The Possibility of Systematic Research Fraud Targeting Under-Studied Human Genes: Causes, Consequences, and Potential Solutions

Jennifer Byrne, Natalie Grima, Amanda Capes-Davis, Cyril Labbé

The Possibility of Systematic Research Fraud Targeting Under-Studied Human Genes: Causes, Consequences, and Potential Solutions

Biomarker Insights
Volume 14: 1–12

DOI: 10.1177/1177271919829162
SAGE

Jennifer A Byrne[1,2], Natalie Grima[1], Amanda Capes-Davis[3] and Cyril Labbé[4]

[1]Molecular Oncology Laboratory, Children's Cancer Research Unit, Kids Research, The Children's Hospital at Westmead, Westmead, NSW, Australia. [2]Discipline of Child and Adolescent Health, The University of Sydney and The Children's Hospital at Westmead, Westmead, NSW, Australia. [3]CellBank Australia, Children's Medical Research Institute and The University of Sydney, Westmead, NSW, Australia. [4]Univ Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France.

ABSTRACT: A major reason for biomarker failure is the selection of candidate biomarkers based on inaccurate or incorrect published results. Incorrect research results leading to the selection of unproductive biomarker candidates are largely considered to stem from unintentional research errors. The additional possibility that biomarker research may be actively misdirected by research fraud has been given comparatively little consideration. This review discusses what we believe to be a new threat to biomarker research, namely, the possible systematic production of fraudulent gene knockdown studies that target under-studied human genes. We describe how fraudulent papers may be produced in series by paper mills using what we have described as a 'theme and variations' model, which could also be considered a form of salami slicing. We describe features of these single-gene knockdown publications that may allow them to evade detection by journal editors, peer reviewers, and readers. We then propose a number of approaches to facilitate their detection, including improved awareness of the features of publications constructed in series, broader requirements to post submitted manuscripts to preprint servers, and the use of semi-automated literature screening tools. These approaches may collectively improve the detection of fraudulent studies that might otherwise impede future biomarker research.

KEYWORDS: Biomarkers, cancer, gene knockdown techniques, research fraud, paper mill, salami publication, under-studied gene

RECEIVED: October 10, 2018. **ACCEPTED:** January 8, 2019.

TYPE: Review

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/ or publication of this article: This work is supported by the funding from the Post-Truth Initiative, a University of Sydney Research Excellence Initiative (SREI 2020) (to J.A.B.) and from the US Office of Research Integrity Project Grant ORIIR180038-01-00 (to J.A.B. and C.L.). This work was supported by donations to the Children's Cancer Research Unit of the Children's Hospital at Westmead.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Jennifer A Byrne, Molecular Oncology Laboratory, Children's Cancer Research Unit, Kids Research, The Children's Hospital at Westmead, Locked Bag 4001, Westmead, NSW 2145, Australia.
Email: jennifer.byrne@health.nsw.gov.au

Introduction

Biomarkers in human health research

Biomarkers represent a type of Holy Grail in research. The concept that single or groups of analytes can illuminate and inform complex biological and disease processes continues to inspire countless researchers and their supporting institutions and investors. Biomarkers that can be measured reliably at low cost in accessible biospecimens have played significant roles in improving human health.[1] Many biomarkers now guide routine medical diagnoses through standardised testing processes that can continue to be refined and improved through ongoing research.

New biomarkers continue to be sought, particularly in the field of cancer.[1–3] The increasing molecular complexity of cancer sub-types, the costs of treatments and the periods of time during which patients can undergo treatment, and post-treatment monitoring, all provide numerous opportunities for reliable biomarkers to improve patient outcomes and/or reduce treatment cost and complexity.[1–3] Because of the cost of cancer treatments and the long periods during which patients may be treated, different types of biomarkers are required.[1–3] Biomarkers can enable disease diagnosis,[1,2] particularly in the era of personalised medicine where it is becoming increasingly apparent that molecular drivers may not always be obvious from the cancer phenotype. Biomarkers can predict patient survival,[1,2] and responses to treatment,[1–3] with a subset of biomarkers representing so-called companion diagnostics used to assign patients to molecularly targeted treatments.[2] Biomarkers can also enable disease monitoring,[1,2] allowing disease progression to be detected earlier and therefore treated more effectively.

Challenges in cancer biomarker research

Quality biomarker research relies on several key factors, including the selection of plausible candidate biomarkers, the availability of appropriate reagents and techniques with which to investigate these candidate biomarkers, and appropriate cohorts of fit-for-purpose biospecimens from informative and relevant populations in which to test these biomarkers.[1-4] The search for cancer biomarkers has, however, been plagued by problems that have meant that outcomes from this field have not always met their expected promise.[1,3-6] Some reasons for biomarker failure lie in the nature of cancer biology itself. Cancer is rarely the result of a single gene, protein, or cellular pathway, but

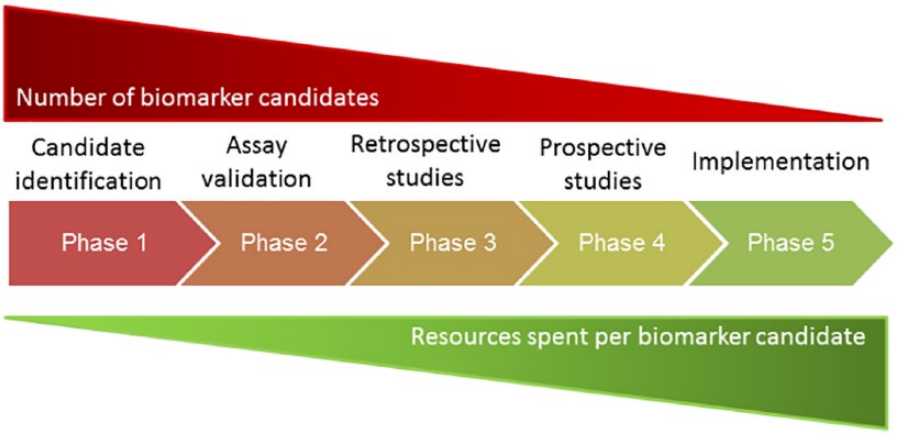

Figure 1. Five phases of biomarker development, shown as the biomarker research pipeline, adapted from Pepe et al.[9] The number of biomarker candidates within the pipeline (shown above the phase diagram) progressively reduces as candidate biomarkers are sequentially analysed, and a proportion of candidates are discarded. At the same time, the resources required to advance each candidate progressively increase (shown below the phase diagram). The selection of unproductive candidates at phase 1 may prevent more productive candidates from entering the pipeline.

instead involves many potential driver genes working in concert to produce complex, interwoven, and sometimes unstable phenotypes.[2,7] As such, it may be difficult to identify single or small numbers of analytes whose measurement accurately reflects very complex composite phenotypes.[7] It can be even more difficult for single biomarkers to outperform existing gold standard descriptors such as tumour stage and grade, which represent the combined action of many individual biological factors that have been operating over time periods ranging from weeks to decades.

Selection of new candidate cancer biomarkers

A fundamental component of precision medicine and biomarker research is the selection of plausible or apparently promising candidate targets and biomarkers for analysis and further testing.[1-3,5-9] The selection of candidate biomarkers is of prime importance (Figure 1), as candidate selection determines many aspects of subsequent downstream analyses and research resource allocation.[2,3,5,6,8–10] Investigating large numbers of candidate biomarkers also increases the likelihood of individual candidates reaching statistical significance through chance alone.[6] Any decision process that improves the selection of productive biomarker candidates, at the expense of incorrect candidates, could greatly improve the efficiency of the biomarker research pipeline (Figure 1).

