Scientific Summarization – Annotation Guidelines

Background

Our task is to create annotations so that when you are reading a paper (the citing paper) that cites another paper (the reference paper), you will have a direct link to those portions of the reference paper that the author was referring to. If we follow this process for all citing papers, then when you look at a reference paper, you will able to see which portions of a reference paper were important to subsequent research, and why (was it the method? the results?). In addition, when reading a reference paper, you will be able to see what people later said about that part of the reference paper, whether they support the reference paper or find fault with it, or both.

Here are two diagrams showing the possibilities. Figure 1 represents the scenario where the reader is deciding whether to read a specific referenced paper, looks at the reference spans in the reference paper; in addition, the reader can also discover other papers that cite the same or a neighboring span. Figure 2 represents the scenario where the reader is committed to reading the reference paper, but is getting a deeper understanding of the impact of the paper from seeing the subsequent discussion of that paper. A full description can be found in this document: SciSumm-background-and-proposal.pdf in the Dropbox directory.

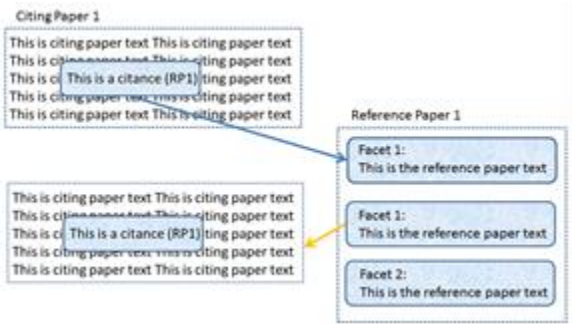


Figure 1. Citance to reference paper summary, where each set of sentences in the reference paper summary preserves links back to the citances that selected those sentences.

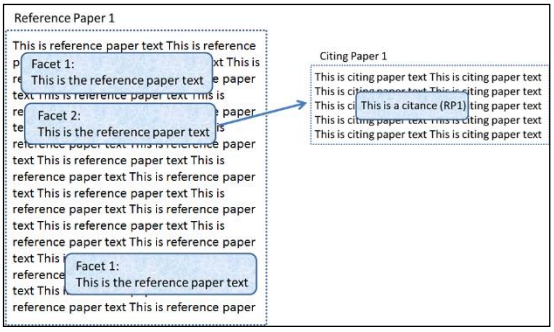


Figure 2. Reference paper with forward links to subsequent literature

Key Concepts

Citing Paper: a paper that cites another paper

Reference Paper: a paper that is cited by another paper

Citance: a citance is made up of the citation text and the citation marker.

Citation text: the sentence(s) in a citing paper that contain the citation and convey the authors' discussion of the citation. The citation text may consist only of the sentence containing the citation, or may include one or more sentences before and/or after the citation. A citation text may also consist of a portion of the sentence containing the citation.

