Jan 19, 2021

Seek & Blastn Standard Operating Procedure

Jennifer A. Byrne[1,2], Yasunori Park[2], Amanda Capes-Davis[2,3], Bertrand Favier[4], Guillaume Cabanac[5], Cyril Labbé[6]

[1]New South Wales Health Statewide Biobank, New South Wales Health Pathology, Camperdown, New South Wales, Australia;

[2]Faculty of Medicine and Health, The University of Sydney, New South Wales, Australia;

[3]CellBank Australia, Children's Medical Research Institute, Westmead, New South Wales, Australia;

[4]Univ. Grenoble Alpes, Team GREPI, Etablissement Français du Sang, EA 7408, BP35, La Tronche, France;

[5]Computer Science Department, IRIT UMR 5505 CNRS, University of Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 9, France;

[6]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France

1 Works for me

dx.doi.org/10.17504/protocols.io.bjhpkj5n

Yasunori Park

Citation: Jennifer A. Byrne, Yasunori Park, Amanda Capes-Davis, Bertrand Favier, Guillaume Cabanac, Cyril LabbÃÂ© (01/19/2021). Seek & Blastn Standard Operating Procedure. https://dx.doi.org/10.17504/protocols.io.bjhpkj5n

ABSTRACT

Seek & Blastn is a semi-automated tool that automatically extracts gene identifiers and nucleotide sequences using named entity recognition techniques. The sentence containing each sequence is automatically analyzed to assign a claimed status (targeting or nontargeting) that is compared with the most likely status according to blastn analysis using the human genomic + transcript database. Subsequently, this tool can be used to identify nucleotide sequence reagent errors in the published or forthcoming scientific literature.

The Seek & Blastn tool was developed by Cyril Labbé from the University of Grenoble Alpes, The National Center for Scientific Research (CNRS), Grenoble, France.

The authors gratefully acknowledge grant funding from the US Office of Research Integrity, grant ID ORIIR180038-01-00 (awarded to Jennifer Byrne and Cyril Labbé) and the National Health and Medical Research Council of Australia, Ideas grant ID APP1184263 (awarded to Jennifer Byrne, Cyril Labbé and Amanda Capes-Davis).

To cite Seek & Blastn or to find further information regarding the tool, please refer to:

- 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 Mar 1;14(3):e0213266. doi: 10.1371/journal.pone.0213266. PMID: 30822319; PMCID: PMC6396917.

For further information regarding this tool:

- Byrne, J.A., Labbé, C. Striking similarities between publications from China describing single gene knockdown experiments in human cancer cell lines. *Scientometrics* **110,** 1471–1493 (2017). https://doi.org/10.1007/s11192-016-2209-6

- Phillips N. Online software spots genetic errors in cancer papers. Nature. 2017 Nov 20;551(7681):422-423. doi: 10.1038/nature.2017.23003. PMID: 29168818.

- Byrne JA, Grima N, Capes-Davis A, Labbé C. The Possibility of Systematic Research Fraud Targeting Under-Studied Human Genes: Causes, Consequences, and Potential Solutions. Biomark Insights. 2019 Feb 5;14:1177271919829162. doi: 10.1177/1177271919829162. PMID: 30783377; PMCID: PMC6366001.

- Byrne JA, Christopher J. Digital magic, or the dark arts of the 21st century-how can journals and peer reviewers detect manuscripts and publications from paper mills? FEBS Lett. 2020 Feb;594(4):583-589. doi: 10.1002/1873-3468.13747. Epub 2020 Feb 17. PMID: 32067229.

- Labbé, C., Cabanac, G., West, R.A. *et al.* Flagging incorrect nucleotide sequence reagents in biomedical papers: To what extent does the leading publication format impede automatic error detection?. *Scientometrics* **124,** 1139–1156 (2020). https://doi.org/10.1007/s11192-020-03463-z

DOI

dx.doi.org/10.17504/protocols.io.bjhpkj5n

PROTOCOL CITATION

Jennifer A. Byrne, Yasunori Park, Amanda Capes-Davis, Bertrand Favier, Guillaume Cabanac, Cyril Labbé 2021. Seek & Blastn Standard Operating Procedure. **protocols.io** https://dx.doi.org/10.17504/protocols.io.bjhpkj5n

LICENSE

CREATED

Aug 10, 2020

LAST MODIFIED

Jan 19, 2021

PROTOCOL INTEGER ID

40207

BEFORE STARTING

Some things to note about using Seek & Blastn for publications:

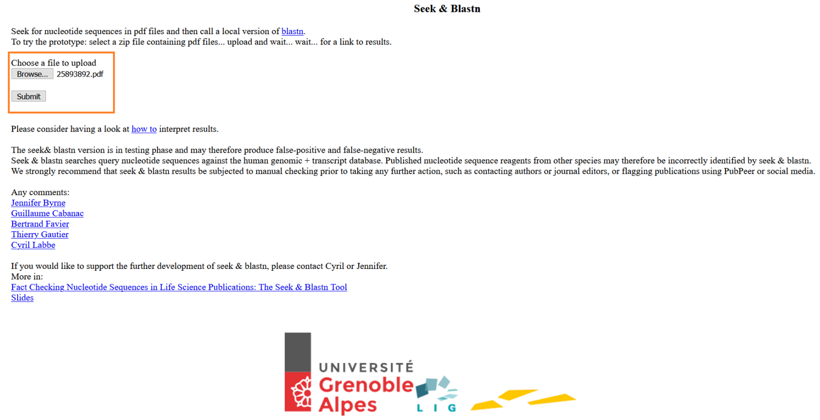
- Seek & Blastn is a program that is designed to extract (i.e. "Seek") and verify the targeting/ non-targeting status of nucleotide sequence reagents.
- The current version of Seek & Blastn fact-checks the identities of nucleotide sequence reagents against the NCBI human genomic and transcript database.
- This protocol is designed to be very detailed and to explain all elements of Seek & Blastn outputs. Don't be put off however, using Seek & Blastn is not difficult.

Upload publication to Seek & Blastn

1 Upload the publication of interest in pdf format to **Seek & Blastn**

1.1 For Seek & Blastn, click here.
Click on the above link and upload the publication of interest by selecting **browse**, then **submit.**
In this procedure, the retracted article: PMID **25893892** will be used as an example.

Multiple publications may be uploaded at once by compressing the pdf of publications into a zip file and uploading the zip file.



Seek & Blastn

Seek for nucleotide sequences in pdf files and then call a local version of blastn.
To try the prototype: select a zip file containing pdf files... upload and wait... wait... for a link to results.

Choose a file to upload
25893892.pdf

Please consider having a look at how to interpret results.

The seek& blastn version is in testing phase and may therefore produce false-positive and false-negative results.
Seek & blastn searches query nucleotide sequences against the human genomic + transcript database. Published nucleotide sequence reagents from other species may therefore be incorrectly identified by seek & blastn.
We strongly recommend that seek & blastn results be subjected to manual checking prior to taking any further action, such as contacting authors or journal editors, or flagging publications using PubPeer or social media.

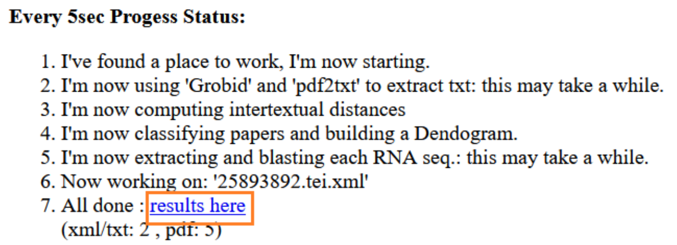
Any comments:
Jennifer Byrne
Guillaume Cabanac
Bertrand Favier
Thierry Gautier
Cyril Labbe

If you would like to support the further development of seek & blastn, please contact Cyril or Jennifer.
More in:
Fact Checking Nucleotide Sequences in Life Science Publications: The Seek & Blastn Tool
Slides

Figure 1.1. The Seek & Blastn homepage after a file has been selected via "Browse...". The file name is shown at the right.

1.2 Wait for Seek & Blastn to finish analyzing the publication, indicated when the window presents the phrase “All done: **results_here**.” Click on the hyperlink to be taken to the Seek & Blastn output page.