Plausible candidate biomarkers can emerge from a number of approaches. Protein targets of targeted therapies or components of affected downstream signalling pathways are obvious candidate biomarkers that can also serve as companion diagnostics.[1,2] Another approach is to mine the experimental results of high-throughput studies conducted in the field of interest. For example, genes that are found to be recurrently mutated or overexpressed may represent molecular drivers of cancer and may also represent prognostic biomarkers or agents for therapeutic monitoring.[2] When selecting biomarkers for follow-up, researchers may integrate the results of high-throughput studies with results from targeted analyses of the function of the gene or protein in question (Figure 2). These targeted studies may have examined the effects of overexpressing the gene and/or of knocking down or inhibiting its function in cancer cell lines and/or non-transformed control cell types. Multiple studies commonly suggesting that a candidate gene drives cancer phenotypes, either in a single cancer type or in different cancer types, could increase the priority of such a biomarker candidate (Figure 2). This may favour the selection of candidates supported by multiple lines of experimental evidence for further analysis to the possible exclusion of other candidates lacking such experimental evidence (Figure 2).

Research fraud and the selection of unproductive biomarker candidates

A major reason for biomarker failure is the selection of candidate biomarkers based on inaccurate or incorrect published experimental results.[5,6] Incorrect research results that drive the selection of failed biomarker candidates are largely considered to derive from myriad forms of unintentional research error.[5,6] The additional possibility that biomarker research may be actively misdirected by research misconduct or fraud has been recognised,[5,11] but given comparatively little consideration in the biomarker literature.

Despite ongoing debate concerning the full definition of research misconduct, the deliberate falsification or fabrication of research findings are broadly recognised as fraudulent practices.[12] Research fraud in the form of data fabrication, falsification, and alteration is admitted by only a small minority of researchers,[13] leading to the perception that research fraud occurs rarely within the scientific community. However, publication retractions also provide a combined measure of research error and fraud, and as the numbers of retracted papers have risen sharply since the late 1990s,[14] the prevalence of research

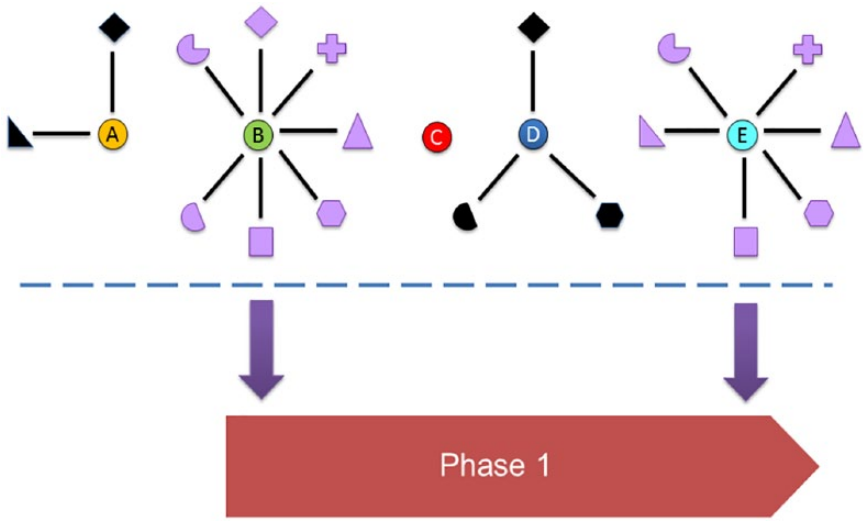

Figure 2. Diagrammatic representation showing how the availability of research studies that support candidate genes can influence their selection to enter the biomarker research pipeline (phase 1, see Figure 1). Five genes (A-E) are shown. Genes A, B, D, and E are connected to functional studies (surrounding symbols) that support that gene as a candidate biomarker within a particular cancer type. Functional studies performed in different cancer types are shown as distinct symbols, with black symbols denoting bona fide published studies and purple symbols denoting fraudulent published studies. Without an understanding that genes B and E have been systematically targeted for the fraudulent production of manuscripts in series, genes B and E appear to be the best biomarker candidates and would be most likely to be selected for further biomarker studies. This decision could exclude genes A and D, which are supported by bona fide research, and which could in fact represent superior candidates.

fraud may also be rising and/or underestimated.[13,14] Numbers of retracted papers are highest from countries which produce the most publications, namely, the United States and China.[15,16]

As research falsification and fabrication are widely recognised to be incorrect practices, they are usually actively concealed by perpetrators and can be difficult to detect.[17] Research fraud is generally considered to be perpetrated by lone actors or teams and to be either driven by factors specific to the researchers themselves (eg, psychiatric illness or outlying research beliefs) or by how rare individuals respond to their research environments.[17-19] Research fraud driven by individual actors or teams can nonetheless give rise to many fraudulent publications, typically dispersed over years to decades.[20,21] More recently, it has been recognised that research fraud can also occur on a wider scale,[22–24] which this review will term systematic fraud. Although the drivers of systematic fraud have not been broadly investigated, fraud on a widespread scale would appear to have less to do with the psychology or behaviour of a minority of researchers, and more to do with the overarching research culture or professional environment.[22,25] By involving many more individual actors and research teams, systematic fraud has the potential to produce very large numbers of fraudulent publications.

A number of factors may render cancer biomarker research as particularly fertile ground for research misconduct and fraud. A major factor is the link between gene dysregulation and cancer. There are approximately 20 000 protein-coding genes in the human genome and a similar estimated number of non-protein-coding or non-coding genes.[26] It has been recognised that biomedical research focuses on only a small proportion of protein-coding genes, most of which were identified before the human genome was first sequenced.[27–32] Most of the human genes therefore remain under-investigated and poorly understood from a functional perspective.[29,31,32] We are concerned that some under-studied human genes are being actively exploited for poor quality and possibly fraudulent published research.[33]

Under-investigated genes represent easy targets for low quality and fraudulent research. By definition, little is known about under-studied genes, so there are many literature gaps that can be filled by individual publications. For example, in the context of cancer research, most genes can be examined in different cancer types through the availability of corresponding cancer cell lines, potentially generating many individual publications around the functions of single genes.[33] Filling minor literature gaps is likely be enabled by the proliferation of specialty journals,[34] some of which may be willing to publish manuscripts of limited value by imposing less rigorous peer review standards. By definition, the lack of publications focussing on under-studied genes also renders the peer review process more challenging. Where a gene has been the subject of hundreds or thousands of publications, this generates a wider pool of expertise from which journal editors can select peer reviewers. In contrast, there will be comparatively few published experts with in-depth knowledge of under-studied genes. Most journal editors are aware of the difficulty in obtaining quality peer reviewers,[35] which may be more acute for manuscripts submitted to specialty journals.[36] Where suitable peer reviewers are already limited in number, this further reduces the likelihood of securing peer reviewers with relevant expertise about topics such as under-studied genes. In summary, the current landscape that combines (1) many thousands of under-studied

Figure	Panel A	Panel B	Panel C
Fig 1	Intracellular GFP detection, microscopic cell images	RT-PCR, graph	Western blot analysis, images of detected bands
	Confirmation of shRNA transfection	Confirmation of gene knockdown at transcript level	Confirmation of gene knockdown at protein level
Fig 2	Cell proliferation assays, graph	Cell staining, microscopic cell images	Anchorage independent growth assays, graph
	Reduced cell proliferation	Visualisation of gene knockdown phenotypes	Reduced anchorage independent growth
Fig 3	FACS cell cycle plots	FACS cell cycle graph	FACS cell cycle graph
	Altered cell cycle distribution	Altered cell cycle distribution	Increased proportion of cells in sub-G1 cell cycle phase

Figure 3. Summary of the conserved series of experimental results shown in 5 *TPD52L2* knockdown studies,[37–41] 4 of which have been retracted from the literature.[37–40] The order of conserved figures is shown vertically and figure panels within each figure are shown horizontally. The data that were shown (upper panels) and the purpose of the experiments (lower panels) are described for each individual figure panel.

human genes, (2) many individual biological or disease systems in which under-studied genes can be examined, (3) growing numbers of specialist journals seeking manuscripts to publish, (4) inadequate peer review standards at some journals, and (5) a lack of content-expert peer reviewers may unwittingly produce a fertile environment for fraudulent publications targeting under-studied genes.

