Scientometrics
DOI 10.1007/s11192-016-2209-6

Striking similarities between publications from China describing single gene knockdown experiments in human cancer cell lines

Jennifer A. Byrne[1,2] · **Cyril Labbé**[3]

Received: 23 August 2016

Abstract Comparing 5 publications from China that described knockdowns of the human *TPD52L2* gene in human cancer cell lines identified unexpected similarities between these publications, flaws in experimental design, and mis-matches between some described experiments and the reported results. Following communications with journal editors, two of these *TPD52L2* publications have been retracted. One retraction notice stated that while the authors claimed that the data were original, the experiments had been out-sourced to a biotechnology company. Using search engine queries, automatic text-analysis, different similarity measures, and further visual inspection, we identified 48 examples of highly similar papers describing single gene knockdowns in 1–2 human cancer cell lines that were all published by investigators from China. The incorrect use of a particular *TPD52L2* shRNA sequence as a negative or non-targeting control was identified in 30/48 (63%) of these publications, using a combination of Google Scholar searches and visual inspection. Overall, these results suggest that some publications describing the effects of single gene knockdowns in human cancer cell lines may include the results of experiments that were not performed by the authors. This has serious implications for the validity of such results, and for their application in future research.

Keywords Gene knockdown · Cancer · Cell lines · Publications · Intertextual distance · China

✉ Jennifer A. Byrne
jennifer.byrne@health.nsw.gov.au

✉ Cyril Labbé
cyril.labbe@imag.fr

[1] Molecular Oncology Laboratory, Children's Cancer Research Unit, Kids Research Institute, The Children's Hospital at Westmead, Locked Bag 4001, Westmead, NSW 2145, Australia

[2] The University of Sydney Discipline of Child and Adolescent Health, The Children's Hospital at Westmead, Locked Bag 4001, Westmead, NSW 2145, Australia

[3] University of Grenoble Alpes, Laboratoire LIG - Bâtiment IMAG - 700 Avenue Centrale, Domaine Universitaire de Saint-Martin-d'Hères, 38058 Grenoble Cedex 9, France

Published online: 28 December 2016

Introduction

Scientific research progresses based upon results described within the peer-reviewed literature. The integrity of scientific publishing is of paramount importance, to ensure future progress, avoid waste, and to maintain stakeholder and public support for scientific research and the use of evidence in broader decision-making (Bik et al. 2016; Bowen and Casadevall 2015; Kreutzberg 2004). Many threats to the integrity of scientific publishing have been identified, which can be broadly categorised according to perceived intent (Bornmann 2013; Fanelli 2009; Smith 2006). Flaws in experimental design, inadequate or incorrect use of standards, incorrect statistical analyses and plagiarism of text may commonly arise through unintentional error (Casadevall et al. 2014). Deliberate attempts to deceive include plagiarism (of text and/or data), duplicate publication, data manipulation or omission and in extreme cases, data or study fabrication (Delgado López-Cózar et al. 2014; Fanelli 2009; Lesk 2015; Smith 2006). More recent efforts to manipulate scientific publishing include selling of manuscripts by third parties to authors in China (Hvistendahl 2013), and so-called peer review scams, where authors propose peer-reviewer email addresses that are either linked to colleagues or themselves (Ferguson et al. 2014). A major driver of scientific fraud is considered to be the "publish or perish" culture that exists in highly competitive environments (van Dalen and Henkens 2012). This is supported by statistical analyses of journal impact factors of retracted articles due to suspected or admitted fraud versus the journal impact factors of articles retracted due to error (Fang et al. 2012; Steen 2011a). However, peer review scams have involved lower-impact journals (Ferguson et al. 2014), possibly due to perceptions of lower peer-review and/or editorial standards.

Although more publications are likely to be flawed through unintentional error than scientific fraud (Steen 2011b), there has been extensive discussion of scientific fraud in the research literature. This is in part because known instances of scientific fraud are likely to represent the tip of a much larger iceberg (Fanelli 2009). Scientific fraud is likely to be actively concealed, and therefore difficult to detect, and may be disguised in retraction notices as inadvertent error (Casadevall et al. 2014). Manipulated or fabricated studies can also be produced more rapidly and/or in greater numbers than genuine studies, and can significantly alter the trajectory of future investigations (White 2005). There is therefore serious and likely underestimated potential for some forms of scientific fraud to misdirect future research efforts (Moore et al. 2010).

Given the damaging effects of scientific misconduct, the detection of actively flawed publications is a major area of investigation. Automated or semi-automated tools are being developed to identify various types of misconduct in science. Some tools aim to identify automatically-generated publications (Amancio 2015; Fahrenberg et al. 2014; Labbé and Labbé 2013), and have been adopted by publishers [Springer Nature with SciDetect[1] as well as the open archive ArXiv (Ginsparg 2014)]. The detection of plagiarism or hidden "intertextuality" between publications is another very active track (Citron and Ginsparg 2015; Labbé and Labbé 2012). Whereas commercial tools are widely employed and efficient for detecting raw and direct plagiarism, targeted rewriting may be sufficient for plagiarism to go undetected. To identify such forms of plagiarism, more complex techniques are being developed to analyse the structure of texts and stylometry characteristics (Amancio et al. 2012; Ausloos et al. 2016; Carpena et al. 2009; Mehri et al. 2012). To highlight errors or dubious publications, one can also employ automatic approaches to

[1] http://scidetect.forge.imag.fr.

check the statistical validity of presented values (Nuijten et al. 2015).[2] It has even been claimed that pseudoscientific theories could be automatically detected using machine learning techniques (Shvets 2014). While mining tools are clearly valuable in their ability to identify or highlight questionable or off-track publications, such tools may also be limited by the science chosen to feed the learning phase. An alternative approach is to highlight errors and misconduct through the power of the crowd, using collaborative websites to promote post-publication peer review by individuals (PubPeer).[3] There is also scope to combine individual and automated approaches, in order to extend the observations or concerns of individual researchers to more comprehensively describe broader phenomena.

The aim of this study was to highlight a set of questionable papers/practices by explaining why these papers are questionable and how they were identified. Through reading the literature, one of us (JAB) identified 5 studies published between 2014 and 2015 that commonly described *TPD52L2* knockdowns performed in 1–2 human cancer cell lines representing different cancer types (Wang et al. 2014; He et al. 2015; Pan et al. 2015; Xu et al. 2015; Yang et al. 2015). The *TPD52L2* gene was first reported by JAB and colleagues (Nourse et al. 1998) and is a member of the *TPD52* gene family (Byrne et al. 2014). Over the past 18 years (1998–2016), only 19 publications in PubMed have referred to *TPD52L2* (and/or a recognised synonymous gene identifier) in the title and/or abstract, and no published study had targeted *TPD52L2* using gene knockdown approaches prior to 2014.

As 5 similar *TPD52L2* publications would not have been expected to have been published in less than 1 year, these were examined and compared in detail. As will be described, visual inspection combined with nucleotide database homology searches identified both striking similarities between the individual publications, fundamental yet apparently undetected flaws in experimental design, and implausible mis-matches between some experiments described and the results shown. Using search engine queries, automatic text-analysis, different similarity measures, and further visual inspection, we identified examples of other highly similar papers published by investigators from China that described the effects of single gene knockdown in human cancer cell lines. When combined with a recent description of the "publish or perish" culture at some Chinese universities (Tian et al. 2016) and the retraction of one *TPD52L2* publication (Pan et al. 2015; Retraction 2016), these results predict that some authors in China may be obtaining data and/or figures from sources such as biotechnology or education companies, and publishing these results without disclosure. This has serious implications for the validity of these results.

