Semi-automated fact-checking of nucleotide sequence reagents in biomedical research publications: the Seek & Blastn tool

Jennifer Byrne PhD
Children’s Cancer Research Unit, Kids Research
The University of Sydney, NSW

@JAByrneSci

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

Jennifer A. Byrne1,2 • Cyril Labbé3

Semi-automated fact-checking of nucleotide sequence reagents in biomedical research publications: The Seek & Blastn tool

Cyril Labbé1,*, Natalie Grima2, Thierry Gautier3, Bertrand Favier4, Jennifer A. Byrne2,5*

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

Jennifer A Byrne1,2id, Natalie Grima1id, Amanda Capes-Davis3id and Cyril Labbé4
What are genes and why do they matter?

• Human cells contain DNA, the instruction manual for the cell
• Instructions are written using nucleotides (A, C, G, T) in specific order
• ~3 billion nucleotides in the same order in each cell = the genome
• The genome is organised into functional units called genes
• Genes are understood by determining the sequence of nucleotides within DNA
• Gene sequences can’t be easily understood by eye (like barcodes), because of redundancy, and three ways to read some sequences

5’-GCGGAGGGTTTGAAAGAATATCTCGAGATATTCTTTTCA-3’
Verifiable experimental reagents

- **Reagents** are research ingredients
- Studying genes requires nucleotide sequences as reagents
- **Nucleotide sequences** can’t be read by eye
- Published nucleotide sequences need to be linked to a text description, to translate the DNA sequence into English
- Links between nucleotide sequences and descriptors can be verified
- **Nucleotide sequences** represent verifiable reagents
Original Article

**Lentivirus-mediated TPD52L2 depletion inhibits the proliferation of liver cancer cells in vitro**

Construction of lentiviral vector expressing TPD52L2-specific shRNA

A candidate short hairpin RNA (shRNA) was screened and validated to be target sequence 5’-GCGGAGGGTTTGAAAGAATATCTCGAGATATTCTTTCAACCCCTCCGCTTTTTT-3’, sequence 1) against human TPD52L2 gene (NM_199360). And the negative control siRNA was 5’-GCGGAGGGTTTGAAAGAATATCTCGAGATATCTTTCAACCCCTCCGCTTTTTT-3’. The stem-loop-stem oligos (shRNAs) were synthesized, annealed, and ligated into the *NheI/Pacl*-linearized shRNA vector *pFH-L* (Shanghai Hollybio, China). The lentiviral-based shRNA-expressing vectors were confirmed by DNA sequencing. The generated plasmids were named as *pFH-L-shTPD52L2* and *pFH-L-shCon*. 
Four *TPD52*-like genes

![Diagram showing four genes: TPD52 at 8q21, L1 at 6q22, L2 at 20q13, and L3 at 9p24.](image)
Early concerns (2015)

• *TPD52L2* is an under-studied gene- 19 publications in PubMed since 1998 (~1 publication per year)
  – 7/19 *TPD52L2* publications from my laboratory (including one published in 2014)

• In 2014-2015, 5 *TPD52L2* publications appeared in <1 year

• These 5 publications showed numerous unexpected similarities
  – All gene knockdown studies in cancer cell lines
  – All from authors based in China
Tumor Protein D52-Like 2 Contributes to Proliferation of Breast Cancer Cells

Mei Yang, Xueyao Wang, Jiaoyuan Jia, Hongwen Gao, Peng Chen, Xianliang Sha, and Shan Wu

Tumor Protein D52-Like 2 Accelerates Gastric Cancer Cell Proliferation In Vitro

Jiapeng Xu, Weimin Wang, Zhenxin Zhu, Ziran Wei, Dejun Yang, and Qingping Cai
Knockdown of tumor protein D52-like 2 induces cell growth inhibition and apoptosis in oral squamous cell carcinoma

Yongchun He, Fengshan Chen*, Ying Cai and Sihui Chen

Department of Orthodontics, Laboratory of Oral Biomedical Science and Translational Medicine, School of Stomatology, Tongji University, Shanghai 200072, P. R. China

