It is often difficult in citation context studies to define the scope of the text relevant to a given cited work. However, for the purposes of this paper, it was sufficient to define the citation context as the single sentence in which the citation was made, the so-called "citance" (Nakov, Schwartz & Hearst, 2004). This is justified by the relatively large number of citing sentences for each paper being analyzed. Furthermore, limiting each context to a single sentence helps insure that there is a close association between the linguistic cues and the reference being cited.

In the prior study biomedical discovery papers were identified by searching for a set of words that denote "discovery" (Small, Tseng & Patek, 2017). However, rather than search citances for some set of "method words", the approach was to take a systematic sample of the most cited papers, namely the top 1000 papers in a biomedical data base, and to then determine how many of these highly cited papers were methods or other types of papers. While searching the entire data base for "method words" would have caste a wider net and potentially identified more methods, focusing on the most cited papers allow us to look at a wider range of paper types including both methods and discoveries.

One of the most important sources of full text for the analysis of scientific papers is PubMed Central® (PMC). This open repository created in 2000 includes papers that were required to be publicly available under the National Institutes of Health public access policy. As in the prior study, we limited the study to the full text from PubMed Central called the "open access subset" constructed from data up to mid-2015. This subset includes 1.1 million full texts, primarily in biomedicine, covering publications mainly in the most recent two decades. About 90% of articles are from years 2002–2015. Over the time period, the coverage rapidly expanded from 4500 articles in the year 2000 to about 200,000 in 2014.

The PMC captures the references cited by these papers, and pre-processing adds codes to the references that allow the user to connect the reference within the text to the bibliographic information at the end of the article, as well as, in most cases, providing a unique article identifier for each reference, the "Pub Med ID", which can be used to find the item in the National Library of Medicine's PubMed data base. The references, of course, span a much wider time period than do the source articles from which they are taken. While 90 percent of the articles are from 2002 to 2015, only 67 percent of the references have publication years in this range.

The PMC "open access subset" obtained from NLM was downloaded and parsed into data fields for loading into a MySQL data base as described in the previous study. In addition to the bibliographic information for each article, all references from the articles were parsed, as well as all sentences and paragraphs from each of the full texts, some of which contained cited references. Roughly 38 million references from the articles were loaded into the MySQL data base, on average 34 references per article. Also 166 million sentences were loaded, on average 149 sentences per article. About 19% of the sentences contained one or more references and thus are citing sentences or "citances".

In addition, the full text was parsed for main section headings together with their paragraph locations so that the section in which citations occurred could be determined. Nearly all the papers (99 percent) were found to have section headings,

Table 1
Subtypes of methods in top 1000 papers.

Method subtype	Percentage of all methods
Computational methods	59.6%
Databases	11.3%
Indexes or scales	9.0%
Guidelines	7.7%
Biochemical methods	7.2%
Sequencing methods	5.2%

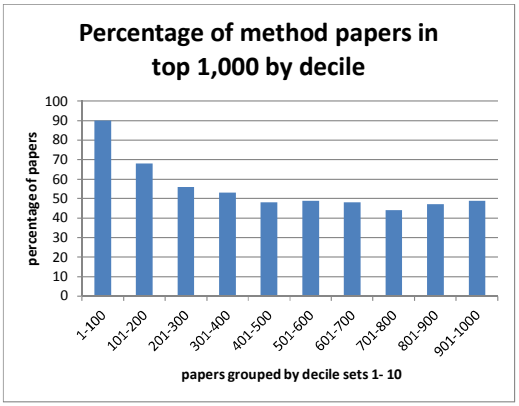

Fig. 1. Percentage of method papers in top 1000 by decile.

although there was a considerable lack of uniformity in section names. The identifier for each citing sentence included a citing article identifier and the paragraph number within the article where the citance occurred. This enabled the identification of the name of the section containing the citance.

The 1000 most cited papers were selected by summing up the number of references per cited item across the PMC subset, and selecting the top 1000 items. These counts include multiple references to the same paper within a given citing paper, sometimes called "op cits". The most cited paper, a method for calculating relative gene expression, the so-called "Ct method", was published in 2001 and was mentioned in 7389 citances. The least cited item has 322 citances and was published in 2013. The oldest paper in the top 1000 is the famous protein determination paper by Lowry from 1951. The most recent paper is from 2014 dealing with the prevalence of prostate cancer. The top 100 papers from the list of 1000 are given in Appendix A.[1]

4. Classification of papers

The first task was to identify the method papers in the top 1000 based on their citances or citing sentences. A total of 646,347 citances were retrieved from the PMC database for the top 1000 papers. The papers were ranked in descending order by total citations and subdivided into ten sets, or deciles, of 100 papers each for convenience in analysis. For example, set 1 consisted of the top 100 papers cited from 7389 to 1077 times (see Appendix A), and set 10 consisted of the bottom 100 papers cited from 341 to 322 times. Manual classification was carried out for all 1000 papers. Each paper was designated as either a method or a non-method based on a scanning of the citances for that paper. Usually it was necessary to read only a small sample of 10–20 citances to determine the nature of a given paper because citing authors tended to repeat similar language when citing the paper (Small, 1978). Later on we will describe a procedure for finding the most characteristic citing sentence for a paper, and a measure of consensus based on the word similarity of a paper's citances.

In the process of method tagging, it was found that methods could be further subdivided into various subtypes including: biochemical methods, sequencing methods, computational methods, indices or scales, databases, and guidelines. The percentage of method papers by subtype is given in Table 1, and reveals that computational methods are by far the most prevalent. This is consistent with the previous study of the top 100 papers from the Web of Science (Van Noorden, Maher & Nuzzo, 2014). The major difference between this distribution of subtypes and earlier compilations by Garfield published in *Current Contents* in the 1970s and '80s is the emergence of computational methods over what had been predominantly a list of biochemical methods. One could argue that we are now in the age of computational biology and medicine.

The distribution of methods across the ten deciles by citation count is given in Fig. 1. The top 100 is heavily skewed toward methods at 90 percent. Starting with the fourth decile this percentage decreases to about fifty percent and remains fairly

[1] The complete list of the 1000 papers including a PubMed identifier is available online or from the author.

steady up to rank 1000. We do not know if this relatively equal division continues to lower citation ranges or if methods begin to decline at some point. Overall 55 percent of the top 1000 are methods.

No systematic attempt was made to further subdivide the non-methods into types, although some potential categories are discoveries and review papers. As in the previous paper on identifying discoveries, a search for "discovery words" in citances was carried out to give a preliminary estimate of the number of discovery papers in the top 1000. Citances for each paper were queried for the string "*discover*" (where asterisks denote a wildcard search) and then manually inspected. A total of 156 or 15 percent of the 1000 papers were identified as discoveries, or 35 percent of the non-methods. This is about twice the number than expected from the previous study which overlapped the top 60 percent of the list of 1000. This underestimate is mainly due to the high cutoff of 20 "discovery" citances used in that previous study. Regarding review papers, we can estimate that about 7 percent of the 1000 papers are reviews given that the number of citances containing the word "review*" is about one-half of the number containing "discovery words".

The 100 papers from set seven, the seventh decile, were used as the basis of an inter-rater reliability study. The second rater was provided with the citances for each of the papers, and used the same guidelines as the first rater on the types of papers to code as methods. The two raters agreed on 96 of the 100 papers; two papers were classified as methods by rater 1 and non-methods by rater 2; and two papers were classified as non-methods by rater 1 and methods by rater 2, for an interrater reliability of 92 percent using Cohen's Kappa (Cohen, 1968).

5. Machine learning

Machine learning had been used in the previous study to distinguish discovery papers from "discovery methods" based on citation contexts. Thus, its application in the present study to distinguish methods from non-methods seemed a natural extension. Furthermore, the procedure could serve as a check on the manual classification by suggesting cases where human judgment was questionable, and also shed light on what words were important in making the distinction.

The 300 papers from sets 8, 9, and 10 were used as a training set for classifying the remaining 700 papers for sets 1 through 7 based on their citances. As can be seen in Fig. 1, the number of methods and non-methods for these three sets are approximately equal. As in the previous study, the citances for each paper were concatenated into a long text string and treated as a "bag of words". This cumulative text was tagged with an identifier for each paper and a code of 1 for method and 0 for non-method. There were a total of 145,408 citances for the 300 papers in the training set, or an average of 487 citances per paper.

The Scikit-learn package was used for machine learning (Pedregosa et al., 2011). This software processes each document, in this case a set of citances for a cited paper, by removing stop-words and computing tf-idf scores for each word (Salton & McGill, 1986). Each document is then represented as a vector consisting of words (coded as numbers) and their associated tf-idf weights. The document vectors define points in a hyper-dimensional space whose axes are individual words. The objective of the training is to find an optimal hyperplane in word space with instances of methods on one side of the plane and non-methods on the other side. Training on the word vectors for the 300 papers using various classifiers defined word coefficients on the hyperplane axes. The resulting solution was applied to the test data consisting of paper identifiers and concatenated citances for the remaining sets 1 through 7 which had not been used in training.