Strikingly similar gene knockdown publications targeting human genes

We have previously proposed that under-studied genes may represent templates for the systematic construction of fraudulent pre-clinical cancer research manuscripts, based on striking similarities between publications, combined with fundamental shared errors in experimental design.[33] We reported 48 examples of publications that described the effects of knocking down a single human gene, typically in 1 to 2 human cancer cell lines that corresponded to a single cancer type.[33] These publications were characterised by unusual levels of textual and organisational similarity.[33] In each case, these 48 publications were authored by teams based in mainland China.[33]

In many respects, the single-gene knockdown papers that we described[33] superficially resemble many other gene knockdown publications that can be identified in the literature. The cohort of papers that we reported usually first demonstrate successful gene knockdown reagent transfection using green fluorescent protein (GFP) staining and then verify knockdown of the gene of interest at transcript and protein levels, typically using some form of reverse-transcriptase polymerase chain reaction (PCR) and Western blot analyses (Figure 3).[33] In some cases, the selection of gene of interest was supported by the analysis of protein expression in clinical patient cohorts, where the results of immunohistochemical analyses were compared with clinicopathological characteristics of the associated tumours and patients.[33] Papers then consecutively analysed the effects of gene knockdown using standard approaches such as cell proliferation and colony formation assays (Figure 3).[33] Another common experiment was the use of fluorescence-activated cell sorting (FACS) to analyse the effects of gene knockdown on cell cycle distributions, which can also specifically analyse the presence of cells in the sub-G_1 phase as a marker for apoptosis (Figure 3).[33] Some papers included additional experiments such as assays of the effects of gene knockdown on cell migration and invasion and/or the detection of the levels of downstream cell signalling proteins through Western blot analyses. Some papers also reported the results of comparing the growth of transfected cell lines in vivo as mouse xenografts. Most of the results included in each paper supported the gene of interest playing a key role in the cancer models analysed,[33] in line with the published biomedical literature's recognised bias towards reporting successful experiments or positive results.[42,43]

Despite their superficial resemblance to conventional gene knockdown papers, a detailed analysis of these papers revealed a number of unusual and concerning features.[33] The first members of the publication cohort to be identified were 5 publications that commonly described the effects of knocking down the *TPD52L2* gene in breast, gastric, glioma, liver, and oral cancers, respectively.[37–41] The *TPD52L2* gene was identified by one of the study authors (J.A.B.)[44] as the third member of the *TPD52* gene family.[45] As the *TPD52L2* gene had been the subject of approximately one paper per year since 1998, the

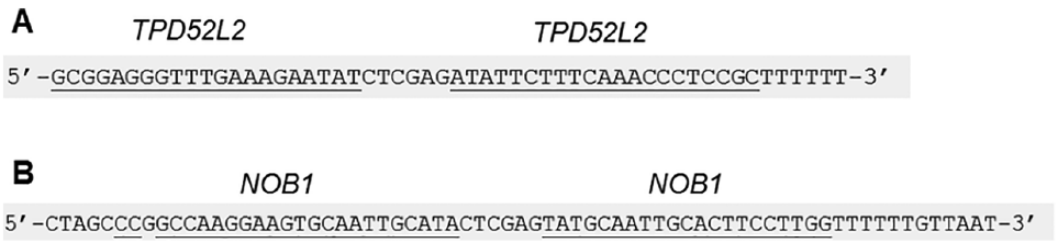

Figure 4. Incorrectly identified short hairpin RNA (shRNA) sequences that were described in *TPD52L2* knockdown studies[37–41] and other single-gene knockdown publications.[33] Nucleotide sequences are shown 5′-3′. Nucleotides that are identical to their indicated target according to blastn[46] search results are underlined. (A) *TPD52L2* shRNA, correctly used as a targeting sequence[37,38,40,41] and incorrectly used as a non-targeting sequence.[33,37,39] (B) *NOB1* shRNA, incorrectly used as a non-targeting sequence.[33,38,40]

appearance of 5 *TPD52L2* knockdown papers in less than 1 year[37–41] seemed unusual. We noted that these 5 papers described a very similar series of figures in the same relative order (Figure 3).[33] The 5 papers had in many cases also used identical RNA interference reagents for gene knockdown experiments (Figure 4).[33]

Nucleotide sequence reagents form the foundation for gene knockdown studies, as every subsequent downstream experiment relies on correct nucleotide sequence reagents having been used to achieve gene knockdown. Because published nucleotide sequences are typically provided near or adjacent to a functional descriptor of the reagent, sequence identities can be cross-checked by performing independent blastn analyses.[33,46] In this sense, published nucleotide sequences represent verifiable facts through the precise relationship that exists between an individual nucleotide sequence, its genetic identity, and therefore its possible experimental use. Blastn analyses of nucleotide sequence reagents described by the 5 *TPD52L2* gene knockdown papers revealed that 4 of these papers were characterised by incorrectly described reagents, including 2 supposedly non-targeting RNA interference reagents that are predicted to target the *TPD52L2* or *NOB1* genes (Figure 4).[33] Mismatches between some blastn-confirmed reagent identities and their described experimental use rendered particular results impossible, such as obtaining opposing experimental results when the same *TPD52L2* short hairpin RNA was employed as both a targeting reagent and a non-targeting control (Figure 4).[33,37]

Following the identification of 5 *TPD52L2* papers,[37–41] other very similar publications describing the effects of knocking down individual human genes in cancer cell lines were identified using PubMed similarity searches and Google Scholar searches employing incorrectly identified nucleotide sequences as search queries.[33] Although each of these papers examined a single human gene in a single cancer type, the cohort included a number of papers that studied common genes. Just as the *TPD52L2* gene was analysed in 5 different cancer types,[37–41] the publication cohort also included 8 other genes that were each examined in 2 to 6 different cancer types.[33] An intertextual distance similarity threshold[47] was used to specify a minimum level of textual similarity between publications and a defined reference cohort and to compare the levels of textual similarity between the papers within the reported cohort.[33] Although the degree of textual similarity between particular studies suggested that text plagiarism had occurred in some cases, the cohort also featured very similarly formatted and presented figures, extending in many cases to the use of a common text font for figure annotation.[33]

Following the published description of 48 single-gene knockdown publications,[33] J.A.B. and C.L. wrote to numerous journal editors to express their concerns about the similarities between and errors within these and other papers. As a result, 17 publications[37–40,48–60] have been retracted, including 14 of the 48 publications originally described.[37–40,48–57] Four other publications have been corrected,[61–64] and 5 Expressions of Concern have been published,[65–69] with other journal investigations still ongoing. The list of retracted papers includes a 6th member of the *TPD52L2* publication series that examined the effects of knocking down the *TPD52L2* gene in lung cancer.[60] In addition to the papers reported by J.A.B. and C.L., another single-gene knockdown publication was recently retracted due to the use of an incorrectly identified nucleotide sequence reagent.[70]

Possible origins of highly similar single-gene knockdown papers. Based on the similarities between the single-gene knockdown papers that we described, we attempted to explain how such strikingly similar papers could have arisen.[33] Any explanation needed to account for the high degree of textual similarity, the description of very similar series of experiments (Figure 3), the presence of highly similar figures, the repeated analysis of particular genes in different cancer types, and shared errors involving common, incorrectly identified nucleotide sequence reagents (Figure 4), without obvious overlaps in authorship between studies examining common genes. Furthermore, as some experimental results were inconsistent with the verified identities of particular nucleotide sequence reagents, at least some experiments could not have been performed as described.[33]

Typically, scientific manuscript preparation benefits from close working relationships between authors, who either individually or collectively derive the data, compile the data into figures and tables and write the manuscript text. The involvement of key individuals who undertake multiple tasks (such as obtaining data, assembling figures, and writing the text) reduces the likelihood of incorporating errors that incorrectly reflect

how experiments were conducted. In the cases of the single-gene knockdown papers that we described, the very similar appearance of figures, combined with inconsistencies between data descriptions versus the data displayed in figures, suggested the possible uncoupling of the production of figures and text.[33] This could arise if the figures and text were being assembled by different individuals or independent groups. Furthermore, the errors included in these papers also suggested the involvement of individuals with an incomplete understanding of the work described. We hypothesised that these features could be consistent with investigators publishing data that had been obtained from third parties.[33]

The possibility that academics and medical doctors in China and other countries may publish results obtained from undeclared third parties has been discussed within the literature over the past 5 years.[16,71-73] Research publication targets or quotas imposed on University academics and medical doctors have been proposed to be largely responsible for driving a significant market for 'assisted' manuscripts in China.[16,71-73] The organisations supplying undeclared data and/or manuscript content have been variably referred to as education companies, biotechnology companies, or paper mills.[16,71,73] Although we recognise that the functions of these organisations may not be identical,[71] we will refer to these organisations as paper mills henceforth.

