



HAL
open science

Identifying biomedical research papers with incorrect nucleotide sequence reagents using targeted and screening approaches

Rachael A. West, Guillaume Cabanac, Amanda Capes-Davis, Jennifer A. Byrne, Cyril Labbé

► To cite this version:

Rachael A. West, Guillaume Cabanac, Amanda Capes-Davis, Jennifer A. Byrne, Cyril Labbé. Identifying biomedical research papers with incorrect nucleotide sequence reagents using targeted and screening approaches. Australasian Interdisciplinary Meta-research & Open Science Conference (AIMOS 2019), Nov 2019, Melbourne, Australia. . hal-02976636

HAL Id: hal-02976636

<https://hal.science/hal-02976636>

Submitted on 23 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Identifying biomedical research papers with incorrect nucleotide sequence reagents using targeted and screening approaches

RA West^{1,2}, G Cabanac³, A Capes-Davis⁴, JA Byrne^{1,2}, C Labbé⁵

¹ The University of Sydney Children's Hospital at Westmead Clinical School, NSW, Australia, ² Kids Research, The Children's Hospital at Westmead, NSW, Australia, ³ Univ. Toulouse, Toulouse, France
⁴ CellBank Australia, children's Medical Research Institute and The University of Sydney, NSW, Australia. ⁵ Univ. Grenoble Alpes, Grenoble, France

1. Background

Nucleotide sequences are verifiable experimental reagents in biomedical publications. We have developed Seek and Blastn (SB) to verify the targeting or non-targeting status of published nucleotide sequences¹.

In this study, we aimed to apply SB to:

1. Identify and verify human gene knockdown publications that were predicted to contain nucleotide sequence errors
2. Screen all papers describing human nucleotide sequence reagents published in the journal *Gene* from 2007-2018

2. Methodology

Selecting Literature Corpora

Single Gene Knockdown (SGK) Corpus: Targeted approach

- SGK papers were previously defined as describing knockdown of one human gene in 1-2 human cancer cell lines²
- 17 human genes were chosen that were studied in at least 2 previous SGK papers^{1,2}
- Keyword searches were performed in PubMed and Google Scholar: "Gene of Interest" AND "cancer" AND "knockdown"
- No publication date restrictions

Gene Corpus: Screening approach

- All Gene papers from 2007-2018

SB Analysis

SGK Corpus

- Papers were downloaded in PDF form and zipped into a compressed file
- Compressed file was analysed by SB on 12th June 2019

Gene Corpus

- Papers were downloaded as PDFs from Elsevier in June 2019
- Compressed file was analysed by SB June 2019
- Non-human papers were excluded from SB analysis

Manual Verification

- Nucleotide sequences were extracted and manually verified using Blastn (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) and UCSC Human Genome Browser (<https://genome.ucsc.edu/>)
- Contaminated or misidentified human cell lines (labelled as 'problematic cell lines') were identified using the Cellosaurus database (<https://web.expasy.org/cellosaurus/>)

3. Seek and Blastn (SB)

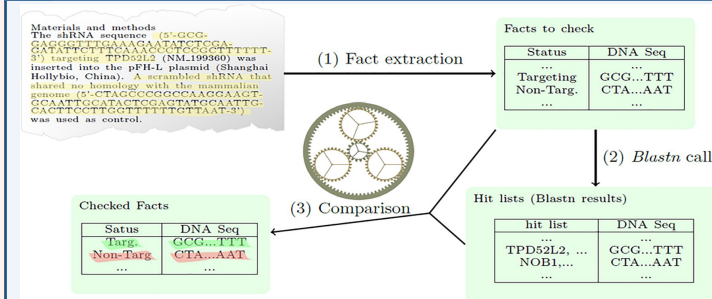


Figure 1: Key steps of the SB semi-automatic fact-checking tool. SB extracts nucleotide sequences and their claimed status (targeting or non-targeting) from text, performs a Blastn analysis and then fact-checks and reports whether the stated claim(s) are correct or incorrect. SB also identifies and reports cell line identifier(s) that correspond to contaminated or misidentified cell lines¹.

4. SGK and Gene Corpora Results

Table 1: Summary of papers with nucleotide sequence errors and/or problematic cell lines. The most frequent error type in both corpora was incorrect nucleotide sequences. Mainland China was the most common country of origin for papers with errors in both corpora.

	SGK Corpus	Gene Corpus
Total verifiable papers screened (n=)	174	873
Journals (n=)	85	1
Journal impact factor(s) (range/median)	0.181-6.854 (2.52)	2.638
Publication range (years)	2007-2019	2007-2018
Screened papers predicted to be erroneous (n=(%))	102/174 (59%)	300/873 (34%)
Papers with incorrect nucleotide sequences (n= (%))	87/102 (87%)	262/300 (87%)
Papers with problematic cell lines (n= (%))	27/102 (26%)	58/300 (19%)
Papers with incorrect nucleotide sequences and problematic cell lines (n= (%))	14/102 (14%)	22/300 (7%)
Other errors	2/102 (2%)	2/300 (1%)
Journals (n=(%))	57/85 (67%)	1
Top 2 countries of origin	China: 94/100 (94%) USA: 3/100 (3%)	China: 115/300 (38%) Iran: 20/300 (7%)

5. Gene Corpus Results



Figure 2: Countries of origin of n=300 Gene papers with nucleotide sequence errors and/or problematic cell lines. Word cloud created using <https://wordart.com/>

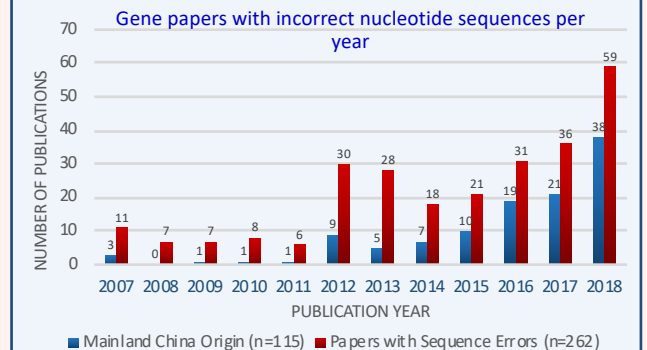


Figure 3: Distribution of n=262 Gene papers containing nucleotide sequence errors over a 12 year period (2007-2018). Numbers of publications from mainland China (blue bars) are compared to total numbers for each year (red bars).

6. Conclusions

- 59% SGK (102/174) and 34% Gene (300/873) papers contained errors
- Most frequent error was incorrect nucleotide sequence reagents
- More papers with errors were published by author teams from mainland China than any other country
- Single Gene Knockdown papers with errors were published in over 50 journals with a range of impact factors

References

1. Labbé C, Grima N, Gautier T, Favier B, Byrne JA. Semi-automated fact-checking of nucleotide sequence reagents in biomedical research publications: The Seek & Blastn tool. *PLOS ONE*. 2019; 14(3): e0213266
2. 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