Citation: Jennifer A. Byrne, Yasunori Park, Amanda Capes-Davis, Bertrand Favier, Guillaume Cabanac, Cyril LabbÃÂ© (01/19/2021). Seek & Blastn Standard Operating Procedure. https://dx.doi.org/10.17504/protocols.io.bjhpkj5n



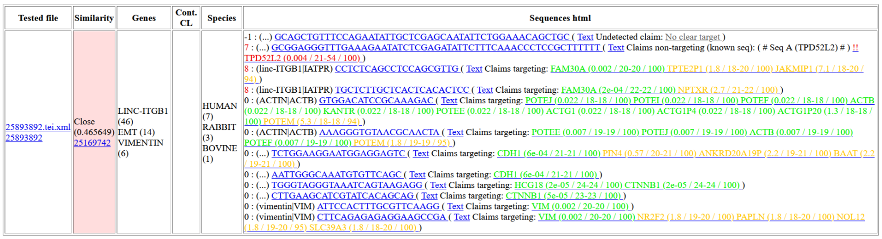
Every 5sec Progess Status:

1. I've found a place to work, I'm now starting.
2. I'm now using 'Grobid' and 'pdf2txt' to extract txt: this may take a while.
3. I'm now computing intertextual distances
4. I'm now classifying papers and building a Dendogram.
5. I'm now extracting and blasting each RNA seq.: this may take a while.
6. Now working on: '25893892.tei.xml'
7. All done : results here
   (xml/txt: 2 , pdf: 3)

Figure 1.2. The Seek & Blastn progress page. The orange box highlights the link indicating that Seek & Blastn has finished analyzing the uploaded publication.

Interpreting Seek & Blastn outputs

2 The output of a Seek & Blastn analysis is presented in tabular form divided into six columns, as shown below for the *retracted* publication PMID **25893892.**

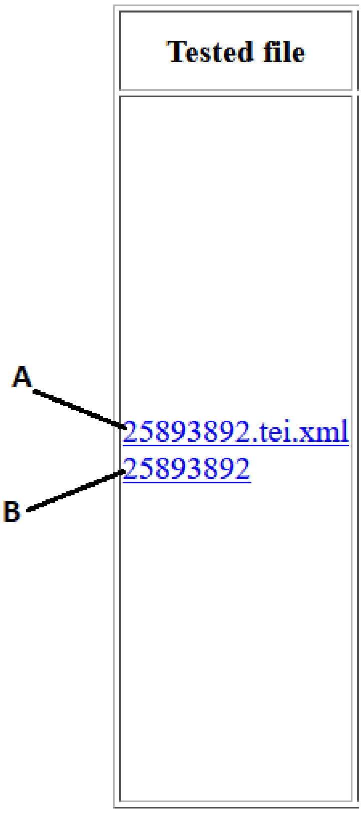


Tested file	Similarity	Genes	Cont. CL	Species	Sequences html
25893892.tei.xml 25893892	Close (0.465649) 25169742	LINC-ITGB1 (46) EMT (14) VIMENTIN (6)		HUMAN (7) RABBIT (3) BOVINE (1)	-1 : (...) GCAGCTGTTTCCAGAATATTGCTCGAGCAATATTCTGGAAACAGCTGC (Text Undetected claim: No clear target) 7 : (...) GCGGAGGGTTTGAAAGAATATCTCGAGATATTCTTTCAAACCCTCCGCTTTTTT (Text Claims non-targeting (known seq): (# Seq A (TPD52L2) #) !! TPD52L2 (0.004 / 21-54 / 100)) 8 : (linc-ITGB1|IATPR) CCTCTCAGCCTCCAGCGTTG (Text Claims targeting: FAM30A (0.002 / 20-20 / 100) TPTE2P1 (1.8 / 18-20 / 100) JAKMIP1 (7.1 / 18-20 / 94)) 8 : (linc-ITGB1|IATPR) TGCTCTTGCTCACTCACACTCC (Text Claims targeting: FAM30A (2e-04 / 22-22 / 100) NPTXR (2.7 / 21-22 / 100)) 0 : (ACTIN|ACTB) GTGGACATCCGCAAAGAC (Text Claims targeting: POTEJ (0.022 / 18-18 / 100) POTEI (0.022 / 18-18 / 100) POTEF (0.022 / 18-18 / 100) ACTB (0.022 / 18-18 / 100) KANTR (0.022 / 18-18 / 100) POTEE (0.022 / 18-18 / 100) ACTG1 (0.022 / 18-18 / 100) ACTG1P4 (0.022 / 18-18 / 100) ACTG1P20 (1.3 / 18-18 / 100) POTEM (5.3 / 18-18 / 94)) 0 : (ACTIN|ACTB) AAAGGGTGTAACGCAACTA (Text Claims targeting: POTEE (0.007 / 19-19 / 100) POTEJ (0.007 / 19-19 / 100) ACTB (0.007 / 19-19 / 100) POTEF (0.007 / 19-19 / 100) POTEM (1.8 / 19-19 / 95)) 0 : (...) TCTGGAAGGAATGGAGGAGTC (Text Claims targeting: CDH1 (6e-04 / 21-21 / 100) PIN4 (0.57 / 20-21 / 100) ANKRD20A19P (2.2 / 19-21 / 100) BAAT (2.2 / 19-21 / 100)) 0 : (...) AATTGGGCAAATGTGTTCAGC (Text Claims targeting: CDH1 (6e-04 / 21-21 / 100)) 0 : (...) TGGGTAGGGTAAATCAGTAAGAGG (Text Claims targeting: HCG18 (2e-05 / 24-24 / 100) CTNNB1 (2e-05 / 24-24 / 100)) 0 : (...) CTTGAAGCATCGTATCACAGCAG (Text Claims targeting: CTNNB1 (5e-05 / 23-23 / 100)) 0 : (vimentin|VIM) ATTCCACTTTGCGTTCAAGG (Text Claims targeting: VIM (0.002 / 20-20 / 100)) 0 : (vimentin|VIM) CTTCAGAGAGAGGAAGCCGA (Text Claims targeting: VIM (0.002 / 20-20 / 100) NR2F2 (1.8 / 19-20 / 100) PAPLN (1.8 / 18-20 / 100) NOL12 (1.8 / 19-20 / 95) SLC39A3 (1.8 / 18-20 / 100))

Figure 2.0. The results page for the PMID: **25893892**. Seek & Blastn outputs can be shared by the URL in the browser (PMID 25893892 result).

2.1 The first column (titled **Tested file**) provides two links as follows:

A) a link to the XML file of the uploaded pdf
B) a link to the uploaded pdf file analyzed by Seek & Blastn.


Tested file

25893892.tei.xml

25893892

Figure 2.1. "Tested file" column of Seek & Blastn output

A) The upper hyperlink is a link to the XML file of the uploaded pdf showing how Seek & Blastn has analyzed the publication.

B) The lower hyperlink takes the user to the pdf file uploaded to Seek & Blastn. Clicking the pdf file link in the example shown below will open the query article "*Long non-coding RNA Linc-ITGB1 knockdown inhibits cell migration and invasion in GBC-SD/M and GBC-SD gallbladder cancer cell lines*" (PMID: 25893892).

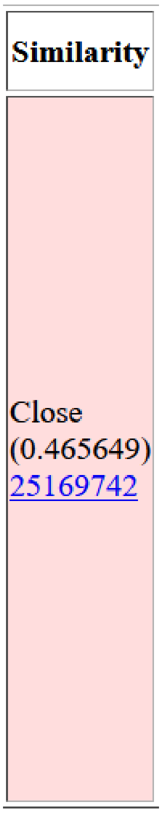
2.2 The second column (titled **Similarity**) shows the results of intertextual distance analysis (Labbé & Labbé 2012). The PMID of the most textually similar publication in the reference corpus (Byrne & Labbé 2017) is provided with the intertextual distance analysis result, divided into the following similarity categories: "OK", "close" and "very close".

Intertextual distance describes the degree of textual similarity between the query and the most similar publication in a reference publication corpus. This distance is provided as a value from 0 to 1, where smaller values indicate greater textual similarity. (Labbé & Labbé 2012)

The degree of similarity for:

- "OK" represents an intertextual distance of > 0.5 to the nearest reference publication.
- "Close" represents an intertextual distance 0.44 - 0.50 to the nearest reference publication as above.
- "Very close" represents an intertextual distance of < 0.44, as above.