Ten different classifiers available in the Scikit-learn package were tested separately giving accuracies ranging from 82.7 percent to 93.3 percent. Six of the ten classifiers were above 90 percent and four were below 90 percent. The median accuracy was 90.3 percent and the highest accuracy was 93.3 percent obtained with the BernoulliNB classifier. This is a naïve Bayes classifier which is based only on the occurrence of a word, not its frequency, that is, binary valued feature vectors. The classifier differs from other Bayes classifiers in that it penalizes for the non-occurrence of a word if that word is an indicator for a particular class (McCallum and Nigam, 1998). The word coefficients of the hyperplane for the BernoulliNB classifier were retrieved from the model and revealed that the top four words having the highest coefficients were "using", "used", "use", and "based". Thus, what we might call utility or utilization words appearing in the citing sentences were strongly indicative of methods papers, and played a prominent role in the classification. This result is consistent with the previous study of discoveries where so-called "discovery methods" were also associated with utility words. However, no clear pattern was found for the words having the lowest coefficients which corresponded to non-methods.

Focusing on the BernoulliNB classifier allows us to study false positives and false negatives. Across the 700 papers from sets 1 through 7, there were 389 or 56% true positives where the manual classification and machine learning agreed that the paper was a method and 264 or 38% true negatives where there was agreement that the paper was not a method. There were 25 (3.6%) false positives where the manual effort considered the paper a non-method and the machine learning classified it as a method, and 22 false negatives where the manual classification indicated a method and the machine gave a non-method. The F1 statistic, computed from the recall and precision, was 0.94 which is identical to the previous study.

It is instructive to examine the false negatives and positives to see where either the manual or machine classification may have gone wrong. The most prevalent false negatives were nine papers describing guidelines or diagnostic criteria which had been included in the definition of papers to be classified as methods, but were not picked up by machine learning. The second most common false negatives were five papers presenting disease statistics. This type of paper had been considered as a type of database and coded as a method but was not consistently identified by the machine learning. For the false positives, the types of papers giving rise to disagreement are more diverse. Ironically, here also the manual classification missed six specialized databases which the machine did classify as methods. Somewhat related are five papers which presented gene

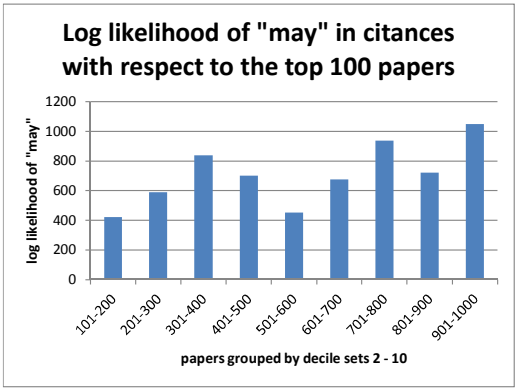

Fig. 2. Each bar represents a set of papers and a corpus of citance words for that set. The log likelihood for the word "may" within each corpus was computed for sets 2 through 10 with respect to set 1, the most cited set with the highest concentration of methods, as the baseline.

expression profiling data, and five papers on genome sequences for organisms which were not manually coded as methods but were classified as such by the machine. Generally, the false positives seemed to be predominantly data compilations of genetic information that were being used in a manner similar to methods.

6. Corpus linguistics

Because no clear indication of the types of words associated with non-methods could be obtained from the hyperplane coefficients, further analyses were undertaken of the relative frequency of words in method and non-method citances. The software package Wordsmith Tools (Scott, 2004) was used to contrast the vocabularies of high method versus low method citances. The software computes keywords ranked by the log likelihood statistic by comparing one corpus of text against another corpus considered as the baseline. As shown in Fig. 1, 90 percent of the papers in set 1 were classified as methods and thus their associated citances can serve as a baseline to compare citances from lower cited sets having a lower proportion of methods, or vice versa, by using the low cited set as the baseline.

Taking the latter approach first, set 1 was compared to set 10 as the baseline. Not surprisingly, the word from set 1 having the highest log likelihood was "using" with a log likelihood of 14,751. The word "used" appeared at rank 33 with a log likelihood of 1692. A similar result was obtained using sets 7, 8 and 9 as the baseline. This result is consistent with machine learning in that utility words had the highest hyperplane coefficients.

The inverse of this comparison is to use set 1 as the baseline and look for words in the lower citation sets that might signal non-method papers. For set 10, the keywords with the highest log likelihoods consist mainly of content or technical words associated with the topics in set 10 that are not represented in set 1. However, a few general, non-technical words appear as well, such as "activity", "associated", and most notably the hedging word "may". In fact, "may" appears at rank 67 with a significant log likelihood of 1049. When the sets 9 through 2 were run through the same analysis using set 1 as the baseline, a similar result was obtained – the word "may" was highly ranked with a statistically significant log likelihood which, however, diminished gradually as the set trended toward the higher citation ranges where methods also became more prevalent. Fig. 2 plots the log likelihood values for "may" for the progression of sets 2 through 10. No other hedging terms as listed, for example, by Hyland appeared as consistently as "may" (Hyland, 2004, 91).

Hedging words have recently been used by Chen and Song as indicators of scientific uncertainty (Chen & Song, 2018). However, we do not know whether uncertainty as indicated by the prevalence of the word "may" was due to the lower citation rate of the papers or the difference between how methods and non-methods are cited. In other words, are methods less likely to be hedged than non-methods? To test this hypothesis we split the citances within a set into those associated with methods and those associated with non-methods. Restricting the citances to a single set means that the citation frequency, while not constant, would nevertheless be limited to the variation within the set. For example, for set 6 the citation count varied from 419 to 463. The keywords obtained for the non-method citances against the methods citations for set 6, again showed that "may" had a significant log likelihood, although the value of the statistic was not as large as obtained when comparing low cited and high cited sets. Splitting method and non-method citances within other sets gave similar results. Thus, it appears that the main reason for the appearance of the hedging term was the heightened uncertainty of the non-method papers or the greater certainty of the methods, rather than some dependence on citation frequency.

To further test whether lower citation rates were associated with greater uncertainty as reflected by the log likelihood of hedging words, a sample of less cited papers was taken. Papers cited around 100 times (>=98 and <=100) gave a sample of 74,175 citances. In contrast, the minimum citation count was 322 in the top 1000. The citances corresponding to non-method papers for sets 1 and 2 were taken as the baseline. If the lower citation rate accounted for the appearance of hedging words, then we would expect to see the word "may" appear with a significant log likelihood. However, this result was not obtained at least for the top 400 words. This supports the notion that hedging or uncertainty is associated with non-methods and is

Table 2
List of Variables.

Variable name	Description
Method	Yes = 1, No = 0
Ln(cites)	Natural log of the number of citations or citances
Age	2015 – Year of publication
Consensus	Mean cosine similarity of each citance for a paper with its cumulation of citances
Using	Percentage of citances for a paper having the words "using", "used" or "use"
May	Percentage of citances for a paper having the word "may"
Show	Percentage of citances for a paper having the word stem "show*"
Suggest	Percentage of citances for a paper having the word stem "suggest*",
Not	Percentage of citances for a paper having the word "not"
Section	Percentage of citances for a paper appearing in "method" sections

less common for methods, and is not a function of citation frequency. We will return to this question later in the paper using a different, more sensitive, approach.

Finally, an attempt was made to expand the list of hedging-like terms by an analysis of citances containing "may". This approach is similar to Chen and Song's in using a term to seed a search for equivalent terms (Chen & Song, 2018). The citances containing "may" were compared to the high method set 1 as the baseline. Some of the hedging-like key words associated with "may" and having significant log likelihoods were "shown", "not", "suggest", "although", "recent" and "however". These words might play some role in distinguishing methods from non-methods. Words such as "may", "suggest", "not", "although", and "however" are considered hedging terms, while a word like "shown" is more an indicator of confirmation, in Hyland's terminology a "booster" word (Hyland, 2004). Whether such terms enhance our ability to differentiate non-methods from methods, however, requires a more sophisticated statistical approach.

7. Descriptive statistics

To more clearly delineate the factors, either linguistic or bibliometric, that differentiate methods from non-methods, a set of variables was defined based on the word analysis, but also including two bibliometric variables and a variable derived from text location. Table 2 gives a list of the variables computed for each paper. Except for the first three, the variable was computed using the citances for each paper. The first variable is a binary variable expressing whether the paper was a method or not, which will be the dependent variable in a logistic regression described below. The age of the paper was computed relative to the last year of data of the data compilation used, namely, 2015 minus the year of publication. The natural log of citation frequency was used instead of the raw citation frequency, although, whether logged or not, it did not prove to be statistically significant.

The degree of consensus among the citing passages was computed as the average cosine similarity of each citance for a paper with the cumulative citances for that paper. This measure is related to one previously used by the author called the "uniformity index" (Small, 1978) and a related metric called "self-cohesion" (Elkiss et al., 2008) which expresses how similar the citing sentences are to one another. If a paper is cited by different authors using the same words, ignoring stop words, then it will have a high consensus score. If there are a wide variety of ways a paper is cited, the consensus will be low. As a by-product of this calculation, the citance having the highest cosine similarity with the citance set as a whole can be identified. This was called previously the "consensus passage" (Small, 1986). This citance is present as the fifth column in Appendix A for the top 100 papers and shows a typical usage by the citing community.