While studying the operations of covert organisations such as paper mills is clearly challenging, it has been claimed that researchers may obtain a number of products or services from these companies.[16,71,73] As possible examples, researchers may pay to have their name added to an existing manuscript,[71] or for data that may be written up with or without assistance from paper mill staff,[16,73] or for supplied data and a fully written manuscript.[16,71,73] Importantly, the retraction notice for one publication within the *TPD52L2* paper series claimed that the described experiments had been outsourced to a biotechnology company.[37] This retraction links the possible involvement of undeclared third parties with the data described by single-gene knockdown papers.[33]

Due to the difficulty in investigating paper mill activities, it is not known whether data supplied by these companies reflect the results of bona fide experiments, data manipulation, and/or data invention. However, the improbable results described by some single-gene knockdown papers mean that it is very unlikely that the associated laboratory experiments were performed as described.[33] Indeed, given the high costs of generating laboratory research data using multiple techniques with the required number of technical replicates, content fabrication would appear to be the quickest and cheapest option for paper mills that aim to profit from selling experimental data and results.[33]

A possible 'theme and variations' approach to content generation. The requirement for some academics and medical doctors in China to publish multiple research articles per year could create a very significant demand for 'assisted' manuscripts.[16,73] Furthermore, the involvement of medical doctors as paper mill clients[16,73] could lead to a focus on research manuscripts that are relevant to human disease. Building on this possibility, we believe that features of the single-gene knockdown publications that we have studied could be consistent with a new and highly concerning form of systematic research fraud that employs a 'theme and variations' model for content generation (Figure 5).[33] From our work to date, we believe that paper mills may build scientific content around under-studied human genes which could serve as individual 'themes' (Figure 5).[33] The active targeting of human genes that have been subjected to limited functional assessment would then permit each gene to be investigated in multiple different cancer types, potentially leading to many manuscript 'variations' (Figure 5). This 'theme and variations' model could account for many poor-quality publications that commonly describe the effects of targeting the functions of single under-studied human genes.[33]

Although the process of data fabrication has been described as requiring significant time and resources,[17] the data presented in single-gene knockdown papers would be easy to fabricate at scale. Data shown in graphical format (demonstrating, for example, changes in gene transcript levels, cell proliferation over time, and cell sorting results in response to gene knockdown) could be created by simply typing data into programmes such as Excel. The creation of figures showing digital images could also be achieved at scale using digital image banks, which could either be created by the paper mill or accessed externally. Data image banks could include macroscopic images (eg, bands detected through Western blot analyses, nude mice bearing tumour xenografts, and associated excised tumours) as well as microscopic images (eg, immunohistochemical staining of tissue sections, cultured cells with and without specific staining, and images from cell migration and invasion assays). The derivation of or paid access to data image banks would be justified if these were used to create many individual manuscripts, and their use could explain the repeated description of particular experimental approaches across different publications (Figure 3). Selected images could be assembled into figures and annotated by staff working for the paper mill, which would also explain recurring figure annotation styles, such as the repeated use of a particular text font.[33] In addition to creating figures, the accompanying manuscript text could also be written by researchers employed by the paper mill.[16] The repeated inclusion of common incorrectly identified targeting sequences as non-targeting sequences (Figure 4) could flag that paper mills contribute to manuscript text, at least by providing text for the methods section. The repeated appearance of nucleotide sequence reagents could also suggest that these reagents are selected from internal databases of RNA interference reagents and PCR primers.

How 'theme and variations' manuscripts may evade detection. Particular features of the production of manuscripts in series could reduce the likelihood of their being detected through editorial practices and peer review (Figure 5).[33] Single-gene knockdown

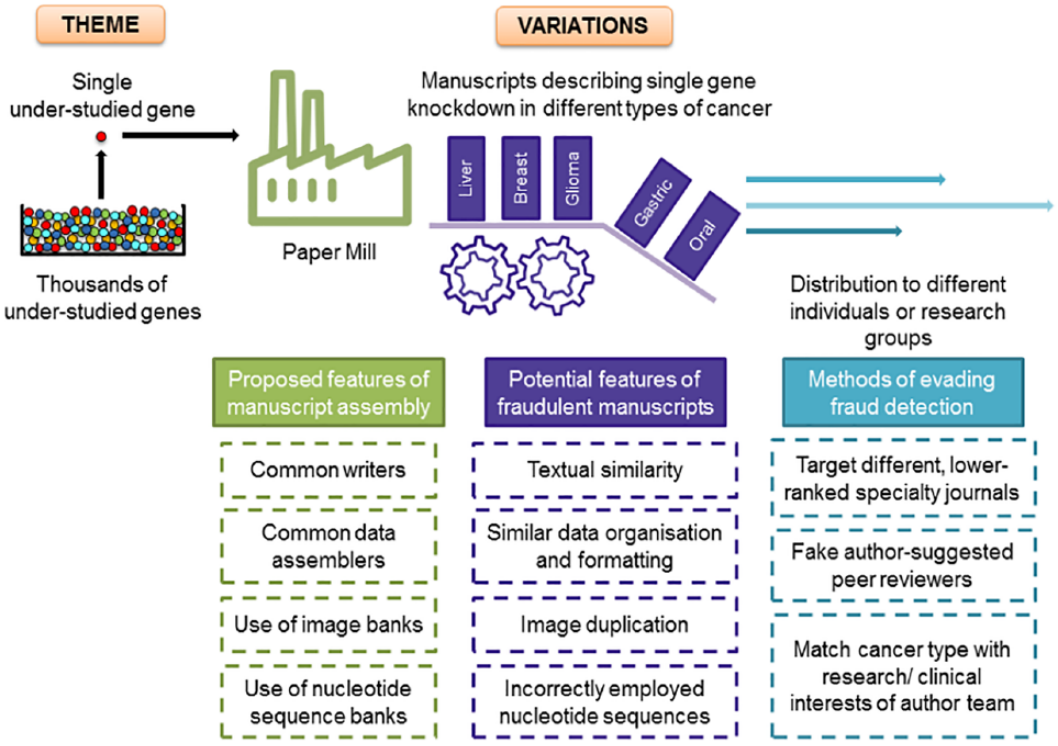

Figure 5. Overview of the proposed key features of the construction of fraudulent manuscript series by paper mills using a 'theme and variations' approach. The 'theme' shown is an under-studied human gene which is examined in different cancer types to produce a number of manuscript 'variations'. The existence of thousands of under-studied human genes means that this process could be repeated many times to produce large numbers of fraudulent manuscripts and ultimately publications.

papers describe widely used and accepted experimental approaches[33] and could therefore appear sound to peer reviewers focussing on the technical aspects of described experiments. As described above, targeting under-studied genes or biological processes may reduce the likelihood that specific content errors will be detected during peer review, due to the lack of available content expertise and/or perceived poor reviewing standards at targeted journals (Figure 5). Paper mills could provide content recipients with information about target journals with vulnerable manuscript screening and/or peer review standards, and paper mills have also been alleged to submit manuscripts on behalf of authors.[16] Paper mills directing manuscript submissions could be particularly advantageous in the case of series of related manuscripts. For example, a co-ordinated submission approach could ensure that most manuscripts constructed around a common gene are submitted to different journals.[37–41,52,59–62,64] Simultaneous construction and then submission of highly related manuscripts to different journals could also prevent unusual similarities from being detected during the peer review process.