Values were based on the intertextual distance score distribution of 4094 publications from the International Journal of Clinical and Experimental Medicine (Byrne & Labbé 2017)



Similarity
Close (0.465649) 25169742

Figure 2.2. "Similarity" column of Seek & Blastn output. The intertextual distance analysis result shown in brackets falls within the category of "close" (0.44 - 0.50). The PMID of the reference publication that is most similar to the query publication is shown below. This reference publication (PMID: **25169742**) comes from a reference corpus of 15 publications, as previously described (Byrne & Labbé 2017).

2.3 The third, fourth, and fifth columns (titled **Genes**, **Cont. CL,** and **Species,** respectively) list identifiers detected by Seek & Blastn within the uploaded publication. If no identifiers are found by Seek & Blastn, the relevant column is left blank.

- "**Genes**" lists individual gene identifiers (GeneCards) from the query publication, with the number of times that Seek & Blastn recognized the identifier shown in brackets.
- "**Cont. CL**" lists any cell line (CL) identifier that has been associated with contamination or misidentification according to the database of Cross-contaminated or Misidentified Cell Lines version 7.2 established by the International Cell Line Authentication Committee (ICLAC).
- "**Species**" lists species identifiers such as "human", "mouse", "rabbit" (Linnaeus) with the number of times each identifier is detected shown in brackets.

It should be noted by users the ICLAC list of cell lines continues to be updated, and the version implemented in Seek & Blastn (version 7.2, 6 October 2014) is **not** the latest version (see https://iclac.org/databases/cross-contaminations/).

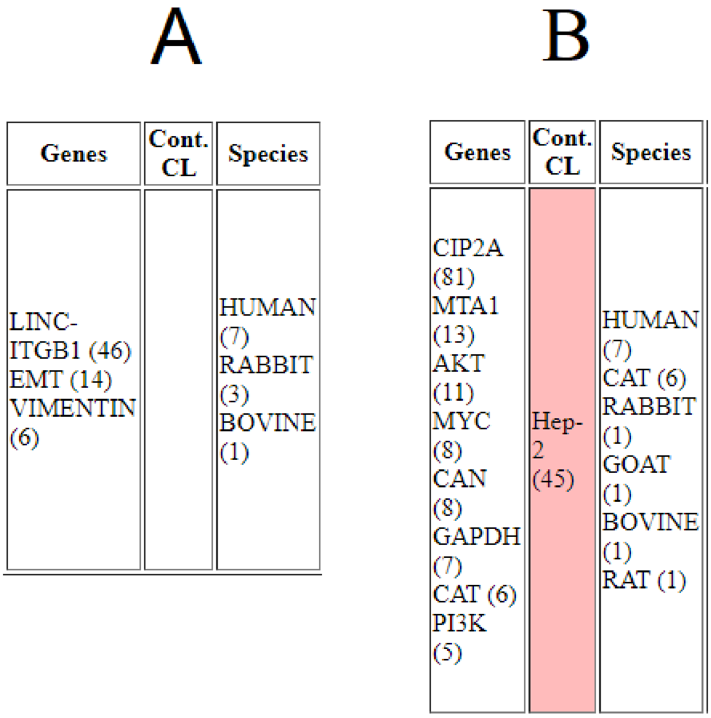
Uploaded publications are considered to describe human research if the frequency of the term "human" is ≥ (the number of times the most frequent species "**not** human" -1) mentioned in the publication.

In publications where this criterion is not met (i.e. "human" < most frequent "not human" species -

Citation: Jennifer A. Byrne, Yasunori Park, Amanda Capes-Davis, Bertrand Favier, Guillaume Cabanac, Cyril LabbÃÂ© (01/19/2021). Seek & Blastn Standard Operating Procedure. https://dx.doi.org/10.17504/protocols.io.bjhpkj5n

1), Seek & Blastn will assume that the publication does not describe human research and will not analyze the publication. This will be indicated by a lack of extracted nucleotide sequences.

In publications that describe the use of human cancer cell lines, other species can be mentioned in association with antibodies. These species may cause Seek & Blastn to not extract sequences and skip the paper. In this case, go to **step 4.1**.



Genes	Cont. CL	Species
LINC-ITGB1 (46) EMT (14) VIMENTIN (6)		HUMAN (7) RABBIT (3) BOVINE (1)

Genes	Cont. CL	Species
CIP2A (81) MTA1 (13) AKT (11) MYC (8) CAN (8) GAPDH (7) CAT (6) PI3K (5)	Hep-2 (45)	HUMAN (7) CAT (6) RABBIT (1) GOAT (1) BOVINE (1) RAT (1)

Figure 2.3. The outputs for "Genes", "Cont. CL", and "Species" columns for PMIDs **25893892** and **28656258.**

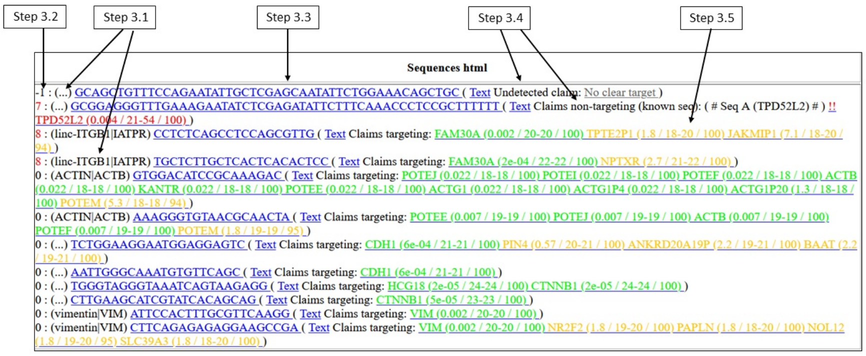
A) In the PMID **25893892**, the output column for "Genes" shows Seek & Blastn identified the gene name "linc-ITGB1" 46 times, "EMT" 14 times, and "Vimentin" 6 times. Note that in this case, EMT (Epithelial-mesenchymal transition) does not represent a gene highlighting the challenges posed by gene identifiers that correspond to other abbreviations. The output column for "Cont. CL" is blank as Seek & Blastn did not identify any contaminated cell lines within this publication. The "Species" column indicates that Seek & Blastn identified the term "human" 7 times, "rabbit" 3 times, and "bovine" once. The publication was analyzed by Seek & Blastn, as human (7) was mentioned more times than rabbit (3) -1 (i.e., 7 ≥ (3-1)).

B) PMID **28656258** has been provided to show Seek & Blastn flagging a potentially misidentified cell line. In this case, Hep-2, a known HeLa derivative that was believed to have originated from laryngeal cancer instead, was detected 45 times within the publication.

Interpreting Seek & Blastn outputs (Sequences)

3 The sixth column (titled "**Sequences.html**") lists the Blastn results for all the nucleotide sequences that were extracted by Seek & Blastn. Individual sequences are listed in rows.

Citation: Jennifer A. Byrne, Yasunori Park, Amanda Capes-Davis, Bertrand Favier, Guillaume Cabanac, Cyril LabbÃÂ© (01/19/2021). Seek & Blastn Standard Operating Procedure. https://dx.doi.org/10.17504/protocols.io.bjhpkj5n