The word variables were computed by calculating the percentage of citances for a paper containing the specific term or terms. For example, the "using" variable was the percentage of citances for a paper that contains one or more of the utility words: "using", "used" or "use". For the top ranked paper in Appendix A on the "Ct method", 64 percent of its citances contained utility words. Variables were also created for other general, non-technical terms or term stems including "may", "suggest*", "show*", and "not", where the asterisk following a word indicates that a wild card search was carried out. This subset of general, non-technical words was found by the corpus linguistic analysis to have the highest log likelihoods in comparisons between non-method and method citances.

The last variable used the section of the paper that the citance was made in, expressed as the percentage of citances for a paper that appeared in sections of citing papers having the word "methods" in their names, for example "Materials and Methods", "Data and Methods", etc. To create this variable the PMC full text was scanned for tags denoting main section headings and at the same time capturing the paragraph number that could be matched with the citance identifiers which also carried the paragraph number. For example, the methods section percentage for the most cited papers in Appendix A on the "Ct method" was 89.1 percent.

It may seem obvious that method papers are cited in the method sections of papers, and this variable came closer to any other to being a "gold standard" for identifying method papers. One limitation is that only about 35 percent of papers have a main section name that contains the word "method" as a stem. Not included were other sections such as an "experimental design" section, but these would have increased the number of relevant citances by only a couple of percent. Nevertheless, this variable turned out to have a high degree of significance in our statistical tests.

Table 3
Descriptive Statistics.

Variable	method	mean	std. dev.	median
ln(cites)	Yes	6.40	0.58	6.20
ln(cites)	No	6.16	0.36	6.10
Age	Yes	14.87	10.56	12.00
Age	No	11.82	6.77	11.00
Consensus	Yes	37.82	6.85	37.40
Consensus	No	31.26	4.85	30.70
Section	Yes	58.43	23.80	65.47
Section	No	7.78	11.18	3.523
Using	Yes	42.66	21.00	43.60
Using	No	6.32	5.84	4.10
May	Yes	1.04	1.00	0.77
May	No	3.79	2.29	3.30
Show	yes	2.25	1.90	1.80
Show	no	6.20	3.92	5.40
Suggest	yes	0.86	1.02	0.46
Suggest	no	2.89	1.81	2.50
Not	yes	2.68	1.85	4.20
Not	no	4.43	2.13	2.30

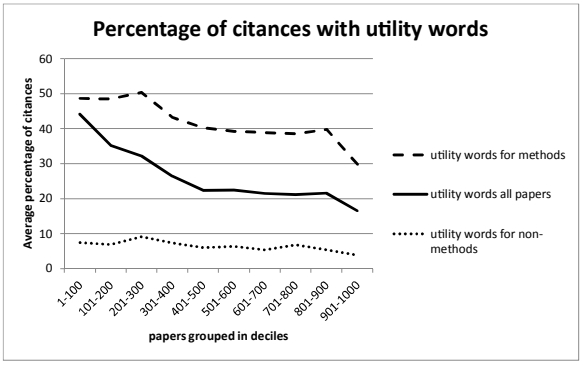

Fig. 3. Average percentage of citances with utility words for methods, non-methods and all papers grouped by decile.

Table 3 is a listing of some descriptive statistics for the bibliometric and linguistic variables contrasting their values for the methods and non-methods across the list of 1000 papers. The first five variables in Table ("ln(cites)", "age", "consensus", "section", and "using") show higher mean and median values for method than for non-method papers, while the last four variables ("may", "show", "suggest" and "not"), all of which are language based, show higher mean and median values for non-methods. The magnitudes of the differences between methods and non-methods are helpful in defining different statistical regression models in the following section.

The trends for these variables across the citation frequency ranges can also help reveal the different behavior of the method and non-method papers. Plotting the mean percentage of citances containing specific words against the papers grouped by citation decile, we can see the effect declining citation frequency has on the rate of word usage. First, for the utility words "using", "used" and "use", the rate of word usage declines with declining citation frequency as shown in Fig. 3, although the extent of the decline is greater for method papers than for non-methods, and there is, of course, a much higher rate of utility word usage for method papers than for non-methods. The situation is reversed for the hedging word "may". In Fig. 4, we see that the rate of "may" increases with declining citation frequency and the non-method papers have a substantially higher "may" rate than the method papers. The major determinant of word usage is thus the type of paper, method or non-method, giving a vertical separation of about 30 percentage points for utility words and a much smaller three percentage point spread for "may". But the trends across declining rates of citation are also of interest and should become more marked at lower citation frequencies.

Fig. 4 shows that there is a tendency for both methods and non-methods to have increasing "may" rates or decreasing utility word rates with decreasing citation frequency. This finding for the hedging term "may" runs counter to our previous finding that citances for lower cited papers are not associated with an increased hedging word rate based on log likelihood, and calls for more research on whether lower cited papers have greater uncertainty. Taking the set of lower cited papers described in the previous section (cited between 98 and 100 times), the average "may" rate was 4.2 percent, which is one percentage point higher than the combined method and non-method "may" rate obtained for set 10 (3.1 percent, see Fig. 4), the lowest cited set in the top 1000. This suggests that the trend toward higher uncertainty with lower citation rate continues, if only gradually. However, no attempt was made in this lower cited sample to control for the method/non-

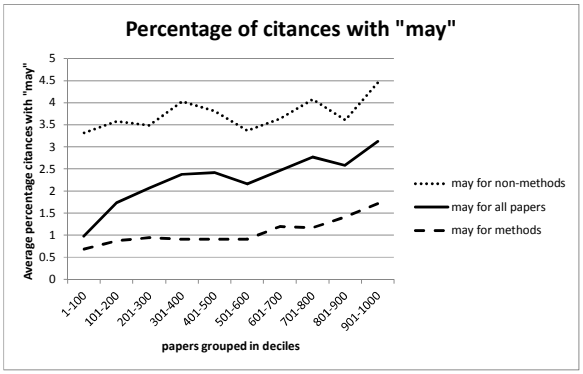

Fig. 4. Average percentage of citances with "may" for methods, non-methods, and all papers grouped by decile.

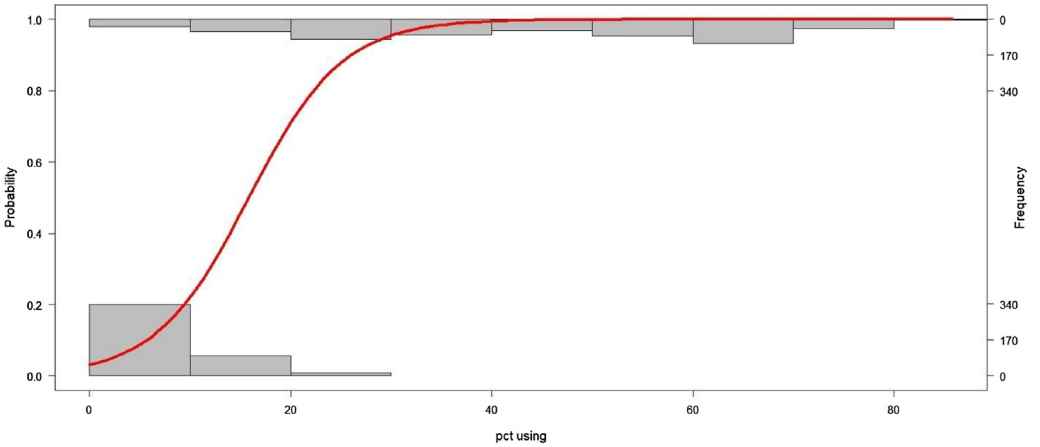

Fig. 5. Logistic curve for the utility words variable labeled as "pct using". The "probability" scale is given on the left vertical axis and the "frequency" of cited papers is on the right.

method mix. In a subsequent experiment, however, a sample of 2000 papers cited 20 times was taken in which methods were differentiated from non-methods.[2] The average percentage of citances with "may" was 5.6% for non-methods and 2.2% for methods, confirming the upward trend at lower citation levels. The way in which word rate percentages were computed, whether by averaging percentages for the papers in a set, or by first summing word citance counts within the set and then dividing by total citances, did not make a difference in the magnitudes or slopes of the curves.

It was also found that a similar trend of increasing word rate with declining citation frequency is displayed by the term "show*" (show, shown, or showed) which goes from 4.4 percent for the most cited set of 100–7.4 percent for the least cited set (set 10). While not considered a hedging term, "show" or "shown" is often used in the presenting of evidence in favor of some hypothesis, suggesting an increase in hypothesis testing at lower citation rates. Aggregating hedging terms also steepened the trend toward higher word rates with declining citation frequency. Doing a combined search for "may", "not", "suggest*" and "show*" the word rate went from 15 percent of citances for non-methods in set 1–21 percent for set 10, a six percentage point increase from high to low citation ranges. These trends suggest that there is a gradual increase in uncertainty with declining citation frequency. However, we do not know how far this trend can be extrapolated.