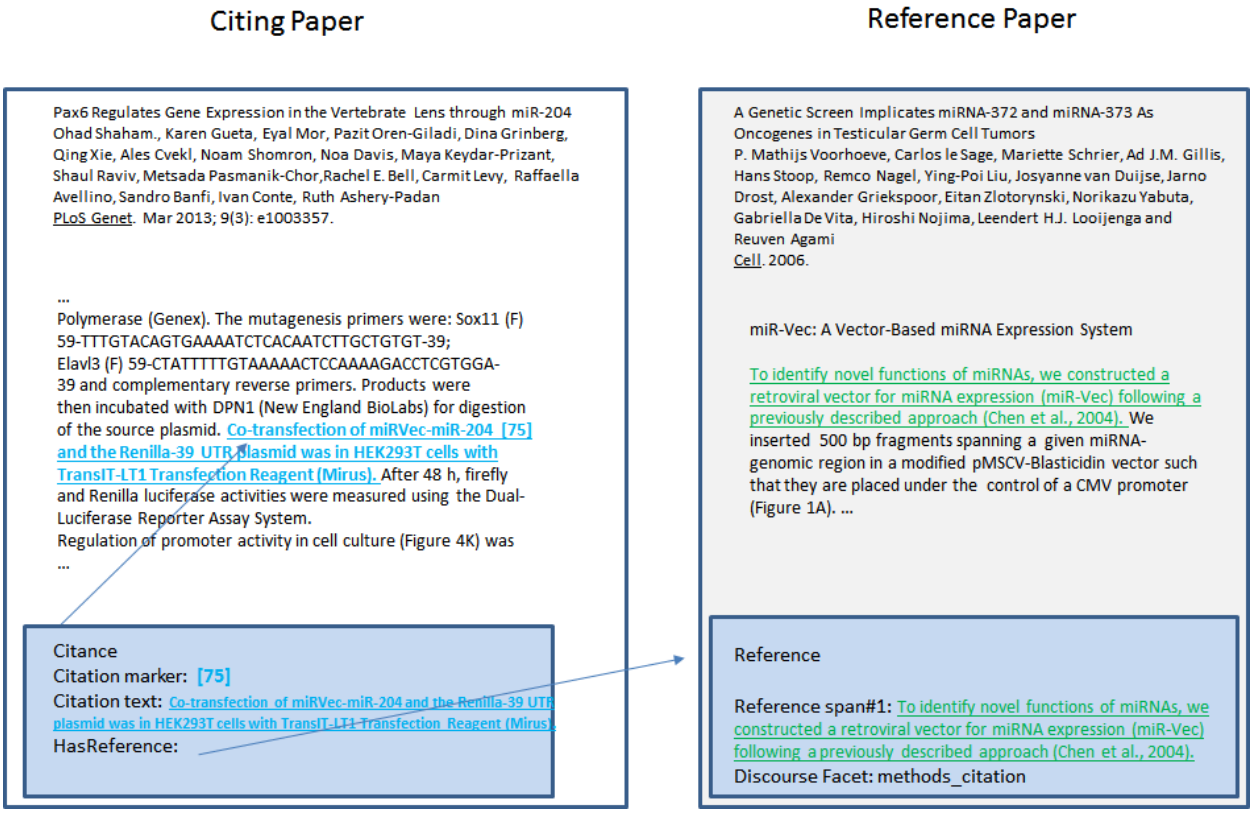
In particular, a citing paper may frequently include discussion of a reference paper in the sentences following the citation. Such discussion may be to express contrast to what was claimed or found in the reference paper, or it may express confirmation or a narrowing/broadening of the findings. We would like you to include all discussion that is pertinent to the reference paper in the citation text.

Citation marker: the citation to the reference paper, usually either (author, year) or [number in refs section]

Reference span: the portions of the reference paper that the authors of the citing paper are discussing. In print publications, there used to be citations not only to the reference paper, but also which paragraph(s) or line number(s) the citation was for. We are trying to reintroduce that level of granularity in this annotation effort. There may be up to three reference spans corresponding to a citance; if there are more, choose the most informative ones.

Facet: a facet is the specific part of the scientific method that the reference span is playing. A reference span might be part of the methods section, for example, or the results section. Currently, this is the set of facets that we identify: hypothesis, method, results, implication, discussion, and data-set-used.

Example from the Biology Domain:



Setting Up

First, you will install Protégé/Knowtator, following the instructions here (or http://knowtator.sourceforge.net/install.shtml).

If you have a Mac and you get an error about PowerPC no longer being supported, download the Unix version and use ls and cd commands at the terminal window (In Applications-Utilities-Terminal) to open up the folder where you have the Unix version saved. (ls lists the files in the current folder, cd folder1 changes the current folder to folder1 as long as folder1 is in the current folder. The command cd .. takes you up one folder level)

Then type the command

sh ./install_protege.bin

Download the contents of the dropbox directory onto your machine.