The deliberate targeting of lower-ranked, specialty journals may present further advantages for the producers of fraudulent scientific content. Although most researchers want their papers to be read and cited, these could be secondary considerations for researchers attempting to meet career publication goals.[74] Hiding fraudulent publications within what has been termed 'the long tail' of research[75] might avoid attracting wide attention to individual publications, which could in turn attract attention to publication series as a whole. This would be further facilitated by these papers targeting under-studied genes, where a limited peer review community might equate to fewer readers. Publishing at least some members of publication series behind journal paywalls may also reduce their visibility to text mining approaches.

Theme and variations manuscripts as 'reverse salami slicing'. The proposed construction of manuscript series around different genes across multiple cancer types shares some features of what has been termed 'salami slicing' of publications.[20,76,77] This questionable practice involves dividing results that could be reported in a single manuscript across multiple smaller manuscripts and is believed to be encouraged by academic reward systems that favour publication quantities over quality.[76] The data described in these individual publications may be otherwise sound, although limited in scope.[77] In the approach that may be taken by paper mills, the lack of prior publications within a field could create the opportunity to generate many small, related studies using a 'theme and variations' model (Figure 5). Thus, rather than dividing a larger body of work into smaller publishable units, which is the model of conventional salami slicing,[20,76,77] paper mills may deliberately construct small but related papers around a common focal point, such as the function of a human gene in cancer.[33] As fraudulent manuscript series are likely to be constructed as multiple small

publishable units, as opposed to being divided into such, the fraudulent construction of manuscript series could be viewed as salami slicing in reverse.

In addition to possibly involving fraudulent data construction, reverse salami slicing would be expected to show other differences from conventional salami slicing.[20,76,77] When a larger body of work is divided into smaller manuscripts by one or more authors, the individual papers will be published by at least some shared authors.[20,76] In this way, papers generated through conventional salami slicing share both highly related topics and common authorship, and these combined attributes allow salami-sliced publications to be recognisable within the literature.[20,76] While this issue needs further research, our work to date suggests that single-gene knockdown papers that examine the same gene in different cancer types do not feature common authors.[33] This possibly fraudulent type of salami slicing is consistent with content being generated by a third party such as a paper mill and then distributed to different author groups, possibly according to their stated cancer research interests (Figure 5). For example, it would be more logical for paper mills to provide manuscripts concerning gene function in colorectal cancer to recipients from clinical departments that focus on gastrointestinal cancers, as opposed to other cancer types. This lack of shared authorship, combined with concordance between research topics and authors' departmental affiliations and past research interests, could render reverse salami slicing to be less easily detected (Figure 5).

Consequences for future biomarker research

The possible construction of fraudulent pre-clinical publications around individual under-studied human genes is of serious concern for several reasons. First, as there are some 20 000 protein-coding human genes alone,[26] most of which have been incompletely functionally annotated,[27–32] this creates literally thousands of opportunities to construct fraudulent manuscript series around the functions of individual genes in different cancer types or other biological systems. As we have described above, producing manuscripts in series may also allow fraudulent manuscripts to be constructed more efficiently and therefore in greater numbers at less cost.[33] Many aspects of manuscript construction, submission, and publication could also be tailored to reduce the probability of fraud detection. Indeed, the more fraudulent manuscripts are produced by a paper mill, the more important evading detection is likely to become.

Most concerningly from a scientific perspective, series of apparently independent yet possibly fraudulent reports have the clear potential to misdirect future research (Figure 2).[33] Large-scale content manufacture has occurred in other settings, notably through the SCIgen algorithm that creates nonsensical abstracts and manuscripts.[22,47] However, the lack of intrinsic sense in SCIGen publications means that they are largely inert contributions to the literature. In contrast, single-gene knockdown papers superficially resemble genuine studies and commonly report that their targeted genes are worthy of further, more clinically focussed research,[33] conclusions that may be unchallenged by other literature where under-studied genes are targeted. Therefore, despite the limited scope of individual single-gene knockdown papers, the broad recapitulation of their results across many different cancer types could reinforce their shared conclusions and encourage further research (Figure 2).

In the short term, fraudulent single-gene knockdown papers could provide biological evidence to support particular human genes entering the cancer biomarker research pipeline, possibly at the expense of superior candidates (Figure 2). As described above, multiple apparently independent studies that commonly show that a particular gene drives cancer phenotypes could favour the selection of such candidate genes for further analysis. Given the high existing rate of biomarker and drug development failure,[1-10] the biomarker research pipeline cannot afford to deal with the additional problem of fraudulent pre-clinical research contributing to weak biomarker candidates entering the pipeline at its earliest stage (Figures 1 and 2). If future investigations confirm the widespread existence of gene knockdown studies that contain paper mill–derived data, it will be vitally important to be able to reliably identify these publications, both to distinguish such papers from genuine contributions and to deter their future publication. The following sections will therefore consider possible approaches to detect and ultimately deter the systematic fraudulent production of manuscripts that target individual genes.

Proposed solutions

Improved awareness and peer review. There are numerous solutions that could help to combat the problem of fraudulent published research and the particular problem of fraudulent manuscripts produced in series. Research fraud is challenging to detect as unlike unintentional error, fraudulent practices are actively concealed by the parties involved.[17] However, the single-gene knockdown papers that have been described to date share numerous common features,[33] and unpublished data suggest that these features could be shared by many other papers. One approach is therefore to raise awareness of the features of single-gene knockdown publications, particularly among specialty journal editors and researchers who examine the functions of human genes. These individuals may be best placed to notice unusual manuscript or publication series, either through their exposure to large numbers of submitted manuscripts or through their detailed knowledge of the literature in their specific fields. Strategies to raise awareness could include publications describing the hallmarks of single-gene knockdown papers, further published research, the investigation of more individual papers by different journals and publishers, and the publication of Expressions of Concern[65–69] and retractions,[37–40,48–60] where appropriate. Awareness of the possibility of systematic research fraud could also be raised among peer reviewers by adding specific questions to manuscript

review checklists. Being asked to consider possible features of research fraud may allow peer reviewers to be more open to the possibility that manuscripts could describe falsified or fraudulent data.[17]

Beyond the similarities that have already been described (Figures 3 to 5), other features of single-gene knockdown papers could enable their detection. The possible simultaneous creation of manuscripts in series (Figure 5), which we have argued would favour efficiency,[33] could also mean that highly related published papers may not cite each other. As checking the contribution of a manuscript relative to existing publications is a broadly accepted component of peer review,[78] manuscripts with features of single-gene knockdown papers (Figures 3 to 5) that also do not cite highly relevant publications should represent yellow flags for editors and peer reviewers. These situations are particularly difficult to explain in the cases of under-studied genes, where failing to reference highly related papers cannot be excused by their being obscured by a wide body of literature.

Another feature of concern is the analysis of a gene in a particular cancer type where a very poor rationale is advanced for these analyses. Given the available amount of high-throughput data available from gene microarray, next-generation sequencing, and proteomics approaches, there is no shortage of data to cite in support of functional analyses of individual gene candidates.[43] It is therefore insufficient for a manuscript or publication to justify functional analyses simply because the gene of interest has never been studied in a particular cancer type. Pre-clinical and early biomarker studies that are not based on relevant, focussed preliminary data are very unlikely to generate biomarkers or pre-clinical research of any real-world value.[4,6] Pre-clinical study justifications should therefore be specific and tailored to either the predicted biological function of the gene of interest or outstanding issues relevant to individual cancer type(s). Journal editors, peer reviewers, and readers should therefore be sceptical of pre-clinical cancer biomarker manuscripts or publications that describe highly generic experiments that are poorly justified from either a biological or a clinical perspective.