Methods

Identification and analysis of index corpus

Google Scholar alerts and PubMed searches were employed to identify index cases (the index corpus, Table 1). Visual inspection of pdf versions of the index corpus focussed upon the results described and their presentation in figures. Homology searches using nucleotide sequences from index papers as queries were performed using Megablast [short

[2] http://CRAN.R-project.org/package=statcheck.

[3] http://pubpeer.com.

Table 1 Descriptions of analysed corpora and numbers of common papers between corpora (diagonal size of the corpus)

Corpus name and description	Index $n = 5$	IJCEM $n = 4094$	Reference $n = 15$	PubMed $n = 88$[a]	Google Scholar sequence A $n = 26$[a]	Google Scholar sequence D $n = 4$	Identified $n = 48$
Index corpus, $n = 5$ papers, all describing *TPD52L2* knockdown in human cancer cell lines	5	1	5	5	2	1	5
IJCEM corpus, $n = 4094$ papers, used as a representative sample of publications in the field		4094	2	1	1	0	2
Reference corpus, $n = 15$ papers, includes index corpus + 10 other highly similar papers identified using PubMed keyword searches and visual inspection			15	9	2	1	15
PubMed corpus, $n = 88$ papers identified by PubMed as "similar" to one index paper (Yang et al. 2015)				88	4	3	18
Google Scholar sequence A corpus, $n = 26$ papers, identified by Google Scholar to contain sequence A					26	0	26
Google Scholar sequence D corpus, $n = 4$ papers, identified by Google Scholar to contain sequence D						4	4
Identified corpus, $n = 48$ papers, includes reference corpus, Google Scholar corpora, $n = 18$ papers from PubMed corpus							48

[a] Numbers of papers represent numbers that were available for download and analysis

hairpin RNA (shRNA) sequences] and/or Blastn [shRNA and reverse transcription polymerase chain reaction (RT-PCR) primer sequences]. During this analysis period, JAB contacted the editors of each of the 4 journals that had published the 5 index papers.

Identification of similar publications

Reference corpus

Initial PubMed searches were performed using different combinations of key words that retrieved index publications ("lentivirus" + "knockdown" ± "mediated" ± "proliferation" ± "cancer"), with and without the gene identifier *NOB1*, which was relevant to one short hairpin RNA (shRNA) sequence described in 2/5 index papers. A subset of the papers identified was visually inspected as described above. Ten papers that showed a high degree of similarity to the index corpus were combined with index cases to form a reference corpus (Table 1).

PubMed corpus

PubMed "similar" searches were performed to identify papers that PubMed recommended as similar to one index paper (Yang et al. 2015). The PubMed "similar" search[4] is based on a similarity function that differently weights words from the abstract, title and MeSH words (Medical Subject Headings). When searching for papers "similar" to Yang et al. (2015), the retrieved set may contain papers that are similar or quite dissimilar, as measured using intertextual distance (see below). Nevertheless, a PubMed "similar" search can be viewed as a first step to identify a set of papers with a greater chance of being similar (within the meaning of intertextual distance) to members of the index corpus. Of the 112 English-language publications identified, 88 publications were downloaded as the PubMed corpus (Table 1). The 24 remaining papers were inaccessible (behind a paywall), and were not analysed further. Publications from open-access publishers may therefore be over-represented in the set of downloaded publications and thus in any subsequent set.

IJCEM corpus

As a testbed to delineate thresholds for further comparisons, 4094 publications were downloaded from the open-access journal International Journal of Clinical and Experimental Medicine (IJCEM), these being all publications from volume 1 number 2 (2008) to volume 9, number 1 (2016). The IJCEM corpus was used a representative sample of publications in the field, as one index publication (Pan et al. 2015) was published in the IJCEM (Table 1).

Intertextual distance analysis

Intertextual distance has been used previously to detect publications produced by the SCIgen computer program (Labbé and Labbé 2013) and is the basis of the SciDetect software[5] currently used by Springer Nature to discover text generated by SCIgen and other fake-paper generators such as Mathgen.[6] The intertextual distance between two texts measures the proportion of word-tokens (strings of alphanumeric characters separated by spaces or punctuation) shared by the two texts. For computing intertextual distances, pdf files were

[4] More information of the PubMed similar function: https://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Computation_of_Similar_Articles.

[5] http://scidetect.forge.imag.fr.

[6] http://thatsmathematics.com/mathgen/.

converted to plain text files using pdftotxt software (part of the Xpdf software suite).[7] During this operation, footers and headers were retained, but figures, graphs and formulae were lost, but often left traces in the extracted text in the form of isolated sequences of words, letters or numbers. As an example of the sizes of texts analysed, reference publications ($n = 15$) contained a mean number of 3384 words per article (range 2973–5223 words), and the total word size of the reference corpus was 50,763. Raw texts were then segmented into word-tokens using the procedure of the Oxford Concordance Program (Hockey and Martin 1988). Stop-words were not removed, as it has been shown that stop-words play an important role in identifying sources and topics (Argamon and Levitan 2005; Ginsparg 2014; Labbé and Labbé 2013; Stamatatos 2009; Tuzzi 2010). Intertextual distance can be interpreted as follows: after randomly choosing 100 words in each text, the δ distance is the expected proportion of common words between the two sets of 100 words. A 0 distance is reached when the same vocabulary is used in the two texts at the exact same frequency. A distance equal to 1 is achieved when the two texts share no words.

Intertextual distances were computed between the 4094 publications of the IJCEM corpus, and the 15 publications of the reference corpus. For the IJCEM corpus, the observed mean (and median) intertextual distance was 0.65 with a standard deviation of 0.039. Only 1% of the observed distances were <0.55, and 0.25% of the observed distances were <0.51 (Fig. 1). It was therefore inferred that an intertextual distance <0.55 is very unusual. In contrast, the mean intertextual distance between the 15 papers of the reference corpus was 0.44 (range 0.34–0.51).

These results were used to define thresholds for the analysis of the PubMed corpus. The adopted approach can be related to a nearest-neighbour classification (Cover and Hart 1967) with similarity thresholds to fix class boundaries. Here, if the intertextual distance between a member of the PubMed corpus and its nearest neighbour within the reference corpus was <0.44 (the mean reference corpus distance), this flagged the paper as being "possibly related to reference papers". A distance of 0.44–0.55 (the most similar 1% of distances in the IJCEM corpus) flagged the paper as having "elements related to reference papers". No decision was taken for papers with distances greater than 0.55. Papers with intertextual distances ≤0.55 were visually examined, and as a result, papers with intertextual distances of <0.50 were subsequently visually inspected for hallmarks of the index corpus.

Google Scholar searches

We performed Google Scholar searches to identify other possible instances of shRNA sequences described in the index corpus within the published literature (Google Scholar corpora, Table 1). Intertextual distance analyses were then performed as described for the PubMed corpus. Additional Blastn searches were also performed using shRNA and RT-PCR primer sequences contained within publications within the Google Scholar and PubMed corpora.

Results

Visual analysis of index corpus

Google Scholar alerts and PubMed searches identified 5 *TPD52L2* studies (the index corpus) published in 4 different journals between 2014 and 2015, with no apparent overlap

[7] http://www.foolabs.com/xpdf/home.html.