Here is what your directory will look like:

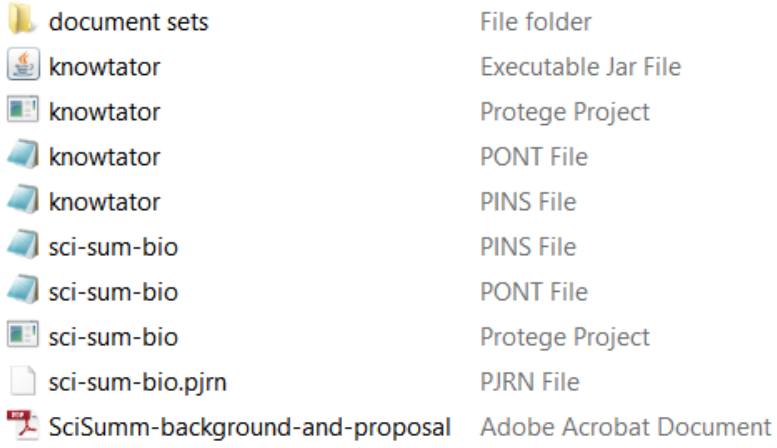


Figure 3. Main Directory structure. Note the 6 Knowtator files required in each task subdirectory. The sci-sum-bio project file is where you'll start your work.

Copy the file knowtator.jar to /plugins/edu.uchsc.ccp.knowtator, in the Protégé file directory (using a default installation, you'll put this in the directory: C:\Program Files (x86)\Protege_3.3.1\plugins\edu.uchsc.ccp.knowtator.

Next follow the directions you will be given for your own task subdirectories. Each subdirectory will correspond to one task. The name of the subdirectory will be the document name for the paper that we'll be studying in each task; in other words, the name of the subdirectory will be the name of the reference paper. Contained in the subdirectory will be the reference paper (the one corresponding to the directory name) and a set of citing papers. This is a subdirectory /text with the corresponding text files which is what Knowtator will use when you annotate.

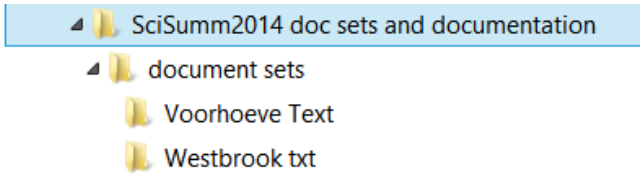
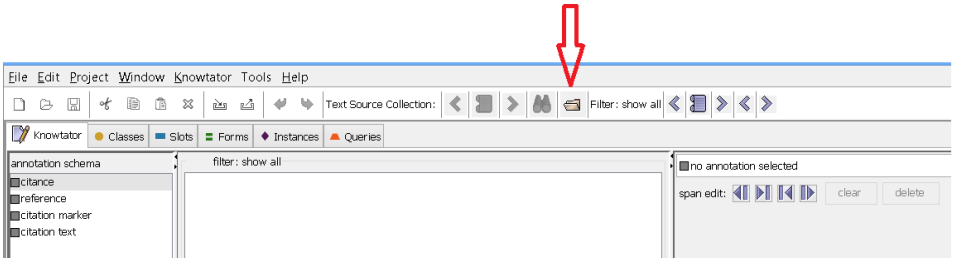


Figure 3. Directory structure where Voorhoeve and Westbrook are task directories.

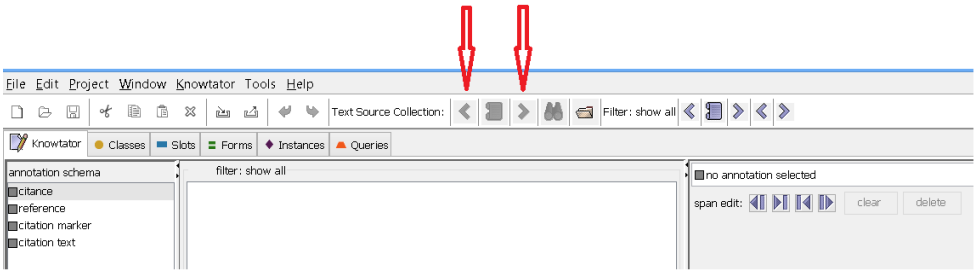
To begin work:

1. Double-click on sci-sum- bio protégé project (see figure 3 above)

2. Select the icon “open text source collection”



3. Select Local files from the pull down menu (it is the default choice), and then navigate to and open the subdirectory that you are assigned to work on. Verify that you can now scroll through all the documents in that directory using these two controls:



Annotation:

There are two steps to the annotation:

1) Identifying the citance. For each citing document, you will find the sentence(s) and the reference to the paper that is the target for the task (i.e. the paper that is the name of the task directory).
2) Identifying the reference span(s). For each citance, identify the span(s) of text in the reference paper that the citance refers to. For each such span, you will be indicating which facet of academic discourse the citance is referencing, whether it is hypothesis, method, results, implication, discussion or data-set-used.

- ❖ Remember: a reference document has the same name as the directory.
- ❖ Remember: a citing document does NOT have the same name as the directory and is one of the documents that cites the reference document. You will create citances for each citing paper. No reference paper will have any citance – reference papers will only be annotated for reference spans and discourse facets.

Process:

Identify citances

Annotate each citance in each of the citing papers. There may be more than one citance per citing paper, in cases where an author cites the reference paper multiple times. For citing papers that include a large number of citations to the reference paper, then select the 5 most contentful citances, i.e., select those citances that refer to specific elements of the reference paper and where the reference paper is not one of a long list of other reference papers. For example, a citance for Abney 1997 where (Abney 1997) was mentioned alone is more contentful than the following citation: "Examples of such techniques are Markov Random Fields (Ratnaparkhi et al., 1994; Abney, 1997; Della Pietra et al., 1997; Johnson et al., 1999), and boosting or perceptron approaches to reranking (Freund et al., 1998; Collins, 2000; Collins and Duffy, 2002).

Create a new citance annotation

In the following example, the full citance will be "**Co-transfection of miRVec-miR-204 [75] and the Renilla-39 UTR plasmid was in HEK293T cells with TransIT-LT1 Transfection Reagent (Mirus)**." We know [75] is (Voorhoeve 2006) by looking up in the references section what [75] refers to.

You may wish to use the Find/Annotate tool under the Knowtator tab to find the citations once you know how the author is referring to the target paper.

- Remember: citance text may be full sentences, multiple sentences, or partial sentences.

For the example below, the citance is the full sentence because, although the citation marker is in the first part of the sentence, the citation text is the full text since it is the co-transfection of miRVec-miR-204 with the Renilla-39 UTR plasmid which is pertinent to [75].

- **Co-transfection of miRVec-miR-204 [75] and the Renilla-39 UTR plasmid was in HEK293T cells with TransIT-LT1 Transfection Reagent (Mirus)**.

Consider also the following citance (38 is the same Voorhoeve article), where we suggest including the full sentence because while only the first portion of the text reports that (38, 39) use microarrays for detecting barcode changes, the second portion reports on an alternative method for barcode detection. As such, while reading the Voorhoeve Reference Paper, it would be interesting to know what alternative methods will be used in subsequent papers.