8. Logistic regression

Since our main goal was to use citation contexts to find the language variables that best predict whether the paper is a method or not, it seemed appropriate to use logistic regression to determine how much each variable contributed to the binary outcome. Logistic regression was carried out using the statistical package "R" (R Core Team, 2017). The binary dependent variable was whether the paper was a method (=1) or not (=0). A logistic curve was fitted to the binary variable and one or more independent variables. For example, the logistic regression curve is shown in Fig. 5 for the utility word variable ("using", "used" and "use"). This graph plots the logistic curve that was fitted to the data. The histogram appearing

[2] This work will be more fully described in a later paper. Details are available from the author.

Table 4
Predictive power of word variables, consensus, and section.

Word or word combination	Accuracy	AIC
using + may + consensus + not + show + suggest	92.0%	390.8
section + using + may + consensus + not + show	91.9%	388.0
using + consensus + show + not	91.9%	401.8
using + may + consensus + not + show	91.6%	390.5
using + consensus + show	91.6%	405.6
using + may + consensus + not + suggest	91.5%	413.8
using + may	91.2%	443.7
using + consensus + suggest	91.1%	423.6
using + may + consensus	91.0%	423.2
section	90.7%	520.4
using	89.5%	493.1
may	83.3%	793.3
suggest	76.6%	957.1
show	75.8%	981.6
consensus	71.5%	1112.0
not	68.0%	1194.8

at the top corresponds to the number of papers categorized as methods with a probability of one across the utility word variable, and at the bottom, the number of non-method papers corresponding to a probability of zero. The papers with a high rate of citances with utility words are shown to be concentrated at the top right, while the papers with a low percent utility are concentrated at the lower left. The logistic curve itself, going from lower left to upper right, represents the observed values for each paper fitted to a probability scale from zero to one on the vertical axis which expresses the likelihood that a given paper is a method based on the utility word variable labeled "using" in this graph. The output of the logistic regression analysis is given in Appendix B. From that output the equation for the logistic curve can be written as: probability of being a method paper = 1/(1 + exp(-(-3.424 + 0.216*pctutility))), where pctutility is the percentage of citances for a paper containing utility words.

The coefficient in the logistic regression equation of 0.216 can be interpreted to mean that the log odds of being a method papers increases by 0.216 with a one percentage point increase in the utility word variable, or alternatively, the odds of the paper being a method increases by a factor of 1.2. The statistical significance of this variable is very high with a p value less than 2*E-16 for the Wald statistic (Allison, 1999).

Not surprisingly, the section variable also gives a highly significant Wald statistic for the method data. In addition, both "using" and "section" had significant goodness-of-fit statistics (p = 0.78 for the utility variable and p = 0.33 for the section variable) with the Hosmer and Lemeshow goodness-of-fit test (Hosmer & Lemeshow, 1989, Chapt. 5). This test compares the fitted model to a maximal model and tests the difference using chi-square. A high p-value here indicates a good fit between the actual and maximal models, and fails to reject the null hypothesis of equivalence. Because the section variable is continuous and not binary, it is also possible to use ordinary linear regression to predict the section variable with the other variables in Table 3 as independent variables. This model was able to explain 80 percent of the variance using the adjusted R-squared.

The other variables listed in Table 3 were also tested as independent variables, either singly or in combination, to predict the binary method variable in logistic regression. It turned out that the consensus variable was statistically significant, but neither the citation count (whether logged or not) nor the age variable was significant. On the other hand, a number of the word variables were statistically significant, namely "using", "may", "show", and "not". Among these, the utility word variable (called "using") had the highest significance.

Results of a logistic regression for the four word variables ("using", "may", "not" and "show") plus the consensus variable are given in the Appendix C. All of these variables are statistically significant and contribute to predicting whether the paper is a method or not. The goodness-of-fit of this model was tested using the Pearson-Windmeijer chi-square test giving a p-value of 0.77 (P.D Allison, personal communication, January 24, 2018) where, like the Hosmer-Lemeshow test, a high value of p indicates a good fit and acceptance of the null hypothesis. (Windmeijer, 1990).

Each word variable or word combination can be evaluated for its predictive ability by converting to probabilities and seeing how many method papers have a probability above 0.5 and how many non-method papers have a probability less than 0.5. For example referring to Fig. 5 for the utility word variable, it turns out that 487 of the papers classified as methods have a probability over 0.5, and 408 non-method papers have a probability of less than 0.5. The sum of these numbers gives an accuracy of 895 out of 1000 or 89.5 percent for the "utility" variable alone (see Table 4). The number of false negatives and false positives (65 and 40 respectively) can also be calculated by counting the number papers below the cutoff for methods and above the cutoff for non-methods. In contrast, the section variable, which was expected to be highly accurate, predicted that 489 of the method papers have a probability of over 0.5, and 418 non-methods have a probability of less than 0.5, for an accuracy of 90.7 percent (63 false negatives and 30 false positives).

Table 4 summarizes the predictive power for various word or word combination variables, including the section and consensus variables. The accuracy in the second column gives the percentage of correct predictions using the 0.50 probability cutoff, and the AIC, or Akaike Criterion, which is a statistic that compares the fit of different models where lower values

indicate a better fit. As can be seen, the accuracy and AIC are approximately inversely related. Note that a random independent variable would give a correct prediction of about one-half of the time, or an accuracy of 50 percent.

Table 4 does not list all possible combinations of the selected words, but it appears that the upper limit on accuracy for these variables is 92 percent. This is comparable to the highest accuracy from machine learning of 93 percent obtained with the Bernoulli Naïve Bayes classifier. However, in contrast to machine learning, the predictive results obtained with logistic regression were achieved with a vocabulary of only a few words plus the section and consensus variables. Also, it is interesting to note that while section is the best performing single variable, some combinations of utility and hedging words achieve a comparable accuracy.

It cannot be claimed that other general marker words (or word combinations) would not have performed equally well in the logistic regression, although our selection was based on words with the highest log likelihood obtained from corpus linguistic. One important point is that the accuracy of prediction works better with each marker word as a separate independent variable in logistic regression, rather than combining the terms with "or" logic in a single search of the citances. For example, an expanded set of hedging terms searched with "or" logic ("may", "not", "suggest*", "although", and "however") did not give statistically significant results in logistic regression. Nor did an expanded set of hedging terms help predict non-methods. The reason why word combinations do not perform as well as single, separate word variables is not clear, but may have to do with the dilution of some strong signals by weaker ones.

9. Discussion

Derek Price, one of the pioneers of scientometrics, was also a historian of technology. In an essay entitled "Of sealing wax and string", he discussed the role of technology and instrumentation in the history of science, and stated the controversial view that ". . historically the arrow of causality is largely from the technology to the science." (Price, 1983) Later in that essay he commented on how simple methods and techniques such as "Lowry's neat method for protein analysis" were among the most cited papers of all time. Our aim has not been to show that methods cause or precede discovery, but rather to show the prominent role methodologies have in biomedical science, especially the new computational methods, and to uncover the language markers that characterize their use. Utility words have been found to be very reliable markers. This is hardly unexpected, but the level of predictive ability is perhaps surprising. Of equal importance are the words associated with non-methods and not associated with methods. Hedging terms, and a few other words not normally considered hedging terms, were found to characterize non-methods but not methods. This suggests that methods, and the information they generate, have a firmer epistemological grounding than the knowledge represented by non-methods.

Why methods are seen to be more certain than non-methods is a topic that requires further investigation. Hanson's view that all experimental observations are theory-laden still applies (Hanson, 1972), but apparently non-methods are perceived as more theory laden and more tentative than methods. Hence, scientific knowledge appears to operate under an implicit or tacit hierarchy of certainty. We may still be skeptical about the outcome or efficacy of a method, but we are likely to be more uncertain about a discovery or research finding. A next step in this research is to explore this latent dimension of scientific knowledge and try to model uncertainty by some means. For example, does certainty derive from frequency of use, dependence on theory, observability, reliability, simplicity, transparency, or possible social interests as constructivists would assert? Also the implications of an uncertainty dimension for scientific change should be explored. For example, does uncertainty influence scientific change, theory choice, hypothesis generation or the assumption modification during scientific anomalies (Duhem, 1962)?

It is also possible to broaden the scope of our analysis from the citance or single sentence to a paragraph of text to see if other factors such as confirmation or corroboration accompany method use, or whether specific tools, such as apparatus or equipment, are associated with application of methods. For example, a glance at a single citing paragraph for the most cited method paper in Appendix A on the "Ct method" shows a myriad of physical devices and materials associated with its application: spectro-photometers, a Cloned AMV First Strand Synthesis kit, SYBR green dye chemistry using the qPCR kit, a CFX96 thermocycler, and Tribolium primers. These are often readily identifiable by their commercial names. This is reminiscent of Latour's description of work in the lab where instruments were invoked as "inscription devices" whose purpose was to generate facts and move scientific ideas up the ladder of "facticity" (Latour and Woolgar, 1979). In addition to "fact" creation, we can expect that methods play a key role in discovery as well as the confirmation and validation of findings as previously noted. For discovery we would need to look for antecedent methods cited in discovery papers, and for confirmation, the subsequent association of methods in citing paragraphs for the discoveries.