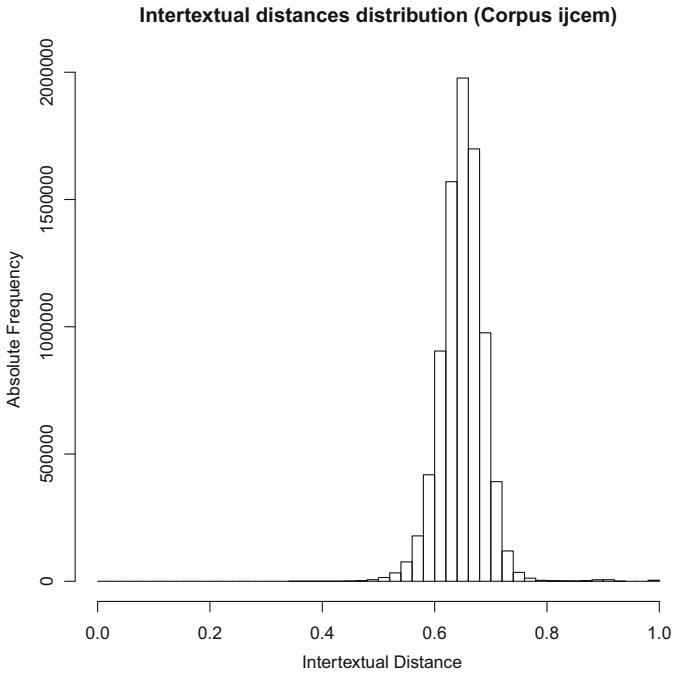

Fig. 1 Histogram showing the absolute frequency of intertextual distances between publications of the IJCEM corpus, representing 4094 papers downloaded from the International Journal of Clinical and Experimental Medicine (publications from volume 1 number 2, 2008 to volume 9, number 1, 2016). The Y axis shows the absolute frequency at which intertextual distance values (shown on the X axis) were observed

in authorship (He et al. 2015; Pan et al. 2015; Wang et al. 2014; Xu et al. 2015; Yang et al. 2015). The articles' stated aims were to perform gene knockdowns of *TPD52L2* in cancer cell line(s) corresponding to different cancer types (breast cancer, gastric cancer, glioma, liver cancer, oral squamous cell carcinoma) (Table 2). Most publications performed gene knockdowns in a single human cancer cell line, although He et al. (2015) repeated some experiments in a second cell line (Table 2). The 5 publications reported successful *TPD52L2* knockdown at the transcript and protein level using lentiviral shRNA vectors, resulting in reduced cell proliferation and colony formation, and altered cell cycle profiles (Table 2). Most studies found that *TPD52L2* knockdown resulted in increased percentages of cells in sub-G1 and G1/G0 phases, and reduced percentages of cells in S phase and G2/M phases (Table 2). Each study concluded that *TPD52L2* is an important gene to study further in the context of cancer, and that *TPD52L2* could have future therapeutic relevance.

More detailed visual analysis revealed other similarities between index publications, in terms of the methods employed and how the results were displayed in individual figures (Tables 3, 4). In an early figure, all index publications included both bright-field and corresponding green fluorescent protein (GFP) images of shRNA-transfected cells, and confirmed *TPD52L2* knockdown at the transcript and protein level (Table 3). In the next figure, cell proliferation results were shown in a graph, and the results of colony formation assays were shown as a histogram (Table 3). As part of the same figure, images were also commonly shown of cell colonies on a six-well plate, and of crystal violet staining, as well as bright field and corresponding fluorescent images of individual cell colonies (Table 3). A subsequent figure showed cell cycle profiles and histograms (Table 3). These

Table 2 Summary of results reported in *TPD52L2* knock-down (KD) studies (index corpus)

Study, cancer type, cell line(s) examined	PubMed ID	*TPD52L2* KD at transcript level	TPD52L2 KD at protein level	Cell proliferation post *TPD52L2* KD	Cell colony formation post *TPD52L2* KD	Cell % in G1/G0 phases post *TPD52L2* KD	Cell % in S phase post *TPD52L2* KD	Cell % in G2/M phases post *TPD52L2* KD	Sub-G1 cell % post *TPD52L2* KD	Other results shown
[a]He et al. (2015) Oral, CAL27, KB	25262828	√	√	↓	↓	↓	↑	↔	↑	↑cleaved PARP detection post-*TPD52L2* KD, TPD52L2 KD using second *TPD52L2* shRNA produced ↓cell proliferation in CAL27 and KB cells
[a]Pan et al. (2015) Liver, SMMC-7721	25932170	√	√	↓	↓	↑	↓	↔	↑	ND
Xu et al. (2015) Gastric, MGC80-3	25746840	√	√	↓	↓	↑	↓	↓	↑	TPD52L2 detection in 4 gastric cancer cell lines, ↑cleaved PARP detection post-*TPD52L2* KD
Yang et al. (2015) Breast, ZR-73-30	24842331	√	√	↓	↓	↑	↓	↓	ND	*TPD52L2* detection in 5 breast cancer cell lines, Pathscan arrays, Rescue shRNA experiment
Wang et al. (2014) Glioma, U251	25629696	√	√	↓	↓	↑	↓	↓	ND	*TPD52L2* KD using second *TPD52L2* shRNA, ↓proliferation post *TPD52L2* KD

ND not done

[a] Article has since been retracted

Table 3 Summary of figures showing experimental results in *TPD52L2* knockdown (KD) studies (index corpus)

Study, cancer type, PubMed ID	Cell images post shRNA transfection	Histogram showing *TPD52L2* KD at transcript level	Immunoblot showing TPD52L2 KD at protein level	↓Cell proliferation post *TPD52L2* KD	Cell images post *TPD52L2* KD	↓Cell colony formation post *TPD52L2* KD	FACS cell cycle plots post *TPD52L2* KD	Cell cycle histograms post *TPD52L2* KD	Cell cycle histograms, ↑sub-G1 population post *TPD52L2* KD	Other results shown
[a]He et al. (2015) Oral, 25262828	Figure 1A	Figure 1B	Figure 1C	Figure 2A	Figure 2B	Figure 2C	Figure 3A	Figure 3B	Figure 3C	Figure 4 Cleaved PARP detection Figure S1 TPD52L2 KD in CAL27 and KB cells
[a]Pan et al. (2015) Liver, 25932170	Figure 1A	Figure 1B	Figure 1C	Figure 2A	Figure 2B	Figure 2C	Figure 3A	Figure 3B	Figure 3C	ND
Xu et al. (2015) Gastric, 25746840	Figure 1B	Figure 1C	Figure 1D	Figure 2A	Figure 2B, 2C	Figure 2D	Figure 3A	Figure 3B	Figure 3C	Figure 1A TPD52L2 detection using immunoblot Figure 3D Cleaved PARP detection

Table 3 continued

Study, cancer type, PubMed ID	Cell images post shRNA transfection	Histogram showing *TPD52L2* KD at transcript level	Immunoblot showing TPD52L2 KD at protein level	↓Cell proliferation post *TPD52L2* KD	Cell images post *TPD52L2* KD	↓Cell colony formation post *TPD52L2* KD	FACS cell cycle plots post *TPD52L2* KD	Cell cycle histograms post *TPD52L2* KD	Cell cycle histograms, ↑sub-G1 population post *TPD52L2* KD	Other results shown
Yang et al. (2015) Breast, 24842331	Figure 2A	Figure 2B	Figure 2C	Figure 3A	Figure 3B	Figure 3C	Figure 4A	Figure 4B	ND	Figure 1 qRT-PCR detection of *TPD52L2* Figure 5 Pathscan arrays Figure S1 Rescue shRNA experiment
Wang et al. (2014) Glioma, 25629696	Figure 1A	Figures 1B, S1A	Figure 1C	Figures 2A, S1B	Figure 2B	Figure 2C	Figure 3A	Figure 3B	ND	Figure S1 *TPD52L2* KD using second *TPD52L2* shRNA

ND not done

[a] Article has since been retracted

Table 4 shRNA identities according to Blastn analyses and use of shRNAs in *TPD52L2* knockdown studies (index corpus)