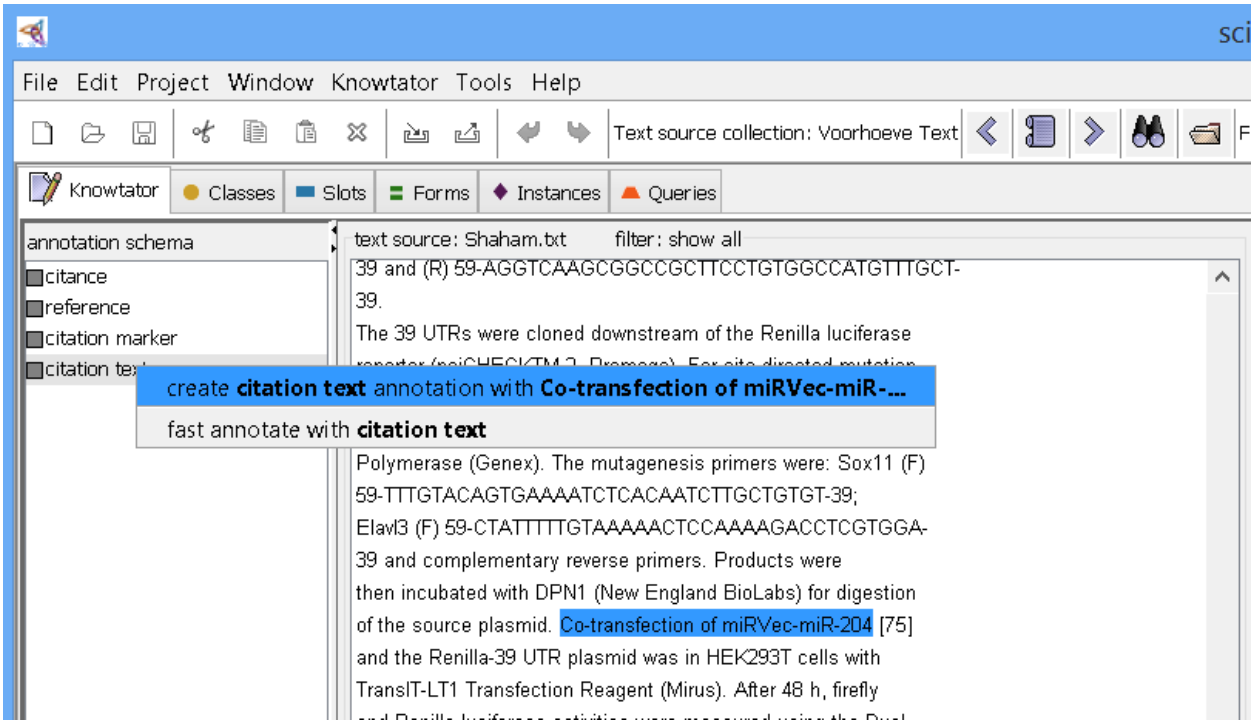
- Unlike previous efforts that used microarrays for detecting barcode changes (38, 39), we used a unique bead-based platform for barcode detection.

In addition, please include in your citation texts those sentences in the CP that express a positive or negative sentiment with respect to the reference paper. Such sentences are often found AFTER the citation marker itself. Here is an example, continuing from the citation text above:

- Unlike previous efforts that used microarrays for detecting barcode changes (38, 39), we used a unique bead-based platform for barcode detection. Compared with microarray or deep

sequencing, the bead-based detection platform is much more cost-effective, and is flexible enough to incorporate additional barcode detectors (1)

This sentence, beginning with “Compared with microarray …” is still pertinent to the Voorhoeve reference of the previous sentence, in that the author is stating that their method is superior to the microarray method of Voorhoeve. Such information would be valuable to someone reading Voorhoeve, and so should be included in the citation text.



- In most cases, the citing text is simply reporting on the reference span and reference text in a neutral, factual manner. There are some cases, however, where the citing text expresses the author’s sentiment about the reference text. After you’ve annotated the extent of the citing text, please consider whether the author is expressing a negative sentiment with regards to the reference paper. <u>If there is a negative sentiment, please make a note of the citing text and citing paper in your lab notebook – we would like to collect these examples for future research and your contribution will be much appreciated.</u>

  Here is an abstract example. We would expect the annotator to include both sentences in the citation text as the subsequent sentence still has discussion relevant to 53 and 54. As you can see the second sentence expresses the author’s view on 53 and 54, as having shortcomings, and so

expresses a negative sentiment. It would be much appreciated if you could make note of such citation text in your lab books for future research. Thank you!

- We followed the method for partial functional inactivation of p53 (see [53,54]). <u>However, it had a few shortcomings</u> as there are now more efficient …

  This phrase is expresses the author's sentiment

Create the citation text annotation **<u>first</u>**. Select that portion of the sentence that is the citation text, then select "citation text" from the annotation schema as below, and create the citation text annotation[1].

