Fact-checking of nucleotide sequences in life science publications: the *seek & blastn* tool

Jennifer Byrne & Cyril Labbé

`@JAByrneSci, jennifer.byrne@health.nsw.gov.au, Cyril.Labbe@imag.fr`

Potential conflicts of interest and financial disclosures : J. A. Byrne has no conflicts of interest or financial disclosures to declare. Springer-Nature is funding a PhD student within the research group of C. Labbé. This PhD project is exploring methods to detect automatically generated scientific papers. Funding from Springer-Nature did not support the work described in this abstract.

UNIVERSITÉ Grenoble Alpes

Eighth International Congress on Peer Review and Scientific Publication

Enhancing the quality and credibility of science

September 10–12, 2017 | Chicago, USA

THE UNIVERSITY OF SYDNEY

Automatic detection of questionable research papers

Scientific ethics

- Plagiarism, auto-plagiarism, content reuse...
- $N - grams$ signature (hashing functions).

Non-sense detection

- Paper generator (SCIgen, physic-gen, MathGen...)
- Authorship detection (inter-textual distance).

Need to detect questionable scientific results

- Fabrications (making up data or results)
- Falsification (manipulating data or results)
- False or unsupported affirmations
- Genuine errors

$\Longrightarrow$

- Error spreading
- Wrong belief
- Research irreproducibility

Starting point : striking similarities, obvious errors

Jennifer Byrne:

- First reported *TPD52L2* (20 years ago)
- 5 Publications with obvious errors!

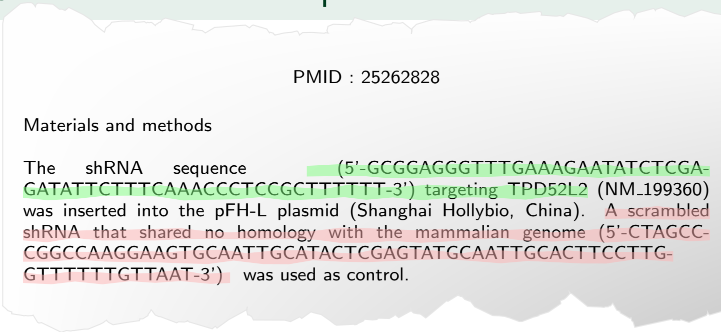
5 Publications from China:

- Single gene knockdown experiments.
- Human cancer cell lines.

Conclusions highlight potential therapy

- ...TPD52L2... novel therapeutic target for glioma treatment.
- ...TPD52L2... novel clues for oral squamous cell carcinoma therapy.
- ...TPD52L2... therapeutic approach for the treatment of breast cancer.
- ...TPD52L2 is indispensable in gastric cancer proliferation.
- ...TPD52L2 could be a novel therapeutic target for human liver cancer.

Obvious errors: example

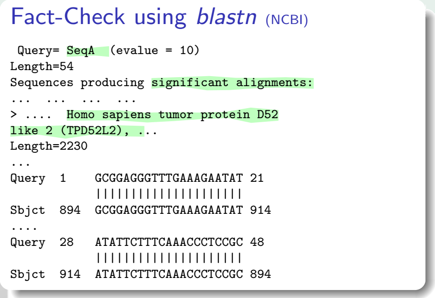


PMID : 25262828

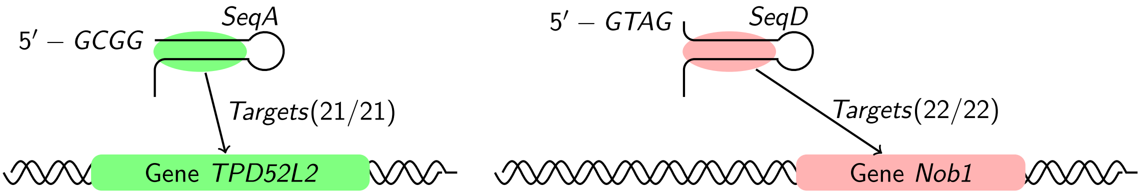
Materials and methods

The shRNA sequence (5'-GCGGAGGGTTTGAAAGAATATCTCGA-GATATTCTTTCAAACCCTCCGCTTTTTT-3') targeting TPD52L2 (NM_199360) was inserted into the pFH-L plasmid (Shanghai Hollybio, China). A scrambled shRNA that shared no homology with the mammalian genome (5'-CTAGCC-CGGCCAAGGAAGTGCAATTGCATACTCGAGTATGCAATTGCACTTCCTTG-GTTTTTTGTTAAT-3') was used as control.