Study Cancer type examined	PubMed ID	Sequence A 5′-GCGG… *TPD52L2* targeting	Sequence B 5′-CCGG… *TPD52L2* targeting	Sequence C 5′-CCAT… *LOC105373896* ncRNA, chr 2	Sequence D 5′-CTAG… *NOB1* targeting	Sequence E 5′-TTCT… *LOC105370714* ncRNA, chr 15	Sequence F 5′-CTCT… *TPD52L2* targeting
[a]He et al. (2015) Oral	25262828	*TPD52L2* targeting			**Non-targeting control**		*TPD52L2* targeting
[a]Pan et al. (2015) Liver	25932170	*TPD52L2* targeting **Non-targeting[b] control**					
Xu et al. (2015) Gastric	25746840	*TPD52L2* targeting			**Non-targeting control**		
Yang et al. (2015) Breast	24842331	**Non-targeting control**	*TPD52L2* targeting	***TPD52L2* rescue**			
Wang et al. (2014) Glioma	25629696	*TPD52L2* targeting				Non-targeting control	*TPD52L2* targeting

[a] Article has since been retracted

[b] Incorrect use of shRNA's is highlighted in bold text

experiments were shown in the same sequence in the figures of each index publication (Table 3). In 4/5 index publications, figures were annotated using a bold font similar to Times New Roman (He et al. 2015; Wang et al. 2014; Xu et al. 2015; Yang et al. 2015).

Blastn analyses of nucleotide sequences within index corpus

As all index publications mentioned the use of lentiviral constructs to perform gene knockdowns, the 6 shRNA sequences described were examined in detail (Table 4). Initial Megablast homology searches failed to identify any homology between shRNA sequences and sequences in the non-redundant nucleotide database, but subsequent Blastn searches identified homology between the 6 shRNA sequences and human sequences (Table 4). All index publications included an shRNA sequence 5′-GCG GAG GGT TTG AAA GAA TAT CTC GAG ATA TTC TTT CAA ACC CTC CGC TTT TTT-3′ (henceforth referred to as sequence A, Table 4). Three index publications employed sequence A as a *TPD52L2*-targeting shRNA, and paired this with "scrambled" shRNA's which were identified to either contain sequences homologous to the *NIN1/PSMD8 Binding Protein 1 Homologue* or *NOB1* gene (sequence D, Table 4) (He et al. 2015; Xu et al. 2015), or *LOC105370714* on chromosome 15 (sequence E, Table 4) (Wang et al. 2014). Pan et al. (2015) employed sequence A as both the *TPD52L2*-targetting and "negative siRNA" control, whereas Yang et al. (2015) employed sequence A as a non-targeting control, and an alternative *TPD52L2*-targeting shRNA (sequence B) with homology to *TPD52L2* (Table 4). Issues with shRNA knockdown and "rescue" experiments performed are briefly summarised in Table 4. In addition, RT-PCR primers with 100% homology to *Anillin* (*ANLN*, Table 5) were used to amplify *TPD52L2* transcripts by Yang et al. (2015). Thus despite the use of the same (Pan et al. 2015) or different (Yang et al. 2015) *TPD52L2* shRNA sequence(s) as both targeting and negative controls, or despite comparing *TPD52L2* knockdowns with those achieved using *NOB1* shRNA instead of non-targeting controls (He et al. 2015; Xu et al. 2015) (Table 4), all index studies reported significant effects of knocking down *TPD52L2* (Table 2).

Retractions

One article (He et al. 2015) was retracted in March 2016, by agreement between the authors, editor and publisher, on account of the use of an incorrectly identified cell line for experiments shown within the supplemental data (He et al. 2016). This cell line mis-identification had been previously reported in 2014 by Dr Amanda Capes-Davis, a recognised expert in cell line authentication (Capes-Davis and Neve 2016), using the Comment function of PubMed. The published retraction notice incorrectly stated the nature of the cell line mis-identification (He et al. 2016), and did not reference concerns raised by JAB. A second article (Pan et al. 2015) was subsequently retracted by the International Journal of Clinical and Experimental Medicine in 2016 following an editorial decision (Retraction 2016). The retraction notice mentioned the existence of other similar articles, and that the experiments of Pan et al. (2015) had been out-sourced to a biotechnology company (Retraction 2016).

Identification of similar publications

Given the incorrect use of a *NOB1* shRNA as a negative control (He et al. 2015; Xu et al. 2015), despite no obvious link between NOB1 and TPD52L2 gene or protein function, we hypothesised that other publications with similar figures and erroneous use of shRNA

Table 5 Summary of characteristics of the 48 articles within the identified corpus

Human gene targeted using shRNA or siRNA	Cancer cell line type studied	Similarity using SCIdetect	Google Scholar or text analysis identifies *TPD52L2* shRNA sequence A used as control?	Core figures from index corpus?	Times New Roman font used within figures?	Other
ADRBK1	Breast	0.40	Google Scholar	Yes	Yes	
ANLN	Breast	0.41	No	Yes	Yes	Google Scholar Sequence D (*NOB1*) used as control
ASNS	Breast	0.40	No	Yes	Yes	
CDK8	Breast	0.36	No	Some elements	No	
CEP55	Breast	Reference[a]	Text analysis	Yes	Yes	
eIF3d	Lung	0.41	Google Scholar	Yes	Yes	
GPR137	Pancreatic	0.40	No	Yes	Yes	Google Scholar Sequence D (*NOB1*) used as control
	Bladder	0.42	Google Scholar	Yes	Yes	
	Colon	0.42	No	Yes	Yes	shRNA very similar to sequence D (*NOB1*) used as control
	Medulloblastoma	0.43	Google Scholar	Yes	Most elements	
HNRNPA1	Lung	0.45	Google Scholar	Most elements	Yes	
ICT1	Glioblastoma	0.45	Google Scholar	Yes	No	
IL1R2	Osteosarcoma	0.43	Google Scholar	Yes	Yes	
Long Noncoding RNA KIAA0125	Gallbladder	0.49	Google Scholar	Some elements	No	
Long Non-coding RNA Linc-ITGB1	Gallbladder	0.43	Google Scholar	Some elements	Some elements	
	Breast	0.46	Google Scholar	Some elements	No	

Table 5 continued

Human gene targeted using shRNA or siRNA	Cancer cell line type studied	Similarity using SCIdetect	Google Scholar or text analysis identifies *TPD52L2* shRNA sequence A used as control?	Core figures from index corpus?	Times New Roman font used within figures?	Other
MPP8	Colon	0.39	Google Scholar	Yes	Yes	
	Thyroid	0.46	Google Scholar	Yes	Yes	
MYO6	Glioma	0.41	Google Scholar	Yes	Yes	
	Lung	0.44	Google Scholar	Yes	Yes	
	Liver	0.43	Google Scholar	Yes	Yes	
	Colorectal	0.42	Google Scholar	Yes	Yes	
NOB1	Glioma	Reference	No	Yes	Yes	One *NOB1* siRNA sequence included within sequence D
	Breast	Reference	No	Yes	No	*NOB1* shRNA targetting sequence includes Sequence E
	Prostate	Reference	Unknown	Most elements	Yes	shRNA sequences not provided
	Liver	Reference	No	Most elements	Yes	
	Ovarian	Reference	No	Most elements	No	
	Osteosarcoma	Reference	No	Yes	Yes	*NOB1* targetting shRNA predicted to target *PNO1*, not *NOB1*
PDLIM5	Gastric	0.38	No	Most elements	Yes	Google Scholar Sequence D (*NOB1*) used as control
PP5	Colorectal	0.42	Google Scholar	Yes	Yes	
PPM1D	Lung	0.42	No	Yes	Yes	
PTGR1	Gastric	0.42	Google Scholar	Yes	No	
PPP4R1	Breast	0.39	Text analysis	Yes	No	