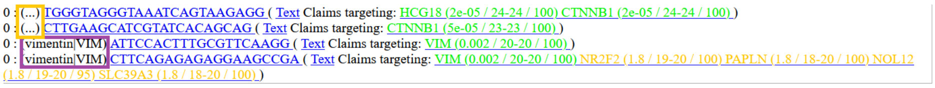


```
Sequences html
-1 : (...) GCAGCTGTTTCCAGAATATTGCTCGAGCAATATTCTGGAAACAGCTGC ( Text Undetected claim: No clear target )
7 : (...) GCGGAGGGTTTGAAAGAATATCTCGAGATATTCTTTCAAACCCTCCGCTTTTTT ( Text Claims non-targeting (known seq): ( # Seq A (TPD52L2) # ) !!
TPD52L2 (0.004 / 21-54 / 100) )
8 : (linc-ITGB1|IATPR) CCTCTCAGCCTCCAGCGTTG ( Text Claims targeting: FAM30A (0.002 / 20-20 / 100) TPTE2P1 (1.8 / 18-20 / 100) JAKMIP1 (7.1 / 18-20 /
94) )
8 : (linc-ITGB1|IATPR) TGCTCTTGCTCACTCACACTCC ( Text Claims targeting: FAM30A (2e-04 / 22-22 / 100) NPTXR (2.7 / 21-22 / 100) )
0 : (ACTIN|ACTB) GTGGACATCCGCAAAGAC ( Text Claims targeting: POTEJ (0.022 / 18-18 / 100) POTEI (0.022 / 18-18 / 100) POTEF (0.022 / 18-18 / 100) ACTB
(0.022 / 18-18 / 100) KANTR (0.022 / 18-18 / 100) POTEE (0.022 / 18-18 / 100) ACTG1 (0.022 / 18-18 / 100) ACTG1P4 (0.022 / 18-18 / 100) ACTG1P20 (1.3 / 18-18 /
100) POTEM (5.3 / 18-18 / 94) )
0 : (ACTIN|ACTB) AAAGGGTGTAACGCAACTA ( Text Claims targeting: POTEE (0.007 / 19-19 / 100) POTEJ (0.007 / 19-19 / 100) ACTB (0.007 / 19-19 / 100)
POTEF (0.007 / 19-19 / 100) POTEM (1.8 / 19-19 / 95) )
0 : (...) TCTGGAAGGAATGGAGGAGTC ( Text Claims targeting: CDH1 (6e-04 / 21-21 / 100) PIN4 (0.57 / 20-21 / 100) ANKRD20A19P (2.2 / 19-21 / 100) BAAT (2.2
/ 19-21 / 100) )
0 : (...) AATTGGGCAAATGTGTTCAGC ( Text Claims targeting: CDH1 (6e-04 / 21-21 / 100) )
0 : (...) TGGGTAGGGTAAATCAGTAAGAGG ( Text Claims targeting: HCG18 (2e-05 / 24-24 / 100) CTNNB1 (2e-05 / 24-24 / 100) )
0 : (...) CTTGAAGCATCGTATCACAGCAG ( Text Claims targeting: CTNNB1 (5e-05 / 23-23 / 100) )
0 : (vimentin|VIM) ATTCCACTTTGCGTTCAAGG ( Text Claims targeting: VIM (0.002 / 20-20 / 100) )
0 : (vimentin|VIM) CTTCAGAGAGAGGAAGCCGA ( Text Claims targeting: VIM (0.002 / 20-20 / 100) NR2F2 (1.8 / 19-20 / 100) PAPLN (1.8 / 18-20 / 100) NOL12
(1.8 / 19-20 / 95) SLC39A3 (1.8 / 18-20 / 100) )
```

Figure 3.0. Results of sequences extracted from PMID **25893892** and analyzed by Seek & Blastn. Results are explained under the following steps:

1. Extracted gene symbols associated with the nucleotide sequence reagent (**step 3.1**)
2. Indication of any error types (**step 3.2**)
3. The extracted nucleotide sequence reagent (**step 3.3**)
4. Link to the extracted nucleotide sequence within the uploaded pdf and claimed the status of the sequence (**step 3.4**)
5. Blastn results with associated values (**step 3.5**)
6. Categorization of Blastn results (**step 3.6**) ***not shown in the figure.***

3.1 The gene symbols and their synonyms are the identifiers associated by Seek & Blastn with the extracted sequence within the uploaded publication. These are from "The approved symbol and synonyms list established by the HUGO Gene Nomenclature Committee (HGNC)". When no gene symbol is extracted by Seek & Blastn, it is indicated by empty brackets (...).



```
0 : (...) TGGGTAGGGTAAATCAGTAAGAGG ( Text Claims targeting: HCG18 (2e-05 / 24-24 / 100) CTNNB1 (2e-05 / 24-24 / 100) )
0 : (...) CTTGAAGCATCGTATCACAGCAG ( Text Claims targeting: CTNNB1 (5e-05 / 23-23 / 100) )
0 : (vimentin|VIM) ATTCCACTTTGCGTTCAAGG ( Text Claims targeting: VIM (0.002 / 20-20 / 100) )
0 : (vimentin|VIM) CTTCAGAGAGAGGAAGCCGA ( Text Claims targeting: VIM (0.002 / 20-20 / 100) NR2F2 (1.8 / 19-20 / 100) PAPLN (1.8 / 18-20 / 100) NOL12
(1.8 / 19-20 / 95) SLC39A3 (1.8 / 18-20 / 100) )
```

Figure 3.1. The last four extracted sequences from the example Seek & Blastn output. The yellow box indicates sequences where Seek & Blastn was unable to identify the gene symbol corresponding to the extracted sequence. The purple box shows sequences where Seek & Blastn associated the extracted sequences with a particular gene symbol, shown correctly as vimentin.

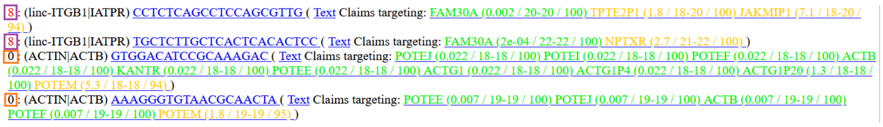
3.2 Where the gene symbol of an extracted sequence matches a Blastn search, the integer to the **left** of the gene symbol remains black.

When a gene symbol however **does not match** any of the significant Blastn hits, the number is highlighted in red, indicating a possible mismatch between the extracted nucleotide sequence and its claimed identity (**figure 3.2**).

Integers next to gene symbols and their respective definitions:

- **-1**: Seek & Blastn was unable to identify whether the sequence is targeting or non-targeting
- **0**: No error detected
- **6**: "Targeting" sequence is predicted to be non-targeting
- **7**: "Non-targeting" sequence is predicted to be targeting
- **8**: Nucleotide sequence reagent is targeting, but is predicted to target the wrong gene or sequence

It is recommended that results flagged with a red integer be manually verified, as per step 5.

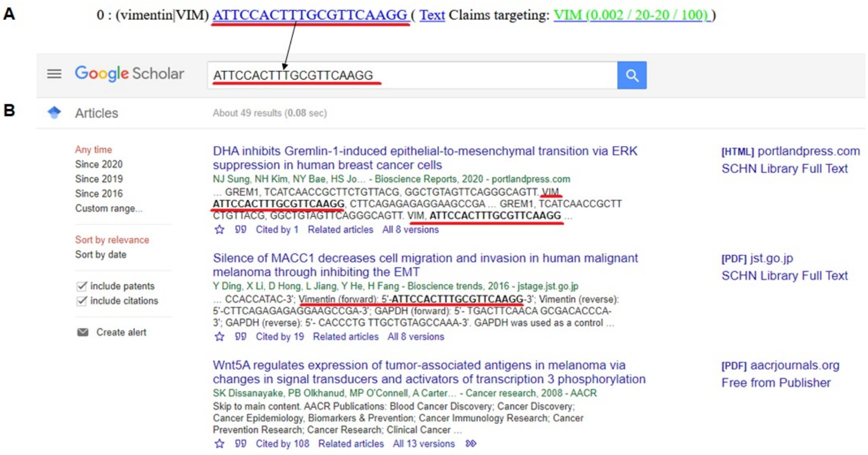


8 : (linc-ITGB1|IATPR) CCTCTCAGCCTCCAGCGTTG (Text Claims targeting: FAM30A (0.002 / 20-20 / 100) TPTE2P1 (1.8 / 18-20 / 100) JAKMIP1 (7.1 / 18-20 / 94))
8 : (linc-ITGB1|IATPR) TGCTCTTGCTCACTCACACTCC (Text Claims targeting: FAM30A (2e-04 / 22-22 / 100) NPTXR (2.7 / 21-22 / 100))
0 : (ACTIN|ACTB) GTGGACATCCGCAAAGAC (Text Claims targeting: POTEJ (0.022 / 18-18 / 100) POTEI (0.022 / 18-18 / 100) POTEF (0.022 / 18-18 / 100) ACTB (0.022 / 18-18 / 100) KANTR (0.022 / 18-18 / 100) POTEE (0.022 / 18-18 / 100) ACTG1 (0.022 / 18-18 / 100) ACTG1P4 (0.022 / 18-18 / 100) ACTG1P20 (1.3 / 18-18 / 100) POTEM (5.3 / 18-18 / 94))
0 : (ACTIN|ACTB) AAAGGGTGTAACGCAACTA (Text Claims targeting: POTEE (0.007 / 19-19 / 100) POTEJ (0.007 / 19-19 / 100) ACTB (0.007 / 19-19 / 100) POTEF (0.007 / 19-19 / 100) POTEM (1.8 / 19-19 / 95))

Figure 3.2. An example of a mismatch between a claimed sequence identity and the Blastn results. The first two sequences are claimed by the publication PMID **25893892** to be RT-PCR primers for linc-ITGB1. These sequences are flagged as a potential error by a red number to the left of the gene symbol, in this example shown by the number **8** (nucleotide sequence reagent "targeting" the wrong gene as per a Blastn search) (purple box). The gene symbol for ACTB was extracted and sequences were analyzed to return a Blastn hit for ACTB as shown above. These sequences are not flagged by Seek & Blastn as shown by the number **0** (no error detected) to the left of the gene symbol in black (orange box).