Original Article

Lentivirus-mediated TPD52L2 depletion inhibits the proliferation of liver cancer cells in vitro

Ze-Ya Pan1, Yun Yang2*, Hao Pan2, Jin Zhang1, Hui Liu1, Yuan Yang1, Gang Huang1, Lei Yin1, Jian Huang1, Wei-Ping Zhou3

Lentivirus-Mediated Knockdown of Tumor Protein D52-like 2 Inhibits Glioma Cell Proliferation

Z. Wang1, J. Sun2*, Y. Zhao2, W. Guo2, K. Lv2 and Q. Zhang2*
### Conserved structure of TPD52L2 papers

<table>
<thead>
<tr>
<th>Figure</th>
<th>Panel A</th>
<th>Panel B</th>
<th>Panel C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fig 1</strong></td>
<td>Intracellular GFP detection, microscopic cell images</td>
<td>RT-PCR, graph</td>
<td>Western blot analysis, images of detected bands</td>
</tr>
<tr>
<td></td>
<td>Confirmation of shRNA transfection</td>
<td>Confirmation of gene knockdown at transcript level</td>
<td>Confirmation of gene knockdown at protein level</td>
</tr>
<tr>
<td><strong>Fig 2</strong></td>
<td>Cell proliferation assays, graph</td>
<td>Cell staining, microscopic cell images</td>
<td>Anchorage independent growth assays, graph</td>
</tr>
<tr>
<td></td>
<td>Reduced cell proliferation</td>
<td>Visualisation of gene knockdown phenotypes</td>
<td>Reduced anchorage independent growth</td>
</tr>
<tr>
<td><strong>Fig 3</strong></td>
<td>FACS cell cycle plots</td>
<td>FACS cell cycle graph</td>
<td>FACS cell cycle graph</td>
</tr>
<tr>
<td></td>
<td>Altered cell cycle distribution</td>
<td>Altered cell cycle distribution</td>
<td>Increased proportion of cells in sub-G1 cell cycle phase</td>
</tr>
</tbody>
</table>
TPD52L2 in oral carcinoma, He et al. 2015
**TPD52L2 in breast carcinoma, Yang et al. 2015**

**A**

- Graph showing the OD at 595 nm over time (days) for ZR-75-30 cells
  - Con
  - Lv-shCon
  - Lv-shTPD52L2

**B**

- Images of cells stained with Crystal violet, fluorescent field, and six-well plate
  - Con
  - Lv-shCon
  - Lv-shTPD52L2
Original Article

Lentivirus-mediated TPD52L2 depletion inhibits the proliferation of liver cancer cells in vitro

Construction of lentiviral vector expressing TPD52L2-specific shRNA

A candidate short hairpin RNA (shRNA) was screened and validated to be target sequence 5'-GCAGGAGGTTTGGAAAGATATCTCGAGATATTCTTTCAACCTCCGCTTTTTTT-3' (sequence 1) against human TPD52L2 gene (NM_199360). And the negative control siRNA was 5'-GCAGGAGGTTTGGAAAGATATCTCGAGATATTCTTTCAACCTCCGCTTTTTTT-3'. The stem-loop-stem oligos (shRNAs) were synthesized, annealed, and ligated into the NheI/Pacl-linearized shRNA vector pFH-L (Shanghai Hollybio, China). The lentiviral-based shRNA-expressing vectors were confirmed by DNA sequencing. The generated plasmids were named as pFH-L-shTPD52L2 and pFH-L-shCon.
“Series” of single gene knockdown papers

TPD52L2

lungs
oral

brain

gastric

liver

breast
“Publication series” for TPD52L2

TPD52L2

R

R

oral

R

gastric

R

liver

R

breast

R

brain

lung
Features of single gene knockdown papers

• Study functions of single genes in 1-2 human cancer cell lines
• Same or recurring experiment types, presented in same order
• High levels of textual similarity
• Can feature incorrect/ contaminated/ misidentified cell lines
• Often feature incorrect nucleotide sequence reagents
• Commonly conclude that the gene is important in the cancer type in question, worth further study
Incorrect nucleotide sequences - different types of hidden errors