Of course, it is possible to be misled by an inaccurate or unreliable method. If, for example, Lowry's methods turned out to give an erroneous determination of proteins, this could have had catastrophic consequences for biomedical research. A critic of science might argue that the frequent use of methods that we observe is just an expression of misguided empiricism. Karl Popper asserted "Science does not rest upon rock-bottom. The bold structure of its theories rises, as it were, above a swamp." (Popper, 1961, 111). For him methods, from which scientists obtain empirical evidence, were no more secure than the theories they supported. On the other hand, this may simply be the way science works, and what sets science apart from mysticism or other ways of knowing. Popper, though rejecting the idea that sense-experience led directly to knowledge, stated ". . . how could we ever reach any knowledge of facts if not through sense-perception?. . Thus perceptual experience must be the sole 'source of knowledge' of all the empirical sciences." (Popper, 1961, 94).

Conclusions

This paper has shown that the citances for highly cited method papers carry much different word markers than non-methods and, conversely, non-method citances carry more hedging terms. Hedging words are, at this point, the best indicators we have to assess the uncertainty of knowledge claims. Some preliminary evidence has been uncovered that the perceived uncertainty of a paper, as measured by the rate of hedging terms in its citances, is inversely related to its citation frequency although the trend is very gradual. Since this dependence on citation frequency is a subtle effect, more careful measurements over a wider range of citation counts are called for. Future studies will need to control for methods and non-methods even at these lower citation ranges since methods can mask and dilute the effect of non-methods and their higher hedging rates.

The relative absence of hedging terms in method paper citances suggests their higher perceived certainty and provides a new explanation for the well-known bibliometric finding that method papers predominate at the high end of the citation ranking. Scientists use methods frequently because they need to generate reliable and certain knowledge. This finding is also consistent with philosophies of science that assert the primacy of experimental and observational evidence provided by methods and tools over theoretical claims. The sources and causes of this primacy pose a new challenge for scientometric and informetric research.

Acknowledgments

I would like to thank Paul D. Allison for advice on logistic regression, Mike Patek for programming, David Pendlebury and Hung Tseng for comments, and Dick Klavans and Kevin Boyack for valuable discussions.

Appendix A.

The columns from left to right are: 1. a numeric id for each paper in inverse order by citation count; 2. the citation count or number of citances for the paper; 3. the year the paper was published; 4. a code indicating the method paper subtype (or N for non-method); 5. the most typical citance based on a cosine similarity analysis with all the citances for the paper; 6. the title of the paper; 7. the authors of the paper, and 8. the journal where the paper was published. The subtype codes are: "C" for computational method; "I" for index or scale; "B" for biochemical method; "S" for sequencing method; "G" for guideline; and "D" for database. A complete list of the 1000 papers including the PubMed Id is available online or from the author.

Appendix A: The top 100 papers

id	cite freq	pub year	code	citance	Title	authors	journal
1	7389	2001	C	Relative gene expression was calculated using the comparative Ct method (2(-Delta Delta Ct).	Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method	Livak, KJ; Schmittgen, TD	Methods
2	6879	1997	C	All 44 sequences were compared against the NCBI non-redundant protein sequence database using PSI-BLAST.	Gapped BLAST and PSI-BLAST: a new generation of protein database search programs	Altschul, SF; Madden, TL; Schaffer, AA	Nucleic Acids Res
3	6174	1990	C	These sequences were used in BLASTP searches using BLAST 2.2.22 at NCBI against the database of non-redundant protein sequences for each organism individually.	Basic local alignment search tool	Altschul, SF; Gish, W; Miller, W; Myers, EW; Lipman, DJ	J Mol Biol
4	4871	2000	C	RTCs that were present in the Gene Ontology (GO) database were used for functional clustering.	Gene ontology: tool for the unification of biology	Ashburner, M; Ball, CA; Blake, JA; Botstein, D; Butler, H; Cherry, JM; Davis, AP; Dolinski, K; Dwight, SS; Eppig, JT	Nature Genet
5	4525	1976	B	The protein concentration was determined by the Bradford method, using bovine serum albumin as the standard protein.	Rapid and sensitive method for quantitation of microgram quantities of protein utilizing principle of protein-dye binding	Bradford, MM	Anal Biochem
6	4342	2011	C	Sequences were aligned by using MUSCLE, and phylogenetic and molecular evolutionary analyses were conducted by using MEGA version 5, using the neighbor-joining tree-building method, with 1000 bootstrap replicates.	MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods	Tamura, K; Peterson, D; Peterson, N; Stecher, G; Nei, M; Kumar, S	Mol Biol Evol

7	4221	1994	C	Sequences were aligned using ClustalW.	ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice	Thompson, JD; Higgins, DG; Gibson, TJ	Nucleic Acids Res
8	3342	2004	C	The sequences were aligned using MUSCLE.	MUSCLE: multiple sequence alignment with high accuracy and high throughput	Edgar, RC	Nucleic Acids Res.
9	3294	2007	C	The phylogenetic tree was constructed using the Mega 4 program using the neighbor-joining method .	MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0	Tamura, K; Dudley, J; Nei, M; Kumar, S	Molecular Biology and Evolution
10	3262	1970	B	Proteins were separated by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) according to the method described by Laemmli.	Cleavage of structural proteins during the assembly of the head of bacteriophage T4	Laemmli, UK	Nature
11	3095	2004	N	MicroRNAs (miRNAs) are a class of small non-coding RNAs that regulate gene expression by binding to their target mRNAs and triggering either protein translation repression or RNA degradation.	MicroRNAs: genomics, biogenesis, mechanism, and function	Bartel, DP	Cell
12	3021	2011	D	Breast cancer is the most common cancer and the leading cause of cancer death among women worldwide, accounting for 23% of the total cancer cases and 14% of the cancer deaths in 2008.	Global cancer statistics	Jemal, A; Bray, F; Center, MM; Ferlay, J; Ward, E; Forman, D	CA Cancer J Clin
13	2841	2007	C	Sequences were aligned using ClustalW software for multiple sequence alignment.	Clustal W and Clustal X version 2.0	Larkin, MA; Blackshields, G; Brown, NP; Chenna, R; McGettigan, PA; McWilliam, H; Valentin, F; Wallace, IM; Wilm, A; Lopez, R; Thompson, JD; Gibson, TJ; Higgins, DG	Bioinformatics
14	2540	2009	C	A gene ontology (GO) enrichment analysis of these genes was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID).	Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources	da Huang, W; Sherman, BT; Lempicki, RA	Nat Protoc
15	2520	1975	I	Cognitive function: Mini-Mental State Examination (MMSE).	"Mini-mental state". A practical method for grading the cognitive state of patients for the clinician	Folstein, MF; Folstein, SE; McHugh, P	Journal of Psychiatric Research
16	2497	2007	C	For the analysis of Data sets 1 through 4, we performed association test on sex using the PLINK whole genome association analysis toolset.	PLINK: a tool set for whole-genome association and population-based linkage analyses	Purcell, S; Neale, B; Todd-Brown, K	Am J Hum Genet
17	2488	1987	C	The phylogenetic tree was constructed using the Neighbor-Joining method.	The neighbor-joining method: a new method for reconstructing phylogenetic trees	Saitou, N; Nei, M	Mol Biol Evol.
18	2478	2009	C	We mapped the reads to the genome using the BWA aligner converting SAM to BAM format using samtools.	The Sequence Alignment/Map format and SAMtools	Li, H; Handsaker, B; Wysoker, A; Fennell, T; Ruan, J; Homer, N; Marth, G; Abecasis, G; Durbin, R	Bioinformatics
19	2425	2011	N	Since the discovery of T cells, B cells, and antibodies specific for tumor antigens, several clinical studies have clearly demonstrated that a high density of T-cell or B-cell subsets within the tumor microenvironment is associated with increased patient survival, such as in colorectal cancer, primary cutaneous melanoma, breast cancer, non-small cell lung cancer (NSCLC), head and neck cancer, and ovarian cancer.	Hallmarks of cancer: the next generation	Hanahan, D; Weinberg, RA	Cell