3.3 Extracted nucleotide sequences are provided as blue hyperlinks, that when clicked, runs a Google Scholar search using the nucleotide sequence as a query (**figure 3.3**).

Next to each sequence is a "Text" hyperlink that aims to identify the location of the extracted sequence within the text. It searches the first three nucleotides of the sequence in the pdf and thereby may not directly lead to the sequence.



0 : (vimentin|VIM) ATTCCACTTTGCGTTCAAGG (Text Claims targeting: VIM (0.002 / 20-20 / 100))

ATTCCACTTTGCGTTCAAGG

About 49 results (0.08 sec)

DHA inhibits Gremlin-1-induced epithelial-to-mesenchymal transition via ERK suppression in human breast cancer cells
NJ Sung, NH Kim, NY Bae, HS Jo... - Bioscience Reports, 2020 - portlandpress.com
... GREM1, TCATCAACCGCTTCTGTTACG, GGCTGTAGTTCAGGGCAGTT. VIM, ATTCCACTTTGCGTTCAAGG, CTTCAGAGAGAGGAAGCCGA ... GREM1, TCATCAACCGCTTCTGTTACG, GGCTGTAGTTCAGGGCAGTT. VIM, ATTCCACTTTGCGTTCAAGG ...
☆ 99 Cited by 1 Related articles All 8 versions
[HTML] portlandpress.com
SCHN Library Full Text

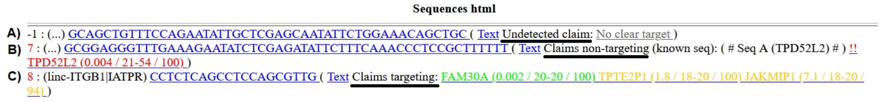
Silence of MACC1 decreases cell migration and invasion in human malignant melanoma through inhibiting the EMT
Y Ding, X Li, D Hong, L Jiang, Y He, H Fang - Bioscience trends, 2016 - jstage.jst.go.jp
... CCACCATAC-3'; Vimentin (forward): 5'-ATTCCACTTTGCGTTCAAGG-3'; Vimentin (reverse): 5'-CTTCAGAGAGAGGAAGCCGA-3'; GAPDH (forward): 5'- TGACTTCAACA GCGACACCCA-3'; GAPDH (reverse): 5'- CACCCTG TTGCTGTAGCCAAA-3'. GAPDH was used as a control ...
☆ 99 Cited by 19 Related articles All 8 versions
[PDF] jst.go.jp
SCHN Library Full Text

Wnt5A regulates expression of tumor-associated antigens in melanoma via changes in signal transducers and activators of transcription 3 phosphorylation
SK Dissanayake, PB Olkhanud, MP O'Connell, A Carter... - Cancer research, 2008 - AACR
Skip to main content. AACR Publications: Blood Cancer Discovery; Cancer Discovery; Cancer Epidemiology, Biomarkers & Prevention; Cancer Immunology Research; Cancer Prevention Research; Cancer Research; Clinical Cancer ...
☆ 99 Cited by 108 Related articles All 13 versions »
[PDF] aacrjournals.org
Free from Publisher

Figure 3.3. Clicking on the hypertext in the Seek & Blastn output (A) uses the sequence as a Google Scholar search query (B).

3.4 Next to each "Text" link are the detected corresponding text claims, shown as either “Claims targeting”, “Claims non-targeting”, or “Undetected”. These claims can be verified as per **step 5**. If a sequence instead has the message “(Seq. not correctly extracted (Char/long/ short)”, go to **step 4.2**.

Citation: Jennifer A. Byrne, Yasunori Park, Amanda Capes-Davis, Bertrand Favier, Guillaume Cabanac, Cyril LabbÃÂ© (01/19/2021). Seek & Blastn Standard Operating Procedure. https://dx.doi.org/10.17504/protocols.io.bjhpkj5n



Sequences html

A) -1 : (...) GCAGCTGTTTCCAGAATATTGCTCGAGCAATATTCTGGAAACAGCTGC (Text Undetected claim: No clear target)
B) 7 : (...) GCGGAGGGTTTGAAAGAATATCTCGAGATATTCTTTCAAACCCTCCGCTTTTTT (Text Claims non-targeting (known seq): (# Seq A (TPD52L2) #) !! TPD52L2 (0.004 / 21-54 / 100))
C) 8 : (linc-ITGB1|IATPR) CCTCTCAGCCTCCAGCGTTG (Text Claims targeting: FAM30A (0.002 / 20-20 / 100) TPTE2P1 (1.8 / 18-20 / 100) JAKMIP1 (7.1 / 18-20 / 94))

Figure 3.4. Underlined next to the hyperlink "text" are the claims extracted by Seek & Blastn.

- A) "Undetected" indicates that Seek & Blastn was unable to detect the predicted claimed status of the nucleotide sequence within the query publication.
- B) "Claims non-targeting" indicates that Seek & Blastn has identified the nucleotide sequence as a claimed non-targeting reagent.
- C) "Claims targeting" indicates that Seek & Blastn has identified the nucleotide sequence as a claimed targeting reagent.

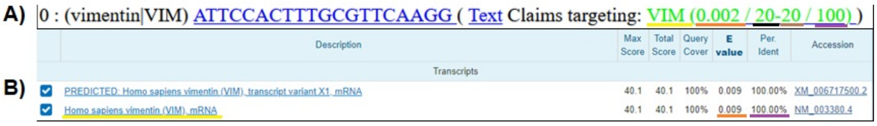
If Seek & Blastn is able to identify a sequence as either targeting or non-targeting, outputs will be provided in either green/orange/red hypertext (see **step 3.6**)

If Seek & Blastn is unable to detect a claim, the output hypertext is shown in grey (**step 3.6**).

Outputs shown in grey, orange, or red hypertext can be verified as per **step 5**.

3.5 Next to the text claims (**step 3.4**) is the gene name corresponding to each targeting Blastn hit that is hyperlinked to the associated Blastn results, shown below in **figure 3.5**. This can support manual validation of results, as per **step 5**.

Where a nucleotide sequence is found to be non-targeting in a Blastn search, this is indicated by "No clear target".



A) 0 : (vimentin|VIM) ATTCCACTTTGCGTTCAAGG (Text Claims targeting: VIM (0.002 / 20-20 / 100))

B)

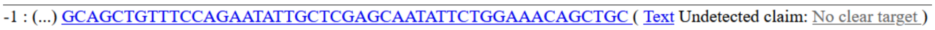
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
	Transcripts						
☑	PREDICTED: Homo sapiens vimentin (VIM), transcript variant X1, mRNA	40.1	40.1	100%	0.009	100.00%	XM_006717500.2
☑	Homo sapiens vimentin (VIM), mRNA	40.1	40.1	100%	0.009	100.00%	NM_003380.4

Figure 3.5. A) Features of each Blastn hit listed by Seek & Blastn. For each hit (underlined *yellow*), the following values are supplied in brackets:

1. The associated e-value of the extracted query sequence produced by Blastn (underlined *orange*)
2. The number of sequential nucleotides in the query sequence mapping to the hit (underlined *black*)
3. The length of the query sequence (underlined *brown*)
4. The percentage of sequence identity between the query sequence and the hit (underlined *purple*)

B) E-values may differ from those produced by the Blastn searches of the "*human genomic plus transcript*" database and optimized for "*somewhat similar sequences*". This is due to Seek & Blastn using an older version of Blastn (BLASTN 2.6.0+). than the Blastn version available at https://blast.ncbi.nlm.nih.gov/Blast.cgi

3.6 Grey hypertext indicates that Seek & Blastn did not detect the targeting/non-targeting status of the sequence. These sequences can be manually verified as per **step 5**.