Table 5 continued

Human gene targeted using shRNA or siRNA	Cancer cell line type studied	Similarity using SCIdetect	Google Scholar or text analysis identifies *TPD52L2* shRNA sequence A used as control?	Core figures from index corpus?	Times New Roman font used within figures?	Other
RPS15A	Lung	0.44	Google Scholar	Yes	Yes	
	Glioblastoma	0.43	Google Scholar	Yes	Yes	
TCTN1	Medulloblastoma	Reference	Text analysis	Yes	Yes	Sequence A missing most 3′ nucleotide
	Glioma	Reference	No	Yes	Yes	
	Pancreatic	Reference	Text analysis	Yes	Yes	
TPD52L2	Oral	Index[b]	No	Index	Yes	Google Scholar identified sequences A, D, used as targeting and control sequences, respectively
	Liver	Index	Google Scholar	Index	No	Sequence A used as both targeting and control sequence
	Gastric	Index	No	Index	Yes	Sequence D (*NOB1*) identified by text analysis, used as control
	Breast	Index	Text analysis	Index	Yes	
	Glioma	Index	No	Index	Yes	
TPTE2P1	Gallbladder	0.45	Google Scholar	Some elements	Yes	
USP39	Thyroid	0.44	Google Scholar	Some elements	Yes	
	Liver	0.42	Google Scholar	Yes	Yes	
ZFR	Pancreatic	0.41	Google Scholar	Yes	Yes	
ZFX	Breast	0.41	No	Some elements	Yes	

[a] Article from reference corpus

[b] Article from index corpus

sequences may exist within the literature. PubMed searches were performed as described in the Methods to identify 10 other similar publications (Tables 1, 5). These publications all examined single genes [*NOB1*, *tectonic family member 1 (TCTN1)*, or *centrosomal protein of 55 kDa (CEP55)*] using gene knockdown approaches in human cancer cell lines, included a similar series of figures, and were published by authors from China (Table 5). Seven (70%) of these papers (Table 5) included all core figures that characterised index cases (Table 3). In addition, 3/10 publications employed the *TPD52L2* shRNA sequence A as a non-targeting control (Table 5), as did 2/5 index publications (Table 4).

As described by in the Methods, intertextual distance analysis was performed to define thresholds for the analysis of the PubMed and other corpora. As a result, publications with intertextual distances of <0.50 were subjected to detailed visual inspection. This identified 18 publications from the PubMed corpus ["similar" to Yang et al. (2015)] that included the index corpus and 4 additional reference publications (Table 1). Nine additional papers were also published by groups of authors from China, reported the consequences of knocking down single genes in 1–2 human cancer cell lines, and included some or all of the data elements common to the index publications (Tables 3, 5).

Google Scholar searches

Google Scholar searches were performed to identify other instances of the shRNA sequences A–F (Table 4) within the published literature. These searches identified that sequence E, used as a non-targeting control by Wang et al. (2014) (Table 4), has been widely used as a non-targeting control, so this sequence was not analysed further. Sequences B, C and F were not identified in any publications beyond index publications (Table 4).

Google Scholar searches identified sequence A (*TPD52L2* shRNA) in 28 publications, which included 2/5 index publications (He et al. 2015; Pan et al. 2015) (Table 4). Two of these 28 publications were behind a paywall, and could not be analysed. In each of the other 24 publications examining a total of 17 different human genes, the *TPD52L2* shRNA sequence A was used as a negative/non-targeting control (Table 5). Visual inspection followed by Blastn analyses identified additional publications where sequence A was employed as a non-targeting shRNA (Table 5). In total, 30/48 (63%) of the identified cohort may have incorrectly used the *TPD52L2* shRNA sequence A as a non-targeting control (Tables 4, 5). Sequence D (*NOB1* shRNA) was also incorrectly employed as a negative control in 2/5 index publications (He et al. 2015; Xu et al. 2015). Google Scholar searches identified sequence D in one index publication (He et al. 2015) (Tables 4, 5), and in 3 additional publications examining 3 other genes (Table 5). In all these cases, sequence D was employed as a non-targeting control (Table 5). All papers within Google Scholar corpora were published by groups of authors from China, described single gene knock-downs in human cancer cell lines, included some or all data elements shared by index publications, and showed intertextual distances <0.5 (Fig. 2; Table 5).

Discussion

Summary of study findings and their potential significance

We have used a number of approaches to identify 48 examples of publications from China that show stringent similarities in terms of topic, text and data presentation, despite the fact

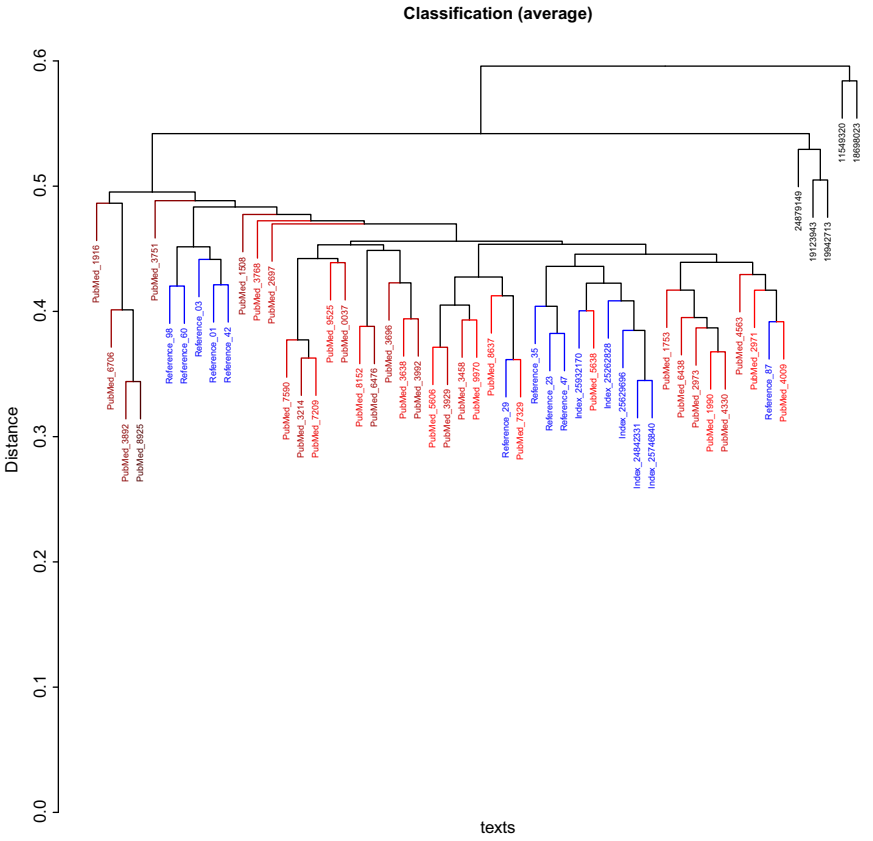

Fig. 2 Dendrogram for cluster analysis of the identified corpus ($n = 48$), including the 15 reference publications (*coloured blue*). Degrees of *red colour* are used to identify those publications that were closest to reference papers. *Black* denotes 5 publications with intertextual distances of $\geq$0.5 that show overlap with either the topics and/or techniques employed by the 15 reference papers. PubMed identifiers are shown for the 5 index publications, and for publications with intertextual distances of $\geq$0.5 (Wilson et al. 2001; Shehata et al. 2008; Verma et al. 2009; Konwisorz et al. 2010; Boyer-Guittaut et al. 2014). Truncated PubMed identifiers are shown for all other publications

that they are authored by individuals with little overlap or apparent relationship. We now consider how the relatively rapid appearance of such highly similar publications may have occurred. It is possible that plagiarism could account for these similar studies (Tables 3, 5), and from the degree of text overlap between studies, text plagiarism may have taken place in some cases. However, a characteristic feature of these papers is the inclusion of similarly formatted and presented figures, which often include a font similar to Times New Roman used in bold (Table 5). It would not be expected that figures deriving from independent laboratories should have such a strikingly similar appearance, particularly when published in different journals. The very similar appearance of figures, combined with inconsistencies between data descriptions versus data shown in figures, suggest some degree of uncoupling between the production of figures versus text. This could be consistent with the use of external suppliers of research data.