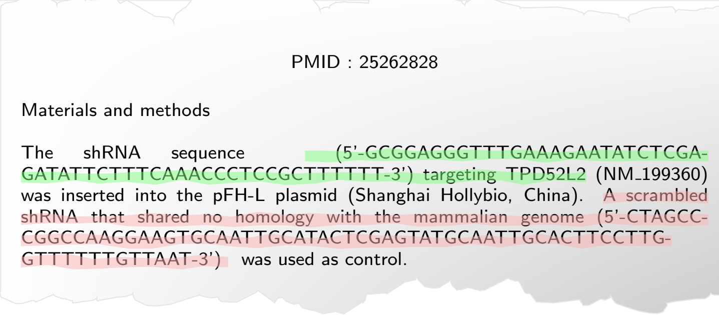
Fact-Check using *blastn* (NCBI)



```
 Query= SeqA  (evalue = 10)
Length=54
Sequences producing significant alignments:
...  ...  ...  ...
> ....  Homo sapiens tumor protein D52
like 2 (TPD52L2), ...
Length=2230
...
Query  1    GCGGAGGGTTTGAAAGAATAT 21
            |||||||||||||||||||||
Sbjct  894  GCGGAGGGTTTGAAAGAATAT 914
....
Query  28   ATATTCTTTCAAACCCTCCGC 48
            |||||||||||||||||||||
Sbjct  914  ATATTCTTTCAAACCCTCCGC 894
```



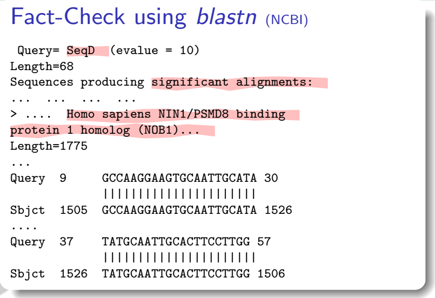
Obvious errors: example



PMID : 25262828

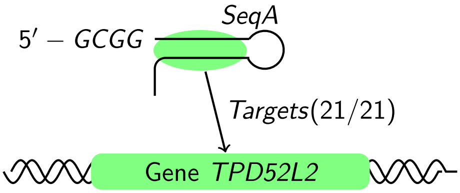
Materials and methods

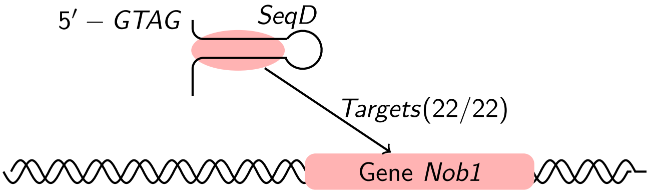
The shRNA sequence (5'-GCGGAGGGTTTGAAAGAATATCTCGAGATATTCTTTCAAACCCTCCGCTTTTTT-3') targeting TPD52L2 (NM_199360) was inserted into the pFH-L plasmid (Shanghai Hollybio, China). A scrambled shRNA that shared no homology with the mammalian genome (5'-CTAGCCCGGCCAAGGAAGTGCAATTGCATACTCGAGTATGCAATTGCACTTCCTTGGTTTTTTGTTAAT-3') was used as control.



Fact-Check using *blastn* (NCBI)

```
 Query= SeqD  (evalue = 10)
Length=68
Sequences producing significant alignments:
...  ...  ...  ...
> ....  Homo sapiens NIN1/PSMD8 binding
protein 1 homolog (NOB1)...
Length=1775
...
Query  9     GCCAAGGAAGTGCAATTGCATA 30
             ||||||||||||||||||||||
Sbjct  1505  GCCAAGGAAGTGCAATTGCATA 1526
....
Query  37    TATGCAATTGCACTTCCTTGG 57
             |||||||||||||||||||||
Sbjct  1526  TATGCAATTGCACTTCCTTGG 1506
```





Nucleotide sequence by *Status* (targeting vs non-targeting)

Targeting

Primers:

Two sets of primers were used for PCR: β-actin (ACTB) forward, 5'-GTGG...AGAC-3' and reverse, 5'-AAAG...AACTA-3'; NOB1 forward, 5'-GAAAG....TGGAG-3' and reverse, 5'-CAGCCTTGAGATGACCTAAGC-3'.

Silencing:

shRNA targeting the NOB1 (CCGGGCTGAACA...TTGTTCAGCTTTTTG).

Positive control:

A NOB1 positive control (5'-CCG...TT-3') was used ...

Non-Targeting

Negative control:

... and negative control (TTCTC...CACGT) sequences were cloned into...