Broader changes to editorial practices could also render manuscript and publication series more visible to journal editors and expert reviewers. The expanded use of central preprint servers could help to identify manuscripts that form part of existing or emerging publication series, recognising that opt-out possibilities would need to be restricted to prevent paper mills from simply avoiding such requirements. Automatic posting of submitted manuscripts to a centralised preprint server would also discourage the submission of the same manuscript to more than one journal, a tactic that may be used by paper mills to increase manuscript acceptance rates. Open and ongoing models of peer review may also allow publication series to be identified more effectively, particularly as awareness of fraudulent publication types grows.

Content error detection. Although the previous section raised some approaches to detect research manuscripts and papers that may include undeclared third-party data, these are unlikely to represent stand-alone solutions. It will take time to build editor and reviewer awareness of manuscripts that have been fraudulently produced and for journals to implement policies to deter their submission. Furthermore, changes in journal editorial policies may not permit the detection of all fraudulent papers that already exist within the literature.

The challenge of inventing entire data sets means that fraudulent manuscripts or papers often contain errors that can facilitate their detection.[17] As we have already described, paper mills are likely to take a number of steps to produce plausible data sets and manuscripts, including employing postdoctoral researchers as expert content producers.[16] Nonetheless, the efficient generation of large numbers of error-free manuscripts will remain very challenging. Most researchers will attest that detecting factual errors in manuscripts typically requires both highly specific content expertise and attention to detail, requirements which may be broadly incompatible with the rapid generation of scientific content by paper mills. As evidence, single-gene knockdown papers have been found to contain errors, such as the description of contaminated and misidentified human cell lines,[38] and incorrectly identified nucleotide sequence reagents (Figure 4).[33] The detection of these and other content errors may therefore facilitate the detection of papers and manuscripts that include content supplied by paper mills. Screening large numbers of manuscripts and publications will in turn require the application of semi-automated tools that have been designed to detect errors in research,[47,79–82] particularly as peer review has been shown to be an ineffective means of detecting errors in test manuscripts.[83,84] Semi-automated tools present advantages of higher throughput, and their use may also lead to the discovery of new features of concern.

To increase the scale and efficiency of manuscript production, we have proposed that paper mills may use images from internal image databases (Figure 5). This practice could result in inadvertent image duplication within and between studies, particularly if large numbers of manuscripts are produced by an individual paper mill. Indeed, image duplication was cited as a reason for the retraction of single-gene knockdown paper to date.[49] A number of tools have recently been described to detect image reuse,[79,80] which may be useful to detect papers produced with the assistance of paper mills, as well as other causes of image duplication. The development of improved algorithms that can reliably identify highly similar figures may also help to identify members of fraudulent publication series.[33]

A more prominent feature of single-gene knockdown papers is their description of incorrect nucleotide sequence reagents, which has characterised most of the highly similar gene knockdown papers that we have described (Figure 4).[33] The repeated description of incorrect nucleotide sequence reagents suggests that these reagents may also be selected from internal

databases. Although mismatches between nucleotide sequence reagents and their stated identities can be detected through cross-checking using search algorithms such as blastn,[46] checking such detailed information may be broadly incompatible with paper mill workflows.

A simple means to identify papers that have used common nucleotide sequence reagents is to employ nucleotide sequences as queries in Google Scholar searches.[33] However, this approach is limited to searching for incorrectly identified nucleotide sequence reagents whose identities are already known. A second limitation of this approach is the variable formatting of nucleotide sequences within publications, and we have previously been unable to detect known instances of incorrectly identified nucleotide sequence reagents through Google Scholar searches.[33] A semi-automated approach to detect incorrectly identified nucleotide sequence reagents would present advantages of both increased throughput and the capacity for knowledge discovery. We have therefore written the Seek and Blastn tool to facilitate the identification of publications where the described identity of a nucleotide sequence does not match its described experimental status (manuscript in revision).[85] This tool is also freely available for researchers to access and test.[86]

Future directions

The possibility that paper mills in China and possibly elsewhere are producing fraudulent manuscripts relevant to human health requires urgent and focussed attention. This review has focussed on the possibility that human protein-coding genes might be deliberately targeted for the production of series of manuscripts that describe gene function in different cancer types. However, there is no reason for paper mills to restrict their attention to protein-coding genes. There are also more than 20 000 non-coding human genes,[26] most of which could also be studied in different cancer types or diseases, and some single-gene knockdown papers have indeed examined non-coding human genes.[49,56] Other genetic features such as gene polymorphisms, which vastly exceed the numbers of both protein-coding and non-coding genes, as well as other entities such as chemotherapeutic drugs and natural products, could equally be examined in multiple different cancer types or biological systems. The mass retraction of publications due to manipulated peer review included publications that examined non-coding human genes, human coding gene polymorphisms, chemotherapeutic drugs, and natural products, all in human cancer types, as well as other publications that tested candidate protein biomarkers in human cancer patient cohorts.[24] Shared topics between papers that were retracted due to compromised peer review,[24] combined with reported attempts by paper mills to actively manipulate peer review,[16] suggest that paper mills may target other entities beyond human protein-coding genes. If such manuscripts were to be produced at scale, fraudulent publications examining non-coding human genes, gene polymorphisms, cancer drugs, and candidate biomarkers could also seriously misdirect cancer and biomarker research.

Investments will be required to define the scale of systematic fraud targeting under-studied genes and other entities, to clear the literature of misleading publications and deter their future publication. At present, improving the integrity of the research literature is not a high priority for biomedical funding agencies, perhaps due to an underestimation of the problem of systematic fraud. Funds will be required to accelerate the development and implementation of semi-automated screening methods and fact-checking tools. However, the testing and use of semi-automated tools require human support, so incentives need to be extended to researchers to test and use these tools. These incentives could include dedicated career support and (re-)training opportunities. Many researchers would be highly motivated to correct the biomedical literature if these activities were valued as highly as original research publications.

While an improved capacity to screen unpublished manuscripts will prevent erroneous research from being published, the broader use of screening tools will have other consequences. The application of literature screening tools could generate large backlogs of publications for journals and peer reviewers to carefully re-evaluate.[80] Increased numbers of papers requiring post-publication review will require drastic improvements to existing slow and time-consuming post-publication review processes.[21,87] The use of semi-automated screening tools could therefore require journals to devote many more resources to post-publication review, which could encourage journals to actively invest in tool development and testing. This could be further stimulated by biomedical research agencies offering industry partnership grants with scientific publishing companies, to support collaborations between academic researchers and journals. However, the widespread implementation of screening tools is also likely to also encourage paper mills to become progressively more adept at evading detection. For this reason, policy measures will also be required to reduce the systemic pressures to publish that have been proposed to drive the paper mill business model.[16,71-73]

On a more positive note, the recognition that under-studied genes and biological processes could be targeted for the production of fraudulent publications could also encourage more investment in basic research. The current focus on translational research places a greater priority on translating existing knowledge about gene function, as opposed to deriving new knowledge about the majority of human genes that have been studied to a limited extent.[29] Increased funding for discovery research to elucidate the functions of under-studied genes could accelerate a fuller understanding of the human genome, while also removing easy targets from being available for research fraud.