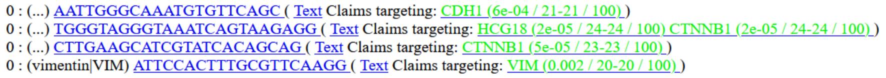


-1 : (...) GCAGCTGTTTCCAGAATATTGCTCGAGCAATATTCTGGAAACAGCTGC (Text Undetected claim: No clear target)

Figure 3.6. An example of a sequence shown as grey hypertext, where no claim was identified by Seek & Blastn. As per **step 3.5**, the Blastn result suggests that this sequence is non-targeting.

Green hypertext indicates that the *claimed status* agrees with the *verified status*, either "Claims targeting" or "Claims non-targeting", shown below in **figure 3.7**.

Citation: Jennifer A. Byrne, Yasunori Park, Amanda Capes-Davis, Bertrand Favier, Guillaume Cabanac, Cyril LabbÃÂ© (01/19/2021). Seek & Blastn Standard Operating Procedure. https://dx.doi.org/10.17504/protocols.io.bjhpkj5n

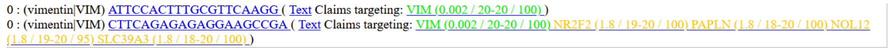


0 : (...) AATTGGGCAAATGTGTTCAGC (Text Claims targeting: CDH1 (6e-04 / 21-21 / 100))
0 : (...) TGGGTAGGGTAAATCAGTAAGAGG (Text Claims targeting: HCG18 (2e-05 / 24-24 / 100) CTNNB1 (2e-05 / 24-24 / 100))
0 : (...) CTTGAAGCATCGTATCACAGCAG (Text Claims targeting: CTNNB1 (5e-05 / 23-23 / 100))
0 : (vimentin|VIM) ATTCCACTTTGCGTTCAAGG (Text Claims targeting: VIM (0.002 / 20-20 / 100))

Figure 3.7. Examples of sequences shown in green hypertext. Despite several gene symbols not being extracted, these four nucleotide sequence reagents were identified to be targeting within the publication, and Blastn results show that they fulfill the criteria of targeting reagents.

The criteria for a sequence to be considered targeting requires either:
(i) 100% sequence identity over at least 15 consecutive nucleotides, including the 3' end nucleotide of the extracted sequence
(ii) two different subsequences of the sequence query 19-21 nucleotides long matching a single target with inverted homology and/or
(iii) 100% sequence identity over at least 17 consecutive nucleotides. (Labbé et al., 2019)

Orange hypertext indicates Blastn hits of lower significance, explained in **figure 3.8.** If Blastn analyses return only low significance hits, these should be manually validated as per **step 5**.



0 : (vimentin|VIM) ATTCCACTTTGCGTTCAAGG (Text Claims targeting: VIM (0.002 / 20-20 / 100))
0 : (vimentin|VIM) CTTCAGAGAGAGGAAGCCGA (Text Claims targeting: VIM (0.002 / 20-20 / 100) NR2F2 (1.8 / 19-20 / 100) PAPLN (1.8 / 18-20 / 100) NOL12 (1.8 / 19-20 / 95) SLC39A3 (1.8 / 18-20 / 100))

Figure 3.8. Examples of sequence(s) shown in orange hypertext.

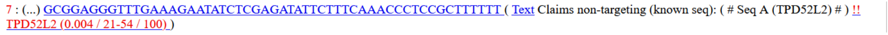
Low significance results are sequences with either: ≥ 90% but < 100% identity over at least 15 nucleotides at a distance of fewer than 3 nucleotides from the 3' end, or, ≤ 100% identity over ≤16 consecutive nucleotides (Labbé et al., 2019)

In the example shown, orange hypertext is shown for NR2F2, PAPLN, NOL12, and SLC39A3, given they showed < 100% sequence identity to the query over < 16 consecutive nucleotides.

Red hypertext indicates Blastn results that conflict with the claimed targeting/non-targeting status of the nucleotide sequence (**figure 3.9**).

Claimed targeting sequences lacking clear targets are indicated by "!! No clear target" or "!! No hits found" in red hypertext.

Claimed non-targeting sequences retrieving significant Blastn results are indicated by "!! (Blastn hit)". These results can be manually verified as per **step 5**.



7 : (...) GCGGAGGGTTTGAAAGAATATCTCGAGATATTCTTTCAAACCCTCCGCTTTTTT (Text Claims non-targeting (known seq): (# Seq A (TPD52L2) #) !! TPD52L2 (0.004 / 21-54 / 100))

Figure 3.9. A nucleotide sequence reagent claimed to be non-targeting ("Claims non-targeting") that is predicted to target a human gene. The extracted sequence is recognized as a previously reported sequence (Seq A) that is predicted to target the TPD52L2 gene (Byrne and Labbé 2017).

Non-targeting sequences should not meet any of the criteria of a targeting reagent (**figure 3.7**).

Validating Seek & Blastn outputs (recommended)

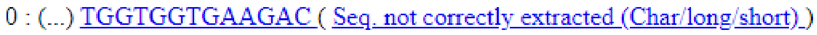
4 Seek & Blastn may not retrieve **all** the relevant information from uploaded publications (Labbé et al. 2019, Labbé et al. 2020). The following steps are recommended to ensure that all Seek & Blastn outputs are analyzed.

4.1 Potentially misidentified cell lines flagged by Seek & Blastn can be found on Cellosaurus. Check the context of any flagged cell lines within the publication to determine if they are being used appropriately. Additionally, Cellosaurus can be used to identify ambiguously named cell lines, to determine whether a misidentified cell line is indeed being used in the publication or if it simply shares the same name as a non-problematic cell line.

If there are any problematic cell lines used within the wrong context, go to **step 7**. To verify any nucleotide sequences, go to **step 4.2**.

4.2 Check that **all** sequences have been extracted by Seek & Blastn. If the number of sequences in Seek & Blastn outputs matches the number of sequences in the publication, go to **step 4.3**. If not, any skipped sequences can be manually verified as per **step 5.**

4.3 Check that all sequences have been extracted correctly. If a nucleotide sequence is either < 14 nucleotides or > 91 nucleotides, Seek & Blastn will recognize this sequence as being incompletely or incorrectly extracted and will flag the reagent with “(Seq. not correctly extracted (Char/long/ short)”. If any sequences have been flagged by Seek & Blastn or incorrectly extracted, manually copy the sequence from the publication and verify, as per step 5.



0 : (...) TGGTGGTGAAGAC (Seq. not correctly extracted (Char/long/short))

Figure 4.0. An example of an incorrectly extracted nucleotide sequence reagent that was not analyzed further by Seek & Blastn (from PMID: **30675243**).

4.4 To manually validate a nucleotide sequence reagent for gene knockdown, PCR, or RT-PCR of a gene other than a microRNA (miR), go to **step 5**.

To manually validate a reverse RT-PCR primer or RT reagent designed for analyzing a miR, got to **step 6**.

Manual verification of Seek & Blastn results (Gene knockdown and/or RT-PCR/PCR primers, excluding miR reagents)

5 Where Seek & Blastn flags a possible incorrect status for any nucleotide sequence reagent, we strongly recommend that such sequences are manually verified prior to taking further action. The recommended steps to manually verify a sequence are summarised in **figure 5.0**.

Citation: Jennifer A. Byrne, Yasunori Park, Amanda Capes-Davis, Bertrand Favier, Guillaume Cabanac, Cyril LabbÃÂ© (01/19/2021). Seek & Blastn Standard Operating Procedure. https://dx.doi.org/10.17504/protocols.io.bjhpkj5n