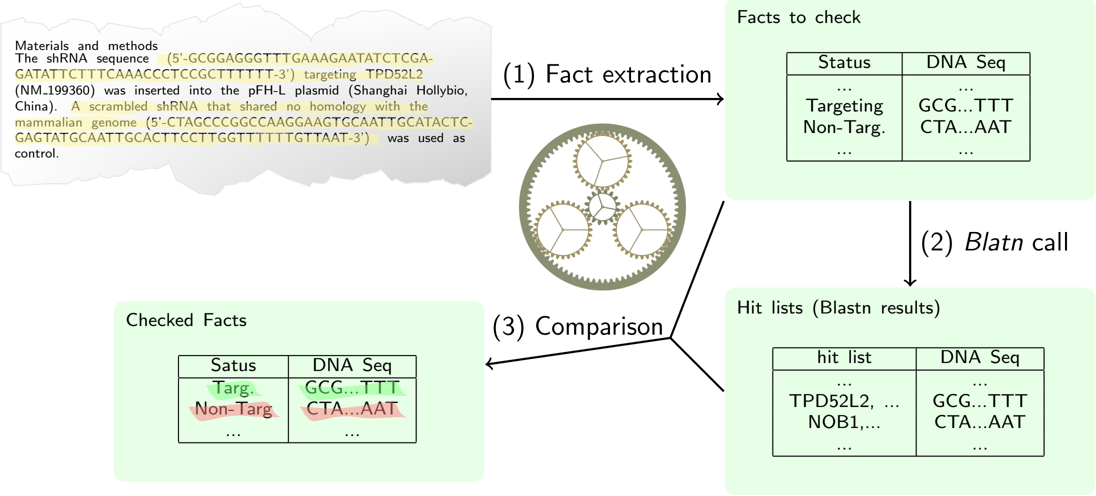
Non-targeting:

Non-targeting shRNA sequence (5-CTAGCC ... TTGTTAAT-3) was used as a control.

Scrambled:

A scrambled sequence (5'-GCGGA ... CTTTTTT-3') that has no significant homology with human gene sequences was used as a negative control.

Seek & Blastn at a glance



Materials and methods
The shRNA sequence (5'-GCGGAGGGTTTGAAAGAATATCTCGA-GATATTCTTTCAAACCCTCCGCTTTTTT-3') targeting TPD52L2 (NM_199360) was inserted into the pFH-L plasmid (Shanghai Hollybio, China). A scrambled shRNA that shared no homology with the mammalian genome (5'-CTAGCCCGGCCAAGGAAGTGCAATTGCATACTC-GAGTATGCAATTGCACTTCCTTGGTTTTTTGTTAAT-3') was used as control.

(1) Fact extraction

Facts to check

Status	DNA Seq
...	...
Targeting	GCG...TTT
Non-Targ.	CTA...AAT
...	...

(2) *Blatn* call

Hit lists (Blastn results)

hit list	DNA Seq
...	...
TPD52L2, ...	GCG...TTT
NOB1,...	CTA...AAT
...	...

(3) Comparison

Checked Facts

Satus	DNA Seq
Targ.	GCG...TTT
Non-Targ	CTA...AAT
...	...

Seek & Blastn steps

(1) Fact extractions

Named entity recognition techniques (thesaurus and rules):

- identifies gene names, contaminated cell lines.

Sequences containing DNA sequences are analyzed (*Finite-state machines*):

- extract nucleotide sequences (15-90 bases),
- assign a status *targeting* or *non-targeting*.

(2) Blastn call

NCBI software gives the hit list for each sequence.

(3) Blastn analysis vs text-extracted information

Set of rules to check whether or not Blastn results are compatible with affirmation detected in the text.

Used Corpora

Problematic Papers (CorpusP)

- A cohort of highly similar cancer research publications.
- 38/48 (79%) included nucleotide sequence(s) that did not match their experimental use (according to *blastn*).

Unknown papers (CorpusU)

154 papers, automatically retrieved using papers from CorpusP and the "PubMed similar" functionality. Mostly open-access[a].

[a] because when fee-based, automatically download is not permitted

Tests and results

Seek & Blastn performances

- In CorpusP and CorpusU, nucleotide sequences were extracted from 48/48 (100%) and 111/154 (73%) papers.
- Claims were not (correctly) identified for 19/341 (5.6%) sequences in CorpusP.
- Identification of the 38/48 (79%) papers in CorpusP incorrectly use nucleotide sequence.

Error detection in scientific literature

- 38/48 (79%) papers in CorpusP appear to have incorrectly employed nucleotide sequence.
- *Seek & Blastn* predicted that 30/154 (19%) CorpusU papers may have incorrectly employed nucleotide sequence reagent(s) but roughly half of them are.

Results suggest ...

that in addition to the "knock down" series, there may exists a "migration series", a "prognosis series", ...

Conclusion

Automatic detection, related works

- Detection of statistically flawed papers
- Fake news detection

Seek & Blastn perspectives

- Online tool : `http://scigendetection.imag.fr/TPD52`
- Avoid false positive.
- Tests of more in-depth analysis of sentences.

How Seek & Blastn could be use

- Pre- and post-publication checking
- Contribute to publishing guidelines
    - Inclusion of sequences within publications, Nucleotide sequence formatting
- Identification of other forms of misconduct