Error types in published nucleotide sequences

Typographic (spelling) errors
- Nucleotide additions
- Nucleotide deletions
- Nucleotide substitutions

Mistaken identity
- Targeting but for different gene
  - Irrelevant data
- “Non-targeting” but with predicted target
  - Failed negative controls
- “Targeting” but predicted to be non-targeting
  - Impossible data
Incorrect nucleotide sequence reagents classified according to error type

<table>
<thead>
<tr>
<th>Nucleotide sequence reagent error type</th>
<th>Consequence of error type</th>
<th>Proportion (%) of incorrect nucleotide sequence reagents (PLOS ONE, n=95 reagents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeting sequence predicted to target different human gene</td>
<td>Data generated for different gene</td>
<td>43/95 (45%)</td>
</tr>
<tr>
<td>“Non-targeting” sequence predicted to target human gene</td>
<td>Incorrect negative control experiments</td>
<td>8/95 (8%)</td>
</tr>
<tr>
<td>“Targeting” sequence predicted to be non-targeting</td>
<td>Impossible/inexplicable results</td>
<td>44/95 (46%)</td>
</tr>
</tbody>
</table>
Semi-automated fact checking- the Seek & Blastn tool

(i) Pdf upload to http://scigendetection.imag.fr/TPD52/
(ii) Text extraction and formatting

Extraction of:
(i) nucleotide sequences
(ii) claimed status
(iii) gene identifiers
(iv) species names
(v) misidentified/contaminated cell line identifiers

(i) Textual similarity analysis (interertextual distance)
(ii) Blastn analysis
(iii) Google Scholar search

(i) Fact checking (blastn results versus claimed description)
(ii) Reporting

Labbé et al. 2019 PLOS ONE
Semi-automated fact checking - the Seek & Blastn tool

(i) Pdf upload to http://scigendetection.imag.fr/TPD52/
(ii) Text extraction and formatting

Extraction of:
(i) nucleotide sequences
(ii) claimed status
(iii) gene identifiers
(iv) species names
(v) misidentified/contaminated cell line identifiers

(i) Textual similarity analysis (intertextual distance)
(ii) Blastn analysis
(iii) Google Scholar search

(i) Fact checking (blastn results versus claimed description)
(ii) Reporting

Labbé et al. 2019 PLOS ONE
### Comparing Seek & Blastn versions

#### Nucleotide sequence metrics

34 papers published in Gene, n=413 nucleotide sequences

<table>
<thead>
<tr>
<th>Nucleotide sequence and targeting/ non-targeting status extraction</th>
<th>Seek &amp; Blastn version</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S&amp;B (PLOS ONE)</td>
<td>New S&amp;B</td>
</tr>
<tr>
<td>Correct sequence/ status extraction</td>
<td>274/413 (66.6%)</td>
<td>358/413 (86.7%)</td>
</tr>
<tr>
<td>Incorrect sequence/ status extraction</td>
<td>139/413 (33.4%)</td>
<td>55/413 (13.3%)</td>
</tr>
<tr>
<td>Incorrect sequence extraction</td>
<td>60/413 (14.5%)</td>
<td>35/413 (8.5%)</td>
</tr>
<tr>
<td>Incorrect gene identification</td>
<td>285/413 (69.0%)</td>
<td>110/413 (26.6%)</td>
</tr>
</tbody>
</table>
Comparing Seek & Blastn versions

Publication metrics

267 papers (n=126 with incorrect nucleotide sequences)

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Seek &amp; Blastn version</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S&amp;B (PLOS ONE)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>108/126 (86%)</td>
</tr>
<tr>
<td>Specificity</td>
<td>86/141 (61%)</td>
</tr>
<tr>
<td>Precision</td>
<td>108/163 (66%)</td>
</tr>
<tr>
<td>False discovery rate</td>
<td>55/163 (34%)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>194/267 (73%)</td>
</tr>
</tbody>
</table>
Applying Seek & Blastn-
Single Gene Knockdown papers