Summary and Conclusions

A key factor to improve the success rate of cancer biomarker translation is the more frequent selection of productive

biomarker candidates at the earliest stage of the biomarker research pipeline (Figure 2). Although it is widely accepted that incorrect published results lead to the selection of unproductive biomarker candidates,[5,6] this review has discussed a new threat to biomarker research, namely, the systematic targeting of under-studied human genes for the fraudulent construction of pre-clinical cancer research publications. This may involve paper mills providing research content or manuscripts describing the functions of under-studied human genes to researchers who publish these results without disclosing their origin. Such fraudulent publications both individually and collectively damage functional genomics research and therefore biomarker research, by generating incorrect information about gene function that appears to be relevant to human disease. This incorrect information will be most damaging where existing knowledge is limited (Figure 2), which will likely be the case for most of the genes that are deliberately targeted.

In summary, systematic research fraud threatens the ethos of science, by undermining the trust required to build on past achievements. Research is therefore urgently needed to transform our awareness of and capacity to both detect and deter fraudulently produced research manuscripts and publications that target under-studied genes. Policy changes are also required to remove unrealistic and ultimately damaging publication quotas, particularly for academics and medical doctors with limited time for research. By combining increased awareness of the problem of fraudulent publications with improved screening and detection methods, we can take meaningful steps to improve the quality of biomarker candidates that enter the biomarker research pipeline at its earliest stage.

Acknowledgements

The authors thank 2 anonymous reviewers for their helpful comments during the peer review process.

Author Contributions

JAB conceived of and designed the content and structure of the manuscript; JAB and NG drafted, revised, and edited the manuscript; AC-D and CL revised and edited the manuscript.

Disclosures and Ethics

The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material.

ORCID iDs

Jennifer A Byrne https://orcid.org/0000-0002-8923-0587

Natalie Grima https://orcid.org/0000-0001-6434-7706

Amanda Capes-Davis https://orcid.org/0000-0003-4184-6339

REFERENCES

1. Boutros PC. The path to routine use of genomic biomarkers in the cancer clinic. *Genome Res*. 2015;25:1508–1513.
2. de Gramont A, Watson S, Ellis LM, et al. Pragmatic issues in biomarker evaluation for targeted therapies in cancer. *Nat Rev Clin Oncol*. 2015;12:197–212.
3. Drucker E, Krapfenbauer K. Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. *EPMA J*. 2013;4:7.
4. Duffy MJ, Sturgeon CM, Sölétormos G, et al. Validation of new cancer biomarkers: a position statement from the European group on tumor markers. *Clin Chem*. 2015;61:809–820.
5. Kern SE. Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. *Cancer Res*. 2012;72:6097–6101.
6. Ioannidis JPA, Bossuyt PMM. Waste, leaks, and failures in the biomarker pipeline. *Clin Chem*. 2017;63:963–972.
7. Borrebaeck CA. Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. *Nat Rev Cancer*. 2017;17:199–204.
8. Bunnage ME. Getting pharmaceutical R&D back on target. *Nat Chem Biol*. 2011;7:335–339.
9. Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst*. 2001;93:1054–1061.
10. Mischak H, Ioannidis JP, Argiles A. Implementation of proteomic biomarkers: making it work. *Eur J Clin Invest*. 2012;42:1027–1036.
11. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials. *Evolution of Translational Omics: Lessons Learned and the Path Forward*. Washington, DC: Institute of Medicine of the National Academies; 2012.
12. Bornmann L. Research misconduct – definitions, manifestations and extent. *Publications*. 2013;1:87–98.
13. Fanelli D. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLoS ONE*. 2009;4:e5738.
14. Steen RG, Casadevall A, Fang FC. Why has the number of scientific retractions increased? *PLoS ONE*. 2013;8:e68397.
15. Ataie-Ashtiani B. World map of scientific misconduct. *Sci Eng Ethics*. 2018;24:1653–1656.
16. Liu X, Chen X. Journal retractions: some unique features of research misconduct in China. *J Scholar Pub*. 2018;49:305–319.
17. Stroebe W, Postmes T, Spears R. Scientific misconduct and the myth of self-correction in science. *Perspect Psychol Sci*. 2012;7:670–688.
18. Kornfeld DS. Perspective: research misconduct: the search for a remedy. *Acad Med*. 2012;87:877–882.
19. Tijdink JK, Bouter LM, Veldkamp CL, van de Ven PM, Wicherts JM, Smulders YM. Personality traits are associated with research misbehavior in Dutch scientists: a cross-sectional study. *PLoS ONE*. 2016;11:e0163251.
20. White C. Suspected research fraud: difficulties of getting at the truth. *BMJ*. 2005;331:281–288.
21. Smith R. Research misconduct: the poisoning of the well. *J R Soc Med*. 2006;99:232–237.
22. Djuric D. Penetrating the omerta of predatory publishing: the Romanian connection. *Sci Eng Ethics*. 2015;21:183–202.
23. Ferguson C, Marcus A, Oransky I. The peer-review scam. *Nature*. 2015;515:480–482.
24. Stigbrand T. Retraction note to multiple articles in Tumor Biology [published online ahead of print April 20, 2017]. *Tumor Biol*. doi:10.1007/s13277-017-5487-6.
25. Sims RR. Linking groupthink to unethical behavior in organizations. *J Bus Ethics*. 1992;11:651–662.
26. Pertea M, Shumate A, Pertea G, et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol*. 2018;19:208.
27. Hoffmann R, Valencia A. Life cycles of successful genes. *Trends Genet*. 2003;19:79–81.
28. Pfeiffer T, Hoffmann R. Temporal patterns of genes in scientific publications. *Proc Natl Acad Sci USA*. 2007;104:12052–12056.
29. Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH. Too many roads not taken. *Nature*. 2011;470:163–165.
30. Rodgers G, Austin C, Anderson J, et al. Glimmers in illuminating the druggable genome. *Nat Rev Drug Discov*. 2018;17:301–302.
31. Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. *Sci Rep*. 2018;8:1362.
32. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol*. 2018;16:e2006643.
33. Byrne JA, Labbé C. Striking similarities between publications from China describing single gene knockdown experiments in human cancer cell lines. *Scientometrics*. 2017;110:1471–1493.