All of the similar publications that we identified feature authorship teams from China. While the assessment of publication numbers as opposed to quality occurs in many countries, sometimes with undesired results (Anderson et al. 2007; Butler 2003; Djuric 2015; van Dalen and Henkens 2012), it has been recognised that Chinese doctors and researchers are under particularly stringent pressure to publish (Hvistendahl 2013; Lin 2013; Tian et al. 2016; Ye et al. 2013; Zeng and Resnik 2010). This can take the form of being required to publish a quota of articles per year, while also being required to teach and fulfil other requirements (Tian et al. 2016). Recent interviews with de-identified young Chinese academics identified a number of strategies that academics were using to meet publication quotas (Tian et al. 2016). These included working very long hours, which has been independently supported (Wang et al. 2012), or repeatedly applying the same method to maximise publication output (Tian et al. 2016), such as performing meta-analyses of existing published data (Ye et al. 2013). Although no medical doctors were interviewed by Tian et al. (2016), reference was made to medical doctors in China using education companies to obtain research results, and then asking others to write the ensuing manuscript. This approach resembles the sale of research manuscripts in China, as uncovered by Hvistendahl (2013). Uncertainty was expressed as to whether education company-supplied results were real, or fabricated (Tian et al. 2016). The retraction notice for Pan et al. (2015) subsequently claimed that the described experiments had been out-sourced to a biotechnology company (Retraction 2016). It is currently unknown as to whether data supplied by any education/biotechnology company would represent the results of bona fide laboratory experimentation, data falsification, invention, or a combination of these activities. However, the fact that the descriptions of some gene knockdown experiments (Tables 4, 5) were entirely inconsistent with the results obtained suggests that at least some of these experiments were not performed as described.

Unrealistic pressure to publish could generate a significant demand for "assisted" manuscripts, particularly if academics in China are expected to publish multiple articles per year (Tian et al. 2016). One strategy to meet this demand could include content development using a "theme and variations" model. For example, content could be developed around particular human genes ("themes") that have been under-investigated relative to others, but can nonetheless be linked with a topic of broad interest, such as human cancer. *TPD52L2* is an example of such a gene, but others may have been similarly targeted (Table 5). Due to the lack of existing publications, understudied genes could then be targeted in multiple cancer types (the "variations"), giving rise to multiple publications. Targeting under-studied genes or processes could also increase the likelihood that content errors go undetected during peer review, although nonsensical "rescue" experiments (Yang et al. 2015) (Table 4), and use of an identical shRNA for *TPD52L2* targeting and as a non-targeting control (Pan et al. 2015) (Table 4) should not have required restricted expertise for their detection. Simultaneous production and then submission of related manuscripts may then prevent text similarities from being detected by plagiarism-detecting software. Journals that do not screen submitted manuscripts for text plagiarism could also be actively targeted. Other developments within scientific publishing may be creating a more receptive market for poor quality research, such as increasing numbers of both academic journals (Michels and Schmoch 2012) and publications in some fields (Pautasso 2012), leading to both information and peer-reviewer overload (Anderson et al. 2007; Parolo et al. 2015; Pautasso 2012; Siebert et al. 2015).

A "theme and variations" model could also allow content to be generated with greater efficiency, at less cost. If the intent of content generation is to meet demand and gain profit, both aims would be more efficiently served if research content were largely or entirely

fabricated. Based on past experience (Roslan et al. 2014; Shehata et al. 2008), the core experiments described by index studies (Tables 1, 2) could require 6 months–1 year of laboratory work. If manuscripts are then sold in China for US$1600–$26,300 (Hvistendahl 2013), the cost of generating bona fide experimental data could exceed the market's capacity to pay. This fact, combined with other errors identified in these papers (Tables 4, 5), suggests that at least some of these gene knockdown data could be fabricated.

Widespread "scientific" content manufacture occurred through use of the SCIgen algorithm (Bohannon 2015; Djuric 2015; Labbé and Labbé 2013), and while such nonsensical manuscripts can be distracting, they are unlikely to lead to further research. However, misreporting the results of gene knockdowns in human cancer cell lines could have very different consequences. Pre-clinical cancer research results are drawn upon for the translation of research to patients. Novel pre-clinical results supported by multiple recent, independent reports (Tables 2, 5) may encourage others to replicate or extend such findings. Indeed, all 5 *TPD52L2* publications encourage further research, and highlight the possible clinical relevance of their findings (He et al. 2015; Pan et al. 2015; Wang et al. 2014; Xu et al. 2015; Yang et al. 2015). At worst, such an approach may endanger patient safety, if only by stopping or delaying more promising research. At best, this could lead to financial and human research resources being wasted in research environments that are increasingly competitive (Anderson et al. 2007; Fang and Casadevall 2015; Kornfeld 2012). Finally, systematic data invention and reporting could have broader adverse consequences to specific trust in pre-clinical cancer research, overall trust in science, and use of scientific results in broader decision making.

Strengths and weaknesses of the approaches used

The combination of content expertise and textual analysis capacity represented a strength of our approach. Sole reliance upon automated tools for the detection of misconduct seems dangerous for many reasons, including the fact that such tools may be used to define boundaries between acceptable and unacceptable practices. A more valuable approach is to employ automatic and mining tools to help to identify novelty, spot errors or highlight inconsistencies, and to combine these approaches with the identification of real and understandable facts. However, the identification of such facts required the visual inspection of articles, and this undoubtedly limited throughput and risked introducing bias. As such, we did not attempt to conduct exhaustive searches for publications that were highly similar to index cases. We also recognise that as there is presently no definitive explanation for the similarity between the studies that we have identified (Tables 3, 5), more extensive searches could be considered to be unwarranted at this point.

Future directions

If future investigations confirm the widespread existence of gene knockdown studies containing education/biotechnology company-derived data, it will be vitally important to be able to robustly identify such publications, in order to measure their prevalence within the research literature, distinguish them from genuine contributions, and deter their future publication. As a first step, the broader application of the methods that we describe is predicted to identify other very similar papers, and our own analyses support this (data not shown). However, we noted that intertextual distance analyses (Fig. 2; Table 5) did not identify the very high level of similarity identified between SCIgen articles (Labbé and

Labbé 2013), possibly because SCIgen works with a limited vocabulary, creating greater degrees of similarity between SCIgen papers. This suggests that the robust detection of articles containing biotechnology or education company-produced data may be more challenging, and may require new search algorithms to be developed. Another key feature of the papers identified was the similar appearance of their figures (Tables 3, 5), and development of algorithms to reliably identify similar figures may be useful.

In the shorter term, the development of particular algorithms may help to identify potentially suspect papers for further analysis. Automated extraction of control nucleotide sequences from papers combined with nucleotide sequence homology searches could identify additional publications that have employed negative control siRNA/shRNA sequences with homology to known genes. However, RNA interference publications often also describe control RT-PCR primer sequences that show high levels of homology to known genes. Matches between control nucleotide sequences and genes could therefore identify large numbers of falsely-positive manuscripts. Another source of error is the splitting of DNA sequences between text lines, which can prevent their identification. As evidence of this problem, Google Scholar searches did not identify every instance of sequences A, B, D or F that we could visually identify within the 5 *TPD52L2* publications (Table 4).