20	2398	2003	C	Bayesian phylogenetic analysis was performed using MrBayes v3.1.2.	MrBayes 3: Bayesian phylogenetic inference under mixed models	Ronquist, F; Huelsenbeck, JP	Bioinformatics
21	2268	2005	C	These gene sets were also used for pathway analysis using Gene Set Enrichment Analysis (GSEA).	Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles	Subramanian, A; Tamayo, P; Mootha, VK	Proc Natl Acad Sci USA
22	2262	2009	C	Reads were mapped to the genome using bowtie.	Ultrafast and memory-efficient alignment of short DNA sequences to the human genome	Langmead, B; Trapnell, C; Pop, M; Salzberg, SL	Genome Biol
23	2235	2009	C	Reads were aligned with the reference genome using BWA.	Fast and accurate short read alignment with Burrows-Wheeler transform	Li, H; Durbin, R	Bioinformatics
24	2221	2004	C	Normalization and analysis of the data was done in R/Bioconductor using the Bioconductor package “limma”.	Bioconductor: open software development for computational biology and bioinformatics	Gentleman, RC; Carey, VJ; Bates, DM; Bolstad, B; Dettling, M; Dudoit, S; Ellis, B; Gautier, L; Ge, Y; Gentry, J; Hornik, K; Hothorn, T; Huber, W; Iacus, S; Irizarry, R; Leisch, F; Li, C; Maechler, M; Rossini, AJ; Sawitzki, G; Smith, C; Smyth, G; Tierney, L; Yang, JY; Zhang, J	Genome Biol
25	2218	1997	C	Sequences were aligned using ClustalX.	The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools	Thompson, JD; Gibson, TJ; Plewniak, F; Jeanmougin, F; Higgins, DG	Nucleic Acids Res
26	2213	1951	B	The protein concentration was determined by the Lowry method using bovine serum albumin as a protein standard.	Protein measurement with the Folin phenol reagent	Lowry, OH; Rosebrough, NJ; Farr, AL; Randall, RJ	J. Biol. Chem.
27	2156	2000	D	Structures are obtained from the Protein Data Bank (PDB).	The protein data bank	Berman, HM; Westbrook, J; Feng, Z	Nucleic Acids Res
28	2148	2003	C	Using this model, a maximum likelihood (ML) tree was inferred using PHYML .	A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood	Guindon, S; Gascuel, O	Syst Biol
29	2035	2001	C	Relative gene expression was calculated using the Pfaffl method.	A new mathematical model for relative quantification in real-time RT-PCR	Pfaffl, MW	Nucleic Acids Res.
30	2008	2000	C	Population genetic structure was also inferred using a Bayesian model-based clustering analysis in the program STRUCTURE 2.3.1.	Inference of population structure using multilocus genotype data	Pritchard, JK.; Stephens, M; Donnelly, P	Genetics
31	1892	2001	N	The preponderance of long interspersed elements type 1 (LINE–1 s or L1s) in the human genome was eloquently revealed by the human genome sequencing project: as the most abundant autonomous transposable element in the human genome.	Initial sequencing and analysis of the human genome	Lander, ES; Linton, LM; Birren, B; Nusbaum, C; Zody, MC; Baldwin, J; Devon, K; Dewar, K; Doyle, M; FitzHugh, W	Nature
32	1884	1998	C	Clustering analysis was performed by using Cluster 3.0 and the hierarchical clustering of genes method, and the clusters were visualized using Java TreeView.	Cluster analysis and display of genome-wide expression patterns	Eisen, MB; Spellman, PT; Brown, PO; Botstein, D	Proc Natl Acad Sci U S A
33	1880	1986	C	In the absence of heterogeneity between the studies, the pooled estimate of each study was calculated using the Mantel-Haenszel method for a fixed-effects model, otherwise, the random-effects model by the DerSimonian and Laird method was used.	Meta-analysis in clinical trials	DerSimonian, R; Laird, N	Control Clin Trials

34	1856	2000	N	While the root of cancer cell proliferation is the result of a loss of growth and cell cycle regulatory controls within the cancer cells themselves, changes in the way cancer cells interact with the surrounding environment are also critical to tumor development and clinical cancer.	The hallmarks of cancer	Hanahan, D; Weinberg, RA	Cell
35	1849	2003	C	Statistical heterogeneity between studies was assessed by both the Q-statistic and the I2 test statistic.	Measuring inconsistency in meta-analyses	Higgins, JP; Thompson, SG; Deeks, JJ; Altman, DG	BMJ
36	1818	1997	C	Publication bias was assessed by funnel plot and Egger's regression test.	Bias in meta-analysis detected by a simple, graphical test	Egger, M; Davey Smith, G; Schneider, M; Minder, C	BMJ
37	1797	2001	C	Differentially expressed genes were identified using Significance analysis of microarrays (SAM).	Significance analysis of microarrays applied to the ionizing radiation response	Tusher, VG; Tibshirani, R; Chu, G	Proc Natl Acad Sci USA
38	1772	2005	C	Linkage disequilibrium (LD) was calculated using Haploview software.	Haploview: analysis and visualization of LD and haplotype maps	Barrett, JC; Fry, B; Maller, J; Daly, MJ	Bioinformatics
39	1765	1985	I	Assessment of insulin resistance was calculated by homeostasis model assessment of insulin resistance (HOMA-IR).	Homeostasis model assessment: insulin resistance and cell function from fasting plasma glucose and insulin concentrations in man	Matthews, DR; Hosker, JP; Rudenski, AS; Naylor, BA; Treacher, DR; Turner, RC	Diabetologia
40	1760	2004	C	Refinement was performed using PHENIX and model building using Coot.	Coot: model-building tools for molecular graphics	Emsley, P; Cowtan, K	Acta Crystallogr D
41	1741	2005	D	Liver cancer is the sixth most common cancer worldwide, accounting for 5.7% of new cancer cases, and the third most common cause of cancer-related death.	Global cancer statistics, 2002	Parkin, DM; Bray, F; Ferlay, J; Pisani, P	CA Cancer J Clin
42	1701	2008	C	Reads were assembled de novo using Velvet.	Velvet: algorithms for de novo short read assembly using de Bruijn graphs	Zerbino, DR; Birney, E	Genome Res
43	1699	2000	D	All genes were firstly mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Gene Ontology database.	KEGG: Kyoto encyclopedia of genes and genomes	Kanehisa, M; Goto, S	Nucleic Acids Research
44	1687	2003	C	Data were normalized using the robust multi-array average (RMA) algorithm.	Exploration, normalization, and summaries of high density oligonucleotide array probe level data	Irizarry, RA; Hobbs, B; Collin, F; Beazer-Barclay, YD; Antonellis, KJ; Scherf, U; Speed, TP	Biostatistics
45	1686	2009	N	MicroRNAs (miRNA) are small non-coding RNAs that regulate the expression of target genes by binding to the untranslated regions of target mRNA	MicroRNAs: target recognition and regulatory functions	Bartel, DP	Cell
46	1685	2008	S	Gene expression levels can be represented as reads per kilobase per million mapped reads (RPKM) in RNA-Seq.	Mapping and quantifying mammalian transcriptomes by RNA-Seq	Mortazavi, A; Williams, BA; McCue, K; Schaeffer, L; Wold, B	Nature Methods
47	1680	1971	I	All participants were right-handed as assessed by the Edinburgh Handedness Inventory.	The assessment and analysis of handedness: the Edinburgh inventory	Oldfield, RC	Neuropsychologia
48	1680	2003	C	The network was visualized by using Cytoscape.	Cytoscape: a software environment for integrated models of biomolecular interaction networks	Shannon, P; Markiel, A; Ozier, O; Baliga, NS; Wang, JT; Ramage, D; Amin, N; Schwikowski, B; Ideker, T	Genome Res
49	1637	1977	C	Kappa-values were interpreted qualitatively according to Landis and Koch: appa-values less than 0 indicate poor agreement, 0.00-0.20 slight agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, and 0.81-1.00 excellent agreement.	The measurement of observer agreement for categorical data	Landis, JR; Koch, GG	Biometrics
50	1596	2006	N	Mouse and human somatic cells can be reprogrammed to embryonic stem cell (ESC)-like cells, known as induced pluripotent stem cells (iPSCs), classically by ectopic expression of transcription factors, Oct4, Klf4, Sox2, and c-Myc, or by other combinations.	Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors	Takahashi, K; Yamanaka, S	Cell

51	1554	2005	N	One miRNA can bind to hundreds of target genes at the UTR of mRNAs (miRNA target genes), and a single miRNA target gene can be targeted by multiple miRNAs.	Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets	Lewis, BP; Burge, CB; Bartel, DP	Cell
52	1552	2006	C	The phylogenetic tree was constructed by maximum likelihood using RAxML with 100 bootstrap replicates.	RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models	Stamatakis, A	Bioinformatics
53	1516	1992	I	For scores of general health-related Quality of Life (QoL) we used the Short Form 36 Health Survey (SF-36) to assess physical and mental health.	The MOS 36-Item short-form health survey (SF-36). Conceptual framework and item selection	Ware, JE; Sherbourne, CD	Med Care
54	1512	2001	C	Bayesian phylogenetic analysis was performed using MrBayes 3.1.2.	MRBAYES: Bayesian inference of phylogenetic trees	Hueslenbeck, JP.; Ronquist, F	Bioinformatics
55	1506	1997	C	tRNA genes were identified using tRNAscan-SE.	tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence	Lowe, TM; Eddy, SR	Nucleic Acids Res
56	1499	1983	I	Depression will be measured using the Hospital Anxiety and Depression Scale (HADS).	The hospital anxiety and depression scale	Zigmond, AS; Snaith, RP	Acta Psychiatr Scand
57	1463	2002	C	Sequences were mapped to the mouse genome using BLAT.	The BLAST-like alignment tool	Kent, WJ	Genome Res
58	1437	2004	C	The molecular figures were produced by using UCSF Chimera.	UCSF Chimera-A visualization system for exploratory research and analysis	Pettersen, EF; Goddard, TD; Huang, CC; Couch, GS; Greenblatt, DM; Meng, EC	J. Comput. Chem.
59	1433	1987	C	Presence of comorbidity was quantified using the Charlson index of comorbidity.	A new method of classifying prognostic comorbidity in longitudinal studies: development and validation	Charlson, ME; Pompei, P; Ales, KL; MacKenzie, CR	J Chronic Dis
60	1414	2001	G	According to the National Cholesterol Education Program (NCEP) Expert Panel ATP-III-criteria (2001) the metabolic syndrome consist of 3 or more of 5 risk factors: hyperglycaemia, hypertension, decreased levels of HDL, elevated levels of triglycerides and abdominal (visceral) obesity.	Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III)	Scott, G; Diane, B; Luther, C; Richard, C; Margo, D; Wm, H; Donald, H; Illingworth, DR; Russell, L; Patrick, M; James, M; Richard, P; Neil, S; Linda, H	JAMA
61	1399	1986	C	The degree of agreement between the two methods is ascertained using a Bland-Altman or Tukey mean-difference plot.	Statistical methods for assessing agreement between two methods of clinical measurements	Bland, JM; Altman, DG	Lancet
62	1398	2003	C	An RNA secondary structure was predicted by using Mfold.	Mfold web server for nucleic acid folding and hybridization prediction	Zuker, M	Nucleic Acids Res
63	1395	2003	C	Data were normalized using quantile normalization.	A comparison of normalization methods for high density oligonucleotide array data based on variance and bias	Bolstad, BM; Irizarry, RA; Astrand, M; Speed, TP	Bioinformatics
64	1363	2002	D	The microarray data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE51980.	Gene Expression Omnibus: NCBI gene expression and hybridization array data repository	Edgar, R; Domrachev, M; Lash, AE	Nucleic Acids Res
65	1346	2007	C	Our analysis is based on genome-wide association study (GWAS) data for seven common diseases genotyped by the Wellcome Trust Case Control Consortium (WTCCC).	Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls	Burton, PR; Clayton, DG; Cardon, LR; Craddock, N; Deloukas, P	Nature