34. Michels C, Schmoch U. The growth of science and database coverage. *Scientometrics*. 2012;93:831–846.
35. Siebert S, Machesky LM, Insall RH. Overflow in science and its implications for trust. *Elife*. 2015;4:e10825. doi:10.7554/eLife.10825.
36. Fox CW. Difficulty of recruiting reviewers predicts review scores and editorial decisions at six journals of ecology and evolution. *Scientometrics*. 2017;113:465–477.
37. Retraction. Lentivirus-mediated TPD52L2 depletion inhibits the proliferation of liver cancer cells in vitro. *Int J Clin Exp Med*. 2016;9:12416.
38. Retracted: knockdown of tumor protein D52-like 2 induces cell growth inhibition and apoptosis in oral squamous cell carcinoma. *Cell Biol Int*. 2016;40:361.
39. Retraction of: tumor protein D52-like 2 contributes to proliferation of breast cancer cells. *Cancer Biother Radiopharm*. 2017;32:387. doi:10.1089/cbr.2014.1723.
40. Retraction of: tumor protein D52-like 2 accelerates gastric cancer cell proliferation. *Cancer Biother Radiopharm*. 2017;32:388. doi:10.1089/cbr.2014.1766.
41. Wang Z, Sun J, Zhao Y, Guo W, Lv K, Zhang Q. Lentivirus-mediated knockdown of tumor protein D52-like 2 inhibits glioma cell proliferation. *Cell Mol Biol (Noisy-le-grand)*. 2014;60:39–44.
42. Rzhetsky A, Foster JG, Foster IT, Evans JA. Choosing experiments to accelerate collective discovery. *Proc Natl Acad Sci USA*. 2015;112:14569–14574.
43. Kaelin WG Jr. Common pitfalls in preclinical cancer target validation. *Nat Rev Cancer*. 2017;17:425–440.
44. Nourse CR, Mattei MG, Gunning P, Byrne JA. Cloning of a third member of the D52 gene family indicates alternative coding sequence usage in D52-like transcripts. *Biochim Biophys Acta*. 1998;1443:155–168.
45. Boutros R, Fanayan S, Shehata M, Byrne JA. The tumor protein D52 family: many pieces, many puzzles. *Biochem Biophys Res Commun*. 2004;325:1115–1121.
46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–410.
47. Labbé C, Labbé D. Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? *Scientometrics*. 2013;94:379–396.
48. Huang WY, Chen DH, Ning L, Wang LW. Retracted: siRNA mediated silencing of NIN1/RPN12 binding protein 1 homolog inhibits proliferation and growth of breast cancer cells. *Asian Pac J Cancer Prev*. 2017;18:2891.
49. Retracted: long noncoding RNA KIAA0125 potentiates cell migration and invasion in gallbladder cancer. *Biomed Res Int*. 2017;2017:3471417.
50. Xu R, Fang XH, Zhong P. Retraction notice to: myosin VI contributes to malignant proliferation of human glioma cells. *Korean J Physiol Pharmacol*. 2017;21:565.
51. Retraction: lentivirus-mediated knockdown of tectonic family member 1 inhibits medulloblastoma cell proliferation. *Int J Clin Exp Med*. 2018;11:2917.
52. Retraction: effect of prostaglandin reductase 1 (PTGR1) on gastric carcinoma using lentivirus-mediated system. *Int J Clin Exp Pathol*. 2018;11:1838.
53. Retraction: siRNA-mediated silencing of CDK8 inhibits proliferation and growth in breast cancer cells. *Int J Clin Exp Pathol*. 2018;11:1836.
54. Qi Y, Hu T, Lin K, Ye R, Ye Z. Retraction note to: lentivirus-mediated short-hairpin RNA targeting protein phosphatase 4 regulatory subunit 1 inhibits growth in breast cancer. *J Breast Cancer*. 2018;21:102.
55. Retraction. Down-regulation of GPR137 expression inhibits proliferation of colon cancer cells [published online ahead of print May 2, 2018]. *Acta Biochim Biophys Sin*. doi:10.1093/abbs/gmy057.
56. Retraction. Long non-coding RNA linc-ITGB1 knockdown inhibits cell migration and invasion in GBC-SD/M and GBC-SD gallbladder cancer cell lines. *Chem Biol Drug Des*. 2018;92:1815.
57. Retraction. si-RNA-mediated knockdown of PDLIM5 suppresses gastric cancer cell proliferation in vitro. *Chem Biol Drug Des*. 2018;92:2035.
58. Retracted: silencing of the COPS3 gene by siRNA reduces proliferation of lung cancer cells most likely via induction of cell cycle arrest and apoptosis. *Asian Pac J Cancer Prev*. 2017;18:2893.
59. Retracted: high expression of PTGR1 promotes NSCLC cell growth via positive regulation of cyclin-dependent protein kinase complex. *Biomed Res Int*. 2017;2017:7640820.
60. Retraction. TPD52L2 silencing inhibits lung cancer cell proliferation by G2/M phase arrest. *Int J Clin Exp Med*. 2018;11:413.
61. You W, Tan G, Sheng N, et al. Corrigendem. Downregulation of myosin VI reduced cell growth and increased apoptosis in human colorectal cancer. *Acta Biochim Biophys Sinica*. 2018;50:731.
62. Corrigendum. Lentivirus-mediated silencing of myosin VI inhibits proliferation and cell cycle progression in human lung cancer cells. *Chem Biol Drug Des*. 2018;92:1717.
63. Corrigendum. Downregulation of TPTE2P1 inhibits migration and invasion of gallbladder cancer cells. *Chem Biol Drug Des*. 2018;92:1816.
64. Corrigendum. Knockdown of myosin VI inhibits proliferation of hepatocellular carcinoma cells in vitro. *Chem Biol Drug Des*. 2018;92:1817.
65. Zhao X, Chen M, Tan J. Expression of concern to: knockdown of ZFR suppresses cell proliferation and invasion of human pancreatic cancer. *Biol Res*. 2018;51:20.
66. Pan Z, Pan H, Zhang J, et al. Expression of concern to: lentivirus mediated silencing of ubiquitin specific peptidase 39 inhibits cell proliferation of human hepatocellular carcinoma cells in vitro. *Biol Res*. 2018;51:19.
67. Liu X, Min L, Duan H. Expression of concern: short hairpin RNA (shRNA) of type 2 interleukin-1 receptor (IL1R2) inhibits the proliferation of human osteosarcoma U-2 OS cells. *Med Oncol*. 2018;35:129.
68. Lin Z, Xiong L, Lin Q. Expression of concern: knockdown of eIF3d inhibits cell proliferation through G2/M phase arrest in non-small cell lung cancer. *Med Oncol*. 2018;35:130.
69. Erratum: knockdown of immature colon carcinoma transcript-I inhibits proliferation of glioblastoma multiforme cells through Gap 2/mitotic phase arrest [expression of concern]. *Onco Targets Ther*. 2018;11:7601.
70. Huang W, Huang H, Wang L, Hu J, Song W. Retraction: SUN1 silencing inhibits cell growth through G0/G1 phase arrest in lung adenocarcinoma. *Onco Targets Ther*. 2017;10:2825–2833.
71. Hvistendahl M. China's publication bazaar. *Science*. 2013;342:1035–1039.
72. Lin S. Why serious academic fraud occurs in China. *Learned Pub*. 2013;26:24–27.
73. Tian M, Su Y, Ru X. Perish or publish in China: pressures on young Chinese scholars to publish in internationally indexed journals. *Publications*. 2016;4:9.
74. Brunette DM. Statistical and other methodological issues in the production of reliable experimental endodontic research and why they matter. *Endod Top*. 2016;34:8–29.
75. Bucchi M. Norms, competition and visibility in contemporary science: the legacy of Robert K. Merton. *J Classic Sociol*. 2015;15:233–252.
76. van Dalen H, Henkens K. Intended and unintended consequences of a publish-or-perish culture: a worldwide survey. *JASIS&T*. 2012;63:1282–1293.
77. Jackson D, Walter G, Daly J, Cleary M. Editorial: multiple outputs from single studies: acceptable division of findings vs. 'salami' slicing. *J Clin Nurs*. 2014;23:1–2.
78. Nicholas KA, Gordon W. A quick guide to writing a solid peer review. *EOS*. 2011;92:233–234.
79. Koppers L, Wormer H, Ickstadt K. Towards a systematic screening tool for quality assurance and semiautomatic fraud detection for images in the life sciences. *Sci Eng Ethics*. 2017;23:1113–1128.
80. Acuna DE, Brookes PS, Kording KP. Bioscience-scale automated detection of figure element reuse. *bioRxiv*. 2018:269415.
81. Nuijten MB, Hartgerink CHJ, van Assen MALM, Epskamp S, Wicherts JM. The prevalence of statistical reporting errors in psychology (1985-2013). *Behav Res Methods*. 2016;48:1205–1226.
82. Georgescu C, Wren JD. Algorithmic identification of discrepancies between published ratios and their reported confidence intervals and P-values. *Bioinformatics*. 2018;34:1758–1766.
83. Schroter S, Black N, Evans S, Godlee F, Osorio L, Smith R. What errors do peer reviewers detect, and does training improve their ability to detect them? *J R Soc Med*. 2008;101:507–514.
84. Cole GD, Shun-Shin MJ, Nowbar AN, et al. Difficulty in detecting discrepancies in a clinical trial report: 260-reader evaluation. *Int J Epidemiol*. 2015;44:862–869.
85. Phillips N. Tool spots DNA errors in papers. *Nature*. 2017;551:422–423.
86. http://scigendetection.imag.fr/TPD52/.
87. Allison DB, Brown AW, George BJ, Kaiser KA. Reproducibility: a tragedy of errors. *Nature*. 2016;530:27–29.