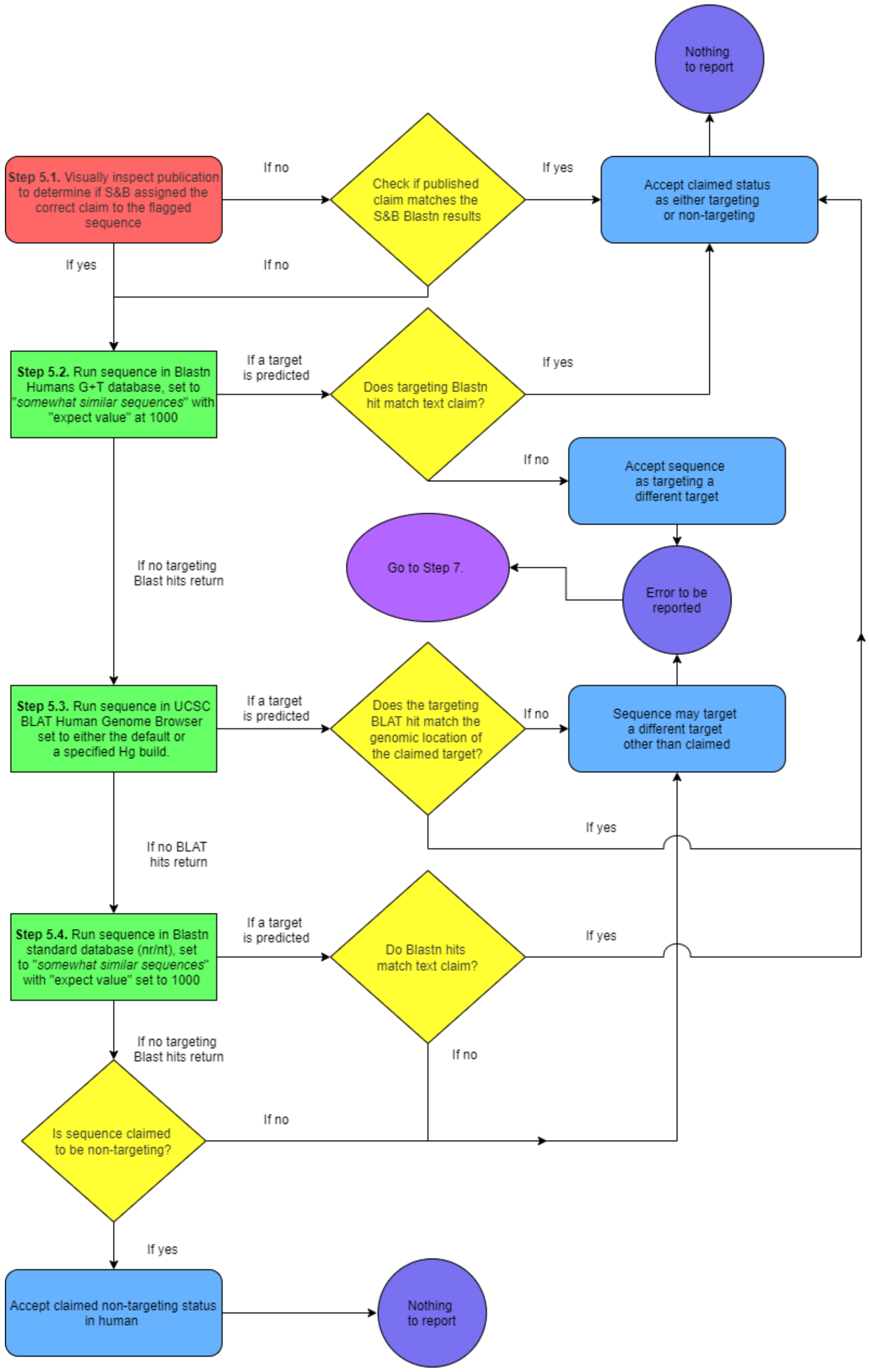

Nothing to report
Step 5.1. Visually inspect publication to determine if S&B assigned the correct claim to the flagged sequence
If no
Check if published claim matches the S&B Blastn results
If yes
Accept claimed status as either targeting or non-targeting
If yes
If no
Step 5.2. Run sequence in Blastn Humans G+T database, set to "somewhat similar sequences" with "expect value" at 1000
If a target is predicted
Does targeting Blastn hit match text claim?
If yes
If no
Accept sequence as targeting a different target
If no targeting Blast hits return
Go to Step 7.
Error to be reported
Step 5.3. Run sequence in UCSC BLAT Human Genome Browser set to either the default or a specified Hg build.
If a target is predicted
Does the targeting BLAT hit match the genomic location of the claimed target?
If no
Sequence may target a different target other than claimed
If yes
If no BLAT hits return
Step 5.4. Run sequence in Blastn standard database (nr/nt), set to "somewhat similar sequences" with "expect value" set to 1000
If a target is predicted
Do Blastn hits match text claim?
If yes
If no targeting Blast hits return
If no
If no
Is sequence claimed to be non-targeting?
If yes
Accept claimed non-targeting status in human
Nothing to report

Figure 5.0. Summary of the workflow used to manually verify Seek & Blastn results for a claimed human nucleotide sequence reagent.

5.1 Visually inspect the publication to determine that the flagged sequence is claimed to be targeting or non-targeting within the text.

- 5.1.1. If the text claim for the sequence was **incorrectly** extracted by Seek & Blastn, check whether the claim matches the Blastn result from Seek & Blastn.

  5.1.1.1. If the text claim matches the Seek & Blastn result in full (e.g., targeting the correct genetic target) accept the reagent as correctly identified.

  5.1.1.2. If the text claim does **not** match the Seek & Blastn result in full (wrong targeting status or genetic target), go to **step 5.2**

Citation: Jennifer A. Byrne, Yasunori Park, Amanda Capes-Davis, Bertrand Favier, Guillaume Cabanac, Cyril LabbÃÂ© (01/19/2021). Seek & Blastn Standard Operating Procedure. https://dx.doi.org/10.17504/protocols.io.bjhpkj5n

- 5.1.2. If the claim was **correctly** extracted by Seek & Blastn, the sequence may be wrongly identified within the text. In this case, go to **step 5.2.**

If a nucleotide sequence reagent is claimed to target a species other than human, the genetic target and species can be confirmed in **step 5.4.**

5.2 Blastn can be accessed **here**.

Run a Blastn analysis for each flagged nucleotide sequence reagent using the following parameters for claimed human sequences: *Human genomic plus transcript (Human G+T)* database, optimized for *somewhat similar sequences*, with the "expect threshold" set to 1000.

- 5.2.1. If the Blastn results suggest that the query is *a targeting sequence* **(step 3.4)**, check whether the Blastn results match the claimed identity within the text.
  5.2.1.1. If the Blastn results **match** the claimed identity, the sequence can be accepted as targeting the claimed target. No further action is needed.
  5.2.1.2. If the Blastn results do **not** match the claimed identity, the query may target another gene or target. In this case, potential recommended courses of action are provided in **step 7.**

Where nucleotide sequence reagents are suggested to target another gene or target, gene identifiers should be crosschecked against potential synonyms in the Genecards (Genecards) database to avoid incorrectly flagging a valid targeting sequence.

- 5.2.2. If the query does not return any significant results in a Blastn search, proceed to **step 5.3.**

5.3 If the query sequence is claimed to target a human gene or genomic sequence, the sequence can be used to query the **UCSC Genome Browser**.

- 5.3.1. The genomic location (i.e. the chromosome and the arm) of the claimed target can be found via both the UCSC Genome Browser or Genecards for the human genome assembly GRCh38/hg38 or for a specified version such as GRCh37/hg19.

- 5.3.2. Open the BLAT search genome available on the UCSC genome browser and copy in the query sequence. Use the following search parameters:
  Genome: *human*
  Assembly: *GRCh38/hg38* (or otherwise as stated by the publication)
  Query type: *BLAT's guess*
  Sort output: *Query, score*
  Output type: *Hyperlink*

- 5.3.3. If the query returns a hit, the location can be viewed in either the "browser" (adjacent to the search bar) or "details" (under the "Side by Side Alignment") option.
  5.3.3.1. If a hit is identified within the genomic range of the claimed target, available within the genome browser, the claimed status may be correct. No further action may be needed.
  5.3.3.2. If a genomic hit is identified at a location **outside** the range of the claimed target, the sequence may be targeting a different target. In this case, potential recommended courses of action are provided in **step 7.**

- 5.3.4. If no result appears, indicated by the BLAT search returning the message "Sorry, no matches found", in either GRCh38 or GRCh37, proceed to **step 5.4.**
  5.3.4.1. This may result from the query being too short ("Warning: Sequence YourSeq is only [**a value <20**] letters long (20 is the recommended minimum)")

Sequences found to be non-targeting in both the Blastn G+T database (**step 5.2**) and the UCSC BLAT

5.4 search genome (**step 5.3.**) can be used as queries in a Blastn analysis, against the *standard database (nucleotide collection (nr/nt))*, optimized for *somewhat similar sequences,* with the "expect threshold" set to 1000. Unrestricted Blastn searches may identify hits in other species.