66	1338	2005	C	Functional annotation of the gene ontology (GO) terms was done using the BLAST2GO program.	Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research	Conesa, A; Götz, S; García-Gómez, JM; Terol, J; Talón, M; Robles, M	Bioinformatics
67	1330	1972	C	The low density lipoprotein cholesterol (LDL-cholesterol) was calculated using the Friedewald formula.	Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge	Friedewald, W T; Levy, R I; Fredrickson, D S	Clin Chem
68	1325	1998	C	To model sequence evolution, we employed the GTR + I + G model of nucleotide substitution, which was identified as the best-fitting model based on the Akaike Information Criterion (AIC) using ModelTest v3.8.	MODELTEST: testing the model of DNA substitution	Posada, D; Crandall, KA	Bioinformatics
69	1318	1983	B	Cell viability tests were performed using the MTT assay with the cell proliferation reagent MTT (3-[4,5-dimethylthiazol-2-yl]-2,5-diphenyl tetrazolium bromide; Sigma), as described previously.	Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays	Mossman, T	J Immunol Methods
70	1310	2007	C	Evolutionary dynamics were estimated using a Bayesian Markov chain Monte Carlo (MCMC) approach implemented in BEAST.	BEAST: bayesian evolutionary analysis by sampling trees	Drummond, A; Rambaut, A	BMC Evol Biol
71	1309	2010	D	Prostate cancer is the most common cancer in men and the second leading cause of cancer deaths in the United States.	Cancer statistics, 2010	Jemal, A; Siegel, R; Xu, J; Ward, E	CA Cancer J Clin
72	1304	2005	C	Genome sequencing was performed using 454 pyrosequencing.	Genome sequencing in microfabricated high-density picolitre reactors	Margulies, M; Egholm, M; Altman, WE	Nature
73	1302	1988	G	These RA patients fulfilled the American College of Rheumatology 1987 criteria for RA.	The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis	Arnett, FC; Edworthy, SM; Bloch, DA; McShane, DJ; Fries, JF; Cooper, NS; Healey, LA; Kaplan, SR; Liang, MH; Luthra, HS; Medsger, TA; Mitchell, DM; Neustadt, DH; Pinals, RS; Schaller, JG; Sharp, JT; Wilder, RL; Hunder, GG	Arthritis and Rheumatism
74	1299	2004	C	Sequence logos were generated using WebLogo.	WebLogo: a sequence logo generator	Crooks, GE; Hon, G; Chandonia, JM; Brenner, SE	Genome Res
75	1291	2010	D	Breast cancer is the most common type of cancer in women, and colon cancer is the third most common cancer in both sexes and the second leading cause of cancer deaths worldwide.	Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008	Ferlay, J; Shin, HR; Bray, F; Forman, D; Mathers, C; Parkin, DM	Int J Cancer
76	1283	2007	C	The aligned codon sequences were used to test positive selection using the branch-site model implemented in the program Codeml of PAML 4.4.	PAML 4: phylogenetic analysis by maximum likelihood	Yang, Z	Molecular Biology and Evolution
77	1278	2007	N	Human induced pluripotent stem (iPS) cells generated by reprogramming of somatic cells with four transcription factors (Oct3/4, Klf4, Sox2, and c-Myc) have properties similar to those of human embryonic stem cells.	Induction of pluripotent stem cells from adult human fibroblasts by defined factors	Takahashi, K; Tanabe, K; Ohnuki, M	Cell
78	1276	1999	N	Bone marrow stromal cells (BMSCs) include mesenchymal stem cells (MSCs), which are tissue stem cells able to differentiate into multiple cell types in mesenchymal tissues such as chondrocytes, osteoblasts, and adipocytes.	Multilineage potential of adult human mesenchymal stem cells	Pittenger, MF; Mackay, AM; Beck, SC; Jaiswal, RK; Douglas, R; Mosca, JD; Moorman, MA; Simonetti, DW; Craig, S; Marshak, DR	Science

79	1273	2000	N	Comprehensive gene expression profiling has identified five major molecular subtypes in breast cancer including luminal A, luminal B, HER2+, basal-like, and normal breast-like subtype.	Molecular portraits of human breast tumours	Perou, CM; Sørlie, T; Eisen, MB; van de Rijn, M; Jeffrey, SS; Rees, CA; Pollack, JR; Ross, DT; Johnsen, H; Akslen, LA; Fluge, O; Pergamenschikov, A; Williams, C; Zhu, SX; Lønning, PE; Børresen-Dale, AL; Brown, PO; Botstein, D	Nature
80	1265	2008	C	Annotation of the genome was performed using the RAST (Rapid Annotation using Subsystem Technology) server.	The RAST server: rapid annotations using subsystems technology	Aziz, RK; Bartels, D; Best, A A; Dejongh, M; Disz, T; Edwards, RA	BMC Genomics
81	1247	2000	C	Two protein sequences of each gene pair were aligned using the global sequence alignment program NEEDLE in the EMBOSS package.	EMBOSS: the European Molecular Biology Open Software Suite	Rice, P; Longden, I; Bleasby, A	Trends Genet
82	1230	2000	G	Tumor response was assessed according to the Response Evaluation Criteria in Solid Tumors (RECIST) criteria.	New guidelines to evaluate the response to treatment in solid tumours	Therase, P; Arbuck, SG; Eisenhauer, EA; Wanders, J; Kaplan, RS; Rubinstein, L	J Natl Cancer Inst
83	1228	2010	D	Our set of data genomes consists of 1000 genomes from the 1000 Genome project.	A map of human genome variation from population-scale sequencing	Genomes Project, C; Abecasis, GR; Altshuler, D; Auton, A; Brooks, LD; Durbin, RM	Nature
84	1226	2009	C	First, the reads are mapped to the reference genome using Tophat.	TopHat: discovering splice junctions with RNA-Seq	Trapnell, C; Pachter, L; Salzberg, SL	Bioinformatics
85	1225	1974	B	C. elegans strains were maintained at 20 °C on nematode growth medium (NGM) seeded with E. coli strain OP50 as described by Brenner.	The genetics of Caenorhabditis elegans	Brenner, S	Genetics
86	1221	2000	C	Primers were designed using the primer3 software.	Primer3 on the WWW for general users and for biologist programmers	Rozen, S; Skaletsky, H	Methods Mol Biol
87	1219	2003	C	Adjusted P values (i.e., Q value < 0.05) were applied to correct for multiple testing using the false discovery rate (FDR) method.	Statistical significance for genomewide studies	Storey, JD; Tibshirani, R	Proc Natl Acad Sci U S A.
88	1206	2009	C	We performed a gene ontology (GO) enrichment analysis for network modules using the Database for Annotation, Visualization and Integrated Discovery (DAVID).	DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways	Shao, Z; Zhao, H; Zhao, H	Nucleic Acids Res
89	1201	2012	D	Prostate cancer accounts for 29% of cancer incidence and 9% of cancer deaths and it is the most common cancer and the second leading cause of cancer death in American men in 2012.	Cancer Statistics 2012	Siegel, R; Naishadham, D; Jemal, A	CA Cancer J Clin
90	1198	2001	C	Transmembrane proteins were predicted using TMHMM.	Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes	Krogh, A; Larsson, B; von Heijne, G; Sonnhammer, EL	Journal of Molecular Biology
91	1192	2002	C	We calculated the Q statistic (P75% large or extreme heterogeneity) to assess heterogeneity across studies.	Quantifying heterogeneity in a meta-analysis	Higgins, JP; Thompson, SG	Statistics in Medicine
92	1176	2004	D	The worldwide prevalence of type 2 diabetes is increasing, and the global number of people with diabetes is estimated to reach 366 million by the year 2030.	Global prevalence of diabetes. Estimates for the year 2000 and projections for 2030	Wild, S; Roglic, G; Green, A; Sicree, R; King, H	Diabetes Care
93	1169	2002	C	Reference gene stability was determined by performing GeNorm analysis which is based on calculation of a reference gene-stability measure M.	Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes	Vandesompele, J; De Preter, K; Pattyn, F; Poppe, B; Van Roy, N; De Paepe, A	Genome Biol
94	1145	1987	B	Total RNA was extracted using the phenol-chloroform-guanidinium-thiocyanate method.	Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction	Chomczynski, P; Sacchi, N	Anal Biochem