All of the papers identified also employed human cancer cell lines, and one index publication (He et al. 2015) was ultimately retracted because some experiments were performed in a mis-identified cell line (He et al. 2016). Computational approaches could be designed to automatically identify papers that include cell lines that have been flagged as contaminated and/or mis-identified. However, variations in cell line identifier use could again lead to both falsely-positive and -negative results. Furthermore, performing experiments in contaminated or mis-identified cell lines may be unintentional, as opposed to a hallmark of misconduct.

Conclusions

In summary, we report preliminary evidence that education/biotechnology companies may be providing content pertaining to gene knockdown experiments in human cancer cell lines to researchers based in China, who then publish these results without disclosing their origin. A driving force behind these publications could be imposed publication quotas, which may be particularly difficult for Chinese medical doctors to meet, given their limited time for research (Hvistendahl 2013; Tian et al. 2016). The possibility that "assisted" manuscripts may be produced on a large scale supports the inadvisability of employing publication quotas for performance management or career progression (Djuric 2015; Lin 2013), and of considering research to be a career necessity for physicians (Altman 2002). The future detection and deterrence of such fraudulent manuscripts will be of vital importance, to ensure that such results do not mis-direct future cancer research efforts, or reduce broader trust in science and research.

Acknowledgements We thank Ms Mara Hvistendahl for invaluable support and assistance. JAB thanks journal editors and peer reviewers for their assistance, and members of the Children's Cancer Research Unit for discussions.

References

Altman, D. G. (2002). Poor-quality medical research: What can journals do? *JAMA, 287*(21), 2765–2767. doi:10.1001/jama.287.21.2765.

Amancio, D. R. (2015). Comparing the topological properties of real and artificially generated scientific manuscripts. *Scientometrics, 105*(4), 1763–1779. doi:10.1007/s11192-015-1637-z.

Amancio, D. R., Aluisio, S. M., Oliveira, O. N., & Costa, L. da F. (2012). Complex networks analysis of language complexity. *Europhysics Letters, 100*(5), 58002. doi:10.1209/0295-5075/100/58002.

Anderson, M. S., Ronning, E. A., De Vries, R., & Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics, 13*(4), 437–461. doi:10.1007/s11948-007-9042-5.

Argamon, S., & Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In *Proceeding of the 2005 ACH/ALLC conference*, June 2005, Victoria, BC, Canada.

Ausloos, M., Nedic, O., Fronczak, A., & Fronczak, P. (2016). Quantifying the quality of peer reviewers through Zipf's law. *Scientometrics, 106*(1), 347–368. doi:10.1007/s11192-015-1704-5.

Bik, E. M., Casadevall, A., & Fang, F. C. (2016). The prevalence of inappropriate image duplication in biomedical research publications. *mBio, 7*(3), e00809–e00816. doi:10.1128/mBio.00809-16.

Bohannon, J. (2015). Hoax-detecting software spots fake papers. *Science, 348*(6230), 18–19. doi:10.1126/science.aab0381.

Bornmann, L. (2013). Research misconduct-definitions, manifestations and extent. *Publications, 1,* 87–98. doi:10.3390/publications1030087.

Bowen, A., & Casadevall, A. (2015). Increasing disparities between resource inputs and outcomes, as measured by certain health deliverables, in biomedical research. *Proceedings of the National Academy of Sciences of the United States of America, 112*(36), 11335–11340. doi:10.1073/pnas.1504955112.

Boyer-Guittaut, M., Poillet, L., Liang, Q., Bôle-Richard, E., Ouyang, X., Benavides, G. A., et al. (2014). The role of GABARAPL1/GEC1 in autophagic flux and mitochondrial quality control in MDA-MB-436 breast cancer cells. *Autophagy, 10*(6), 986–1003. doi:10.4161/auto.28390.

Butler, L. (2003). Explaining Australia's increased share of ISI publications—The effects of a funding formula based on publication counts. *Research Policy, 32*(1), 143–155. doi:10.1016/S0048-7333(02)00007-0.

Byrne, J. A., Frost, S., Chen, Y., & Bright, R. K. (2014). Tumor protein D52 (TPD52) and cancer—Oncogene understudy, or understudied oncogene? *Tumour Biology, 35*(8), 7369–7382. doi:10.1007/s13277-014-2006-x.

Capes-Davis, A., & Neve, R. M. (2016). Authentication: A standard problem or a problem of standards? *PLoS Biology, 14*(6), e1002477. doi:10.1371/journal.pbio.1002477.

Carpena, P., Bernaola-Galván, P., Hackenberg, M., Coronado, A. V., & Oliver, J. L. (2009). Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, 79*(3 Pt 2), 035102. doi:10.1103/PhysRevE.79.035102.

Casadevall, A., Steen, R. G., & Fang, F. C. (2014). Sources of error in the retracted scientific literature. *The FASEB Journal, 28*(9), 3847–3855. doi:10.1096/fj.14-256735.

Citron, D. T., & Ginsparg, P. (2015). Patterns of text reuse in a scientific corpus. *Proceedings of the National Academy of Sciences of the United States of America, 112*(1), 25–30. doi:10.1073/pnas.1415135111.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, IT-13,* 21–27.

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journla of the Association for Information Science and Technology, 65,* 446–454. doi:10.1002/asi.23056.

Djuric, D. (2015). Penetrating the omerta of predatory publishing: The Romanian connection. *Science and Engineering Ethics, 21*(1), 183–202. doi:10.1007/s11948-014-9521-4.

Fahrenberg, U., Biondi, F., Corre, K., Jégourel, C., Kongshøj, S., & Legay, A. (2014). Measuring global similarity between texts. In L. Besacier et al. (Eds.), *Statistical language and speech processing, lecture notes in computer science* (Vol. 8791, pp. 220–232). Switzerland: Springer International Publishing. doi:10.1007/978-3-319-11397-5_17.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE, 4*(5), e5738. doi:10.1371/journal.pone.0005738.

Fang, F. C., & Casadevall, A. (2015). Competitive science: Is competition ruining science? *Infection and Immunity, 83*(4), 1229–1233. doi:10.1128/IAI.02939-14.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences of the United States of America, 109*(42), 17028–17033. doi:10.1073/pnas.1212247109.