- 5.4.1. If Blastn suggests the query sequence is targeting, check whether the species and genetic target match the text claim.
  5.4.1.1. If both the species and genetic identity match the text claim, the query can be accepted as a correct targeting reagent.
  5.4.1.2. As described in **step 5.2.1**, if Blastn results do not match the claimed status or target, nor any synonyms under the Genecards database, the query may be non-targeting in the claimed species. Potential recommended courses of action are provided in **step 7.**

- 5.4.2. If there are no significant Blastn hits, the query can be accepted as a non-targeting sequence. If the sequence was claimed to be non-targeting, no further action is needed.
  5.4.2.1. If the sequence was claimed to be targeting but is found to be non-targeting, recommended courses of action are provided in **step 7**.

Manual verification of Seek & Blastn results (analysing reverse transcriptase (RT) primers and RT-PCR primers for miR analysis)

6 Amplification of miRs via RT-PCR typically involves the addition of a 3' sequence using an RT primer and the use of a reverse RT-PCR primer that binds only the 3' 4-8 nucleotides of the target miR (Dellett and Simpson, 2016). Consequently, RT-PCR miR reagents are more likely to be flagged by Seek & Blastn as non-targeting.

6.1 **miR RT-PCR forward primers** can be verified as per **step 5.2**, and/or as a query in miRBase.

- 6.1.1. If the 3' end of a miR forward RT-PCR primer shows 100% sequence identity to the claimed miR, over at least 12 consecutive ribonucleotides (Busk 2014), it can be accepted as targeting.

- 6.1.2. If the forward miR RT-PCR primer does **not** show 100% sequence identity over at least 12 consecutive nucleotides to the claimed miR, it is potentially a non-targeting reagent. In this case, potential courses of action are provided in **step 7.**

6.2 To analyze **miR RT primers**, retrieve the relevant miR sequence from miRBase. If the miR RT primer is not provided, go to **step 6.3.**

- 6.2.1. If the miR strand is specified (e.g., "hsa-miR-145-5p"), only the relevant strand needs to be analyzed.
  6.2.1.1. If the specific miR strand is **not** specified (e.g., "hsa-miR-145"), analyze both the 5p and 3p strands of the miR.

- 6.2.2. Manually align the nucleotides on the 3' end of the RT primer to the 3' end of the claimed miR target.
  6.2.2.1. If the RT primer shows at least 4 consecutive nucleotides of complementarity, including the 3' end of the miR target (Balcells et al., 2011), the RT primer is accepted as a targeting reagent.
  6.2.2.2. If the RT primer does **not** show at least 4 consecutive nucleotides of complementarity, including the 3' end of the miR target, the RT primer is unlikely to bind the miR target and may be non-targeting. In this case, potential courses of action are provided in **step 7.**

6.3 **miR reverse RT-PCR primers** are analyzed as in **step 6.2.** Where an RT primer is provided, the reverse RT-PCR primer can also be compared to the RT primer as well.

- 6.3.1. If the reverse RT-PCR primer has 100% sequence identity to the RT primer over at least 17 consecutive nucleotides, it is likely to represent a targeting reagent, assuming the RT primer binds the miR as per **step 6.2.2.1.**

- 6.3.2. If the miR RT-PCR reverse primer does **not** share 100% sequence identity with the RT primer over at least 17 consecutive nucleotides, it may represent a non-targeting reagent. Potential courses of action are provided in **step 7.**

Citation: Jennifer A. Byrne, Yasunori Park, Amanda Capes-Davis, Bertrand Favier, Guillaume Cabanac, Cyril LabbÃÂ© (01/19/2021). Seek & Blastn Standard Operating Procedure. https://dx.doi.org/10.17504/protocols.io.bjhpkj5n

- 6.3.3. If the RT primer sequence is not provided, manually align the 3' end of the reverse RT-PCR primer to the 3' end of the claimed miR target.

6.3.3.1. If the reverse RT-PCR primer shows, as per **step 6.2.2.1** at least 4 consecutive nucleotides that are complementary to the 3' end of the miR target (Balcells et al., 2011), it is likely to represent a targeting reagent.

6.3.3.2. If the reverse RT-PCR primer shows < 4 nucleotides of complementarity to the miR target, go to **step 6.3.4.**

- 6.3.4. If the RT primer is not provided, and the reverse RT-PCR primer 3' end is not indicated to bind to the claimed miR target, use the reverse RT-PCR primer as a query in Google Scholar (figure 3.3 B)

6.3.3.1. If the reverse RT-PCR primer has been used as a reverse RT-PCR primer for **other miRs**, or has been described as a universal primer, it may represent an appropriate reverse RT-PCR primer. However, it's possible that it's stated genetic identity (as a reverse RT-PCR primer for a specific miR) is incorrect.

6.3.3.2. If the reverse RT-PCR primer has not been used as a reverse RT-PCR primer by other publications, or listed as a universal primer, it may represent a non-targeting reagent. Potential courses of action are provided in **step 7.**

6.4 An example of an analysis of miR RT-PCR reagents is shown in **figure 6.0.** An example of an analysis of miR RT-PCR reagents shown to be **incorrect** is shown in **figure 6.1.**

MIr1208LineUp.png

Figure 6.0: An example of the analysis of miRNA RT and RT-PCR reagents.

A) RT-PCR reagents for human miR-1208 were obtained from the supplementary file of PMID: **30341811**, which included the RT primer, the forward RT-PCR primer, and the reverse RT-PCR primer.

B) The mature sequence miR-1208 was obtained by searching for "hsa-miR-1208" in miRBase. Only one mature miR-1208 sequence is provided by miRBase.

C) Comparison of the **forward RT-PCR primer** (**step 6.1**) and the miR-1208 sequence revealed 100% sequence identity over 12 nucleotides (*highlighted blue*). Manual comparison of the 3' end of the **RT primer** (**step 6.2**) and the 3' end of hsa-miR-1208 revealed complementarity over at least 4 consecutive nucleotides (*highlighted green*). Manual comparison of the **reverse RT-PCR primer** (**step 6.3**) and the RT-primer revealed 100% sequence identity over 18 nucleotides (*highlighted purple*). Consequently, all nucleotide sequence reagents are likely to be appropriate for the analysis of human miR-1208.

61figa2Resized.png

Figure 6.1. An example of the analysis of incorrect miRNA RT and RT-PCR reagents.

A) RT-PCR reagents for human miR-145-5p were obtained from the supplementary file of PMID: **25578496**, which included the stem-loop RT-primer and two RT-PCR primers for miR-145-5p labeled as a "forward" and "reverse" primer.

B) The mature sequence of miR-145-5p was obtained by searching for "hsa-miR-145" in miRBase. Only the 5p sequence was analyzed as specified within the supplementary file (A).

C) Manual comparison of the **RT primer** (**step 6.2**) and the 3' end of hsa-miR-145-5p (*highlighted green*) revealed complementarity over at least 4 consecutive nucleotides. Manual comparison of the **"reverse" RT-PCR primer** did not show any sequence identity to the RT primer, and instead showed 100% sequence identity to the 5' end of hsa-miR-145-5p over 12 consecutive nucleotides (*highlighted blue*). Manual comparison of the **"forward" RT-PCR primer** (*highlighted grey*) showed no sequence identity to the RT primer or hsa-miR-145-5p.

D) The "**forward" RT-PCR primer** sequence was used as a search query in Google Scholar. Search results revealed that this sequence is commonly used as a forward RT-PCR primer for BMI1.

E) A Blastn search using the parameters of **step 5.2** confirmed that the "forward" RT-PCR primer sequence is likely to target BMI1 instead of hsa-miR-145-5p.

Recommended steps in the event of incorrect nucleotide sequence reagents

7 After the identification of an incorrect nucleotide sequence reagent or problematic cell line, possible further actions include:

Citation: Jennifer A. Byrne, Yasunori Park, Amanda Capes-Davis, Bertrand Favier, Guillaume Cabanac, Cyril LabbÃ© (01/19/2021). Seek & Blastn Standard Operating Procedure. https://dx.doi.org/10.17504/protocols.io.bjhpkj5n

- Contacting publishing journals, publication authors, and/or their institutions.

- Publications can be commented on in PubPeer (https://pubpeer.com/).

Citation: Jennifer A. Byrne, Yasunori Park, Amanda Capes-Davis, Bertrand Favier, Guillaume Cabanac, Cyril LabbÃÂ© (01/19/2021). Seek & Blastn Standard Operating Procedure. https://dx.doi.org/10.17504/protocols.io.bjhpkj5n