95	1142	2000	B	E. coli CFT073 gene deletion mutants were constructed using the Red mediated homologous recombination system as previously described .	One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products	Datsenko, KA; Wanner, BL	Proc Natl Acad Sci USA
96	1107	2009	C	The number of haplotypes and nucleotide diversity were calculated using DnaSP software (version 5.10).	DnaSP v5: a software for comprehensive analysis of DNA polymorphism data	Librado, P; Rozas, J	Bioinformatics
97	1097	2003	C	Gene ontology (GO) analysis was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID).	DAVID: database for annotation, visualization, and integrated discovery	Dennis, G; Sherman, BT; Hosack, DA; Yang, J; Gao, W; Lane, HC; Lempicki, RA	Genome Biol
98	1090	2005	C	The most likely number of clusters was determined using the log-probability of the data (lnPr(X/K); and the method of Evanno et al.	Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study	Evanno, G; Regnaut, S; Goudet, J	Mol Ecol
99	1088	2007	C	The structure was solved by molecular replacement using PHASER.	Phaser crystallographic software	McCoy, AJ; Grosse-Kunstleve, RW; Adams, PD; Winn, MD; Storoni, LC; Read, RJ	J Appl Crystallogr
100	1077	2010	C	Differential expression analysis of genes was performed using the DESeq R package.	Differential expression analysis for sequence count data	Anders, S; Huber, W	Genome Biol

Appendix B. logistic regression for the percent using words variable

glm(formula = method ∼ using, family = binomial, data = tbl)
Deviance Residuals:

Min	1Q	Median	3Q	Max
−2.35071	−0.36255	0.00445	0.12538	2.60904

	Coefficients	Estimate Std. Error	z value	Pr(>|z|)
(Intercept)	−3.42372	0.23039	−14.86	<2e-16 ***
using	0.21600	0.01503	14.37	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.
(Dispersion parameter for binomial family taken to be 1).
Null deviance: 1375.46 on 999 ◦ of freedom.
Residual deviance: 489.15 on 998 ◦ of freedom.

Appendix C. Logistic model for four word variables and consensus

Logistic regression:
glm(formula = method ∼ using + may + cons + not + show, family = binomial, data = tbl)
Deviance Residuals:

Min	1Q	Median	3Q	Max
−2.15395	−0.27560	0.00323	0.08082	3.07973

	Coefficients	Estimate Std. Error	z value	Pr(>|z|)
(Intercept)	−5.51691	1.08446	−5.087	3.63e-07 ***
using	0.19103	0.01738	10.993	<2e-16 ***
may	−0.37446	0.11185	−3.348	0.000814 ***
cons	0.10668	0.02500	4.267	1.98e-05 ***
not	0.26754	0.07640	3.502	0.000462 ***
show	−0.29136	0.05798	−5.025	5.03e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.
(Dispersion parameter for binomial family taken to be 1).
Null deviance: 1375.46 on 999 ◦ of freedom.
Residual deviance: 378.55 on 994 ◦ of freedom.
AIC: 390.55.
Number of Fisher Scoring iterations: 8.
Pearson-Windmeijer goodness-of-fit (GOF) test.
p-value = 0.77 (higher values better).

References

Allison, P. D. (1999). *Logistic regression using SAS system: Theory and application*. Cary, NC: SAS Institute Inc.

Bertin, M., Atanassova, I., Gingras, Y., & Lariviére, V. (2016). The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology*, *67*(1), 164–177.

Bertin, M., Atanassova, I., Sugimoto, C., & Lariviére, V. (2016). The linguistic patterns and rhetorical structure of citation context: An approach using N-grams. *Scientometrics*, *109*(3), 1417–1434. http://dx.doi.org/10.1007/s11192-016-2134-8

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80.

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citation in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*, 59–73.

Chen, C., & Song, M. (2018). *Representing scientific knowledge: The role of uncertainty*. London, UK: Springer.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scale disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.

DiMarco, C., Kroon, F. W., & Mercer, R. E. (2006). Using hedges to classify citations in scientific articles. In J. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing attitude and affect in text: Theory and applications, vol. 20. The Information Retrieval Series* (pp. 247–263). Netherlands: Springer. http://dx.doi.org/10.1007/1-4020-4102-0_19

Duhem, P. (1962). *The aim and structure of physical theory*. New York: Atheneum.

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science*, *59*(1), 51–62. http://dx.doi.org/10.1002/asi.20707

Garfield, E. (1973). Citation frequency as a measure of research activity and performance. *Current Contents*, 5 (January 31, 1973). [Reprinted in: Garfield, E. *Essays of an Information Scientist*, 1962-73, Vol. 1 (pp. 406–408). Philadelphia, PA.: ISI Press].

Garfield, E. (1977). Highly cited articles. 39. Biochemistry papers published in the 1950. *Current Contents*, 25 (June 20, 1977), pp. 5–12. [Reprinted in: Garfield, E. (1977-78). *Essays of an Information Scientist*, Vol. 3 (pp. 147–154). Philadelphia, PA : ISI Press].

Garfield, E. (1990). The most-cited papers of all time, SCI 1945–1988. Part 1A. The SCI top 100–Will the Lowry method ever be obliterated? *Current Contents*, 7, pp. 3–14, February 12, 1990. [Reprinted in: *Essays of an information Scientist*, Vol. 13 (p. 45). Philadelphia, PA : ISI Press].

Garfield, E. (1991). The most-cited papers of all time, SCI 1945–1988, Part 4. The papers ranked 301–400. *Current Contents*, 21 (May 27, 1991), pp. 5–16. [Reprinted in: Garfield, E. (1991). *Essays of an Information Scientist*, Vol. 14 (p. 79). Philadelphia, PA : ISI Press].

Hanson, R. N. (1972). *Patterns of discovery*. Cambridge: Cambridge University Press.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: John Wiley & Sons.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor: The University of Michigan Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago: The University of Chicago Press.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*. Beverly Hills, CA: Sage Publications.

Latour, B. (1987). *Science in action*. Cambridge, MA: Harvard University Press.

Lowry, O. H. (1977). *Citation classic commentaries*. [1, January 3, 1977]. http://garfield.library.upenn.edu/classics1977/A1977DM02300001.pdf

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naïve Bayes text classification. *In proceedings of the AAAI/ICML-98 workshop on learning for text categorization*, 41–48.

Mercer, R. E., & DiMarco, C. (2003). The importance of fine-grained cue phrases in scientific citations. *Lecture Notes in Artificial Intelligence*, *2671*, 550–556.

Nakov, P., Schwartz, A., & Hearst, M. (2004). Citances: Citation sentences for semantic analysis of bioscience text. *SIGIR workshop of search and discovery on bioinformatics*.

Olby, R. (1974). *The path to the double helix*. Seattle: University of Washington Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Popper, K. R. (1961). *The logic of scientific discovery*. New York: Science Editions, Inc.

Price, D. J. (1983). Of sealing wax and string: A philosophy of the experimenters craft and its role in the genesis of high technology. In *Little science, big science and beyond*. pp. 237–253. New York: Columbia University Press.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Scott, M. (2004). *WordSmith tools version 4*. Oxford: Oxford University Press.

Small, H., & Griffith, B. C. (1974). The structure of scientific literatures 1: Identifying and graphing specialties. *Science Studies*, *4*(1), 17–40.

Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics*, *11*, 46–62.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science*, *8*, 327–340.

Small, H. (1986). The synthesis of specialty narratives from co-citation clusters. *Journal of the American Society for Information Science*, *37*(3), 97–110.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation functions. *EMNLP '06: Proceedings of the 2006 conference on empirical methods in natural language processing, association for computational linguistics, July, 2006*.

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers: Nature explores the most-cited research of all times. *Nature*, *514*(7524), 550–553. http://www.nature.com/news/the-top-100-papers-1.16224

Windmeijer, F. A. G. (1990). The asymptotic distribution of the sum of weighted squared residuals in binary choice models. *Statistica Neerlandica*, *44*(2), 69–78.

Wouters, P. (1999). *The citation culture. Amsterdam: University of Amsterdam. [Ph.D. dissertation]*. http://garfield.library.upenn.edu/wouters/wouters.pdf

Ziman, J. M. (1968). *Public knowledge: An essay concerning the social dimension of science*. Cambridge: Cambridge University Press.