Ferguson, C., Marcus, A., & Oransky, I. (2014). The peer-review scam. *Nature, 515*(7528), 480–482.
Ginsparg, P. (2014). Automated screening: Arxiv screens spot fake papers. *Nature, 508*(7494), 44. doi:10.1038/508044a.
He, Y., Chen, F., Cai, Y., & Chen, S. (2015). Knockdown of tumor protein D52-like 2 induces cell growth inhibition and apoptosis in oral squamous cell carcinoma. *Cell Biology International, 39*(3), 264–271. doi:10.1002/cbin.10388.
He, Y., Chen, F., Cai, Y., & Chen, S. (2016). Retracted: Knockdown of tumor protein D52-like 2 induces cell growth inhibition and apoptosis in oral squamous cell carcinoma. *Cell Biology International, 40*(3), 361. doi:10.1002/cbin.10593.
Hockey, S., & Martin, J. (1988). *OCP users' manual*. Oxford: Oxford University Computing Service.
Hvistendahl, M. (2013). China's publication bazaar. *Science, 342*(6162), 1035–1039. doi:10.1126/science.342.6162.1035.
Konwisorz, A., Springwald, A., Haselberger, M., Goerse, R., Ortmann, O., & Treeck, O. (2010). Knockdown of ICB-1 gene enhanced estrogen responsiveness of ovarian and breast cancer cells. *Endocrine-Related Cancer, 17*(1), 147–157. doi:10.1677/ERC-09-0095.
Kornfeld, D. S. (2012). Perspective: Research misconduct: The search for a remedy. *Academic Medicine, 87*(7), 877–882. doi:10.1097/ACM.0b013e318257ee6a.
Kreutzberg, G. W. (2004). The rules of good science. *EMBO Reports, 5*(4), 330–332. doi:10.1038/sj.embor.7400136.
Labbé, C., & Labbé, D. (2012). Detection of hidden intertextuality in the scientific publications. In *11th International conference on textual data statistical analysis*, 2012, Liège, Belgium (pp. 537–551). Liège: LASLA - SESLA.
Labbé, C., & Labbé, D. (2013). Duplicate and fake publications in the scientific literature: How many SCIgen papers in computer science? *Scientometrics, 94*(1), 379–396. doi:10.1007/s11192-012-0781-y.
Lesk, M. (2015). How many scientific papers are not original? *Proceedings of the National Academy of Sciences of the United States of America, 112*(1), 6–7. doi:10.1073/pnas.1422282112.
Lin, S. (2013). Why serious academic fraud occurs in China. *Learned Publishing, 26*(1), 24–27. doi:10.1087/20130105.
Mehri, A., Darooneh, A. H., & Shariati, A. (2012). The complex networks approach for authorship attribution of books. *Physica A: Statistical Mechanics and its Applications, 391*(7), 2429–2437. doi:10.1016/j.physa.2011.12.011.
Michels, C., & Schmoch, U. (2012). The growth of science and database coverage. *Scientometrics, 93*(3), 831–846. doi:10.1007/s11192-012-0732-7.
Moore, R. A., Derry, S., & McQuay, H. J. (2010). Fraud or flawed: Adverse impact of fabricated or poor quality research. *Anaesthesia, 65*(4), 327–330. doi:10.1111/j.1365-2044.2010.06295.x.
Nourse, C. R., Mattei, M. G., Gunning, P., & Byrne, J. A. (1998). Cloning of a third member of the D52 gene family indicates alternative coding sequence usage in D52-like transcripts. *Biochimica et Biophysica Acta, 1443*(1–2), 155–168.
Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*. doi:10.3758/s13428-015-0664-2.
Pan, Z. Y., Yang, Y., Pan, H., Zhang, J., Liu, H., Yang, Y., et al. (2015). Lentivirus-mediated *TPD52L2* depletion inhibits the proliferation of liver cancer cells in vitro. *International Journal of Clinical and Experimental Medicine, 8*(2), 2334–2341.
Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., & Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics, 9*(4), 734–745. doi:10.1016/j.joi.2015.07.006.
Pautasso, M. (2012). Publication growth in biological sub-fields: Patterns, predictability and sustainability. *Sustainability, 4*(12), 3234–3247. doi:10.3390/su4123234.
Retraction. (2016). Lentivirus-mediated *TPD52L2* depletion inhibits the proliferation of liver cancer cells in vitro [Retraction]. *International Journal of Clinical and Experimental Medicine, 9*(6), 12416.
Roslan, N., Bièche, I., Bright, R. K., Lidereau, R., Chen, Y., & Byrne, J. A. (2014). TPD52 represents a survival factor in ERBB2-amplified breast cancer cells. *Molecular Carcinogenesis, 53,* 807–819. doi:10.1002/mc.22038.
Shehata, M., Bieche, I., Boutros, R., Weidenhofer, J., Fanayan, S., Spalding, L., et al. (2008). Non-redundant functions for tumor protein D52-like proteins support specific targeting of TPD52. *Clinical Cancer Research, 14,* 5050–5060. doi:10.1158/1078-0432.
Shvets, A. (2014). A method of automatic detection of pseudoscientific publications. In: D. Filev et al. (Eds.), *Intelligent systems'2014, advances in intelligent systems and computing* (Vol. 323, pp. 533–539). Switzerland: Springer International Publishing. doi:10.1007/978-3-319-11310-4_46.

Siebert, S., Machesky, L. M., & Insall, R. H. (2015). Overflow in science and its implications for trust. *Elife*. doi:10.7554/eLife.10825.

Smith, R. (2006). Research misconduct: The poisoning of the well. *Journal of the Royal Society of Medicine, 99*(5), 232–237. doi:10.1258/jrsm.99.5.232.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538–556. doi:10.1002/asi.21001.

Steen, R. G. (2011a). Retractions in the scientific literature: Do authors deliberately commit research fraud? *Journal of Medical Ethics, 37*(2), 113–117. doi:10.1136/jme.2010.038125.

Steen, R. G. (2011b). Misinformation in the medical literature: What role do error and fraud play? *Journal of Medical Ethics, 37*(8), 498–503. doi:10.1136/jme.2010.041830.

Tian, M., Su, Y., & Ru, X. (2016). Perish or publish in China: Pressures on young Chinese scholars to publish in internationally indexed journals. *Publications, 4,* 9. doi:10.3390/publications4020009.

Tuzzi, A. (2010). What to put in the bag? Comparing and contrasting procedures for text clustering. *Statistica Applicata-Italian Journal of Applied Statistics, 22*(1), 81–98.

van Dalen, H., & Henkens, K. (2012). Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. *JASIS&T, 63*(7), 1282–1293. doi:10.1002/asi.22636.

Verma, S., Tabb, M. M., & Blumberg, B. (2009). Activation of the steroid and xenobiotic receptor, SXR, induces apoptosis in breast cancer cells. *BMC Cancer, 9,* 3. doi:10.1186/1471-2407-9-3.

Wang, Z., Sun, J., Zhao, Y., Guo, W., Lv, K., & Zhang, Q. (2014). Lentivirus-mediated knockdown of tumor protein D52-like 2 inhibits glioma cell proliferation. *Cellular and Molecular Biology (Noisy-le-grand), 60*(1), 39–44.

Wang, X., Xu, S., Peng, L., Wang, Z., Wang, C., Zhang, C., et al. (2012). Exploring scientists' working timetable: Do scientists often work overtime? *Journal of Informetrics, 6*(4), 655–660. doi:10.1016/j.joi.2012.07.003.

White, C. (2005). Suspected research fraud: Difficulties of getting at the truth. *BMJ, 331*(7511), 281–288. doi:10.1136/bmj.331.7511.281.

Wilson, S. H., Bailey, A. M., Nourse, C. R., Mattei, M. G., & Byrne, J. A. (2001). Identification of MAL2, a novel member of the MAL proteolipid family, though interactions with TPD52-like proteins in the yeast two-hybrid system. *Genomics, 76*(1–3), 81–88. doi:10.1006/geno.2001.6610.

Xu, J., Wang, W., Zhu, Z., Wei, Z., Yang, D., & Cai, Q. (2015). Tumor protein D52-like 2 accelerates gastric cancer cell proliferation in vitro. *Cancer Biotherapy and Radiopharmaceuticals, 30*(3), 111–116. doi:10.1089/cbr.2014.1766.

Yang, M., Wang, X., Jia, J., Gao, H., Chen, P., Sha, X., et al. (2015). Tumor protein D52-like 2 contributes to proliferation of breast cancer cells. *Cancer Biotherapy and Radiopharmaceuticals, 30*(1), 1–7. doi:10.1089/cbr.2014.1723.

Ye, X.-F., Yu, D.-H., & He, J. (2013). The rise in meta-analyses from China. *Epidemiology, 24*(2), 335–336. doi:10.1097/EDE.0b013e31828264be.

Zeng, W., & Resnik, D. (2010). Research integrity in China: Problems and prospects. *Developing World Bioethics, 10*(3), 164–171. doi:10.1111/j.1471-8847.2009.00263.x.