- Analysed 17 human genes featuring in ≥2 flagged papers (PLOS One, 2019)
- Performed PubMed and Google Scholar searches using gene identifiers and key words
- Identified 162 papers, published in 75 journals
  - 3-20 papers per gene
- 32/162 (20%) incorrect cell lines
- 92/162 (57%) incorrect nucleotide sequences
  - 129 incorrect reagents, 76 unique reagents

<table>
<thead>
<tr>
<th>Nucleotide sequence reagent error type</th>
<th>Consequence of error type</th>
<th>Proportion (%) of incorrect nucleotide sequence reagents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeting sequence predicted to target different human gene</td>
<td>Data generated for different gene</td>
<td>37/76 (49%)</td>
</tr>
<tr>
<td>“Non-targeting” sequence predicted to target human gene</td>
<td>Incorrect negative control experiments</td>
<td>19/76 (25%)</td>
</tr>
<tr>
<td>“Targeting” sequence predicted to be non-targeting</td>
<td>Impossible/ inexplicable results</td>
<td>20/76 (26%)</td>
</tr>
</tbody>
</table>
“Publication series” for NOB1

15 papers in 10 journals
Applying Seek & Blastn-Journal screening approach

- Downloaded papers published in Gene 2014-2018
- S&B extracted nucleotide sequences or flagged incorrect cell lines in 559 papers
- 299/559 (53%) papers flagged for incorrect nucleotide sequences
- Manually verified 91 papers
  - 49/91 (54%) papers had incorrect nucleotide sequences
  - 231 incorrect reagents, 220 unique reagents

<table>
<thead>
<tr>
<th>Nucleotide sequence reagent error type</th>
<th>Consequence of error type</th>
<th>Proportion (%) of incorrect nucleotide sequence reagents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeting sequence predicted to target different human gene</td>
<td>Data generated for different gene</td>
<td>133/220 (60%)</td>
</tr>
<tr>
<td>“Non-targeting” sequence predicted to target human gene</td>
<td>Incorrect negative control experiments</td>
<td>5/220 (3%)</td>
</tr>
<tr>
<td>“Targeting” sequence predicted to be non-targeting</td>
<td>Impossible/ inexplicable results</td>
<td>82/220 (37%)</td>
</tr>
</tbody>
</table>
Most human genes remain under-studied

~90% of the genes that make protein
~18,000 genes

~90% of the genes that make RNA
~19,000 genes
Why should we care???

• Incorrect sequence and cell line reagents cast serious doubts regarding the validity of reported research

• Multiple “independent” reports may encourage others to either replicate or extend such findings
  – *Single knockdown papers encourage translational research*

• At worst, may compromise patient safety, if only by stopping/ delaying more promising research

• At best, wasted time and resources
  – Pursuit of unproductive gene candidates
  – Overestimation of knowledge of gene function
  – Distortion of results from text mining
Incorrect nucleotide sequences occur in other publication types

- Combination of non-coding + protein-coding genes
- Gene variations in 20,000 protein-coding genes
- Naturally occurring or synthetic drugs + single genes
Thank-you to

Université Grenoble Alpes, France
Dr Bertrand Favier
Dr Thierry Gautier
Dr Cyril Labbé

University of Sydney, Australia
Dr Amanda Capes Davis
Ms Natalie Grima
Ms Rachael West

Université Toulouse, France
A/Professor Guillaume Cabanac

Northwestern University, USA
Dr Thomas Stoeger
Dr Martin Gerlach
Professor Luis A.N. Amaral

Journal editors and peer reviewers

kidsresearch.org.au
Children’s Cancer Research Unit (CCRU)
“Publication series” for ICT1 and MYO6

ICT1

leukemia

gastric

prostate

lymphoma

breast

brain

EoC

7 papers in 7 journals

MYO6

oral

colon

gastric

liver

brain

prostate

breast

8 papers in 7 journals