The citation text should include the citation marker in its span (even though we will indicate what the citation marker is separately as well.) Remember, a citation text will usually be a full sentence or multiple sentences, but it can also be a partial sentence when the full sentence discusses concepts not relevant to the reference paper.

- Here is an example of a case where the citation text is a partial sentence. The concern is that the same sentence contains two separate citations to the same reference paper.

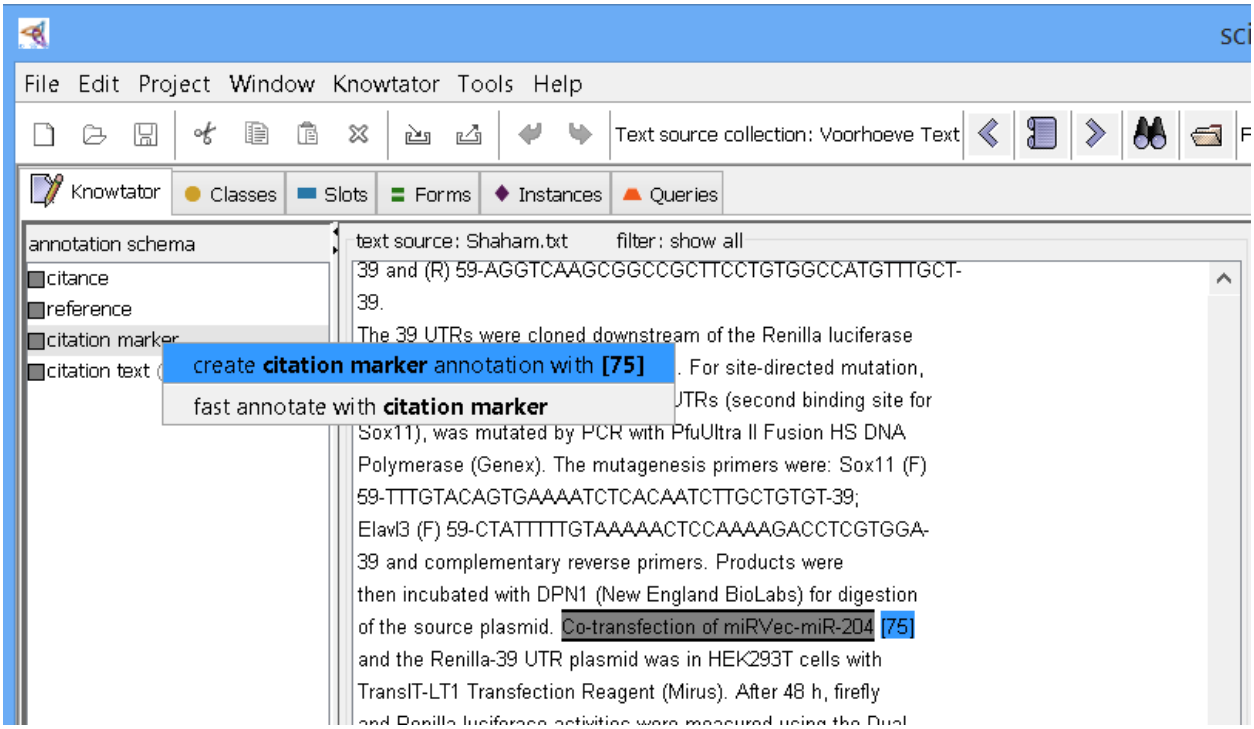
  - The following gene-specific sequences were used to generate siRNAs (Dharmacon): siBub3 50-AGCGACUGUGCCAAUUCCA-30; siMad2 50-GGAAGAGUCGGACCACAG-30; siCdc20 50-CGGCAGGACUCCGGGCCGA-30 [**11];** siCdh1 50-UGAGAAGUCUCCCAGUCAG-30 **[11];**

  If both citation markers had the same reference span(s) in the reference paper, then we would recommend creating a single citation text with the maximum citation text. However, in this case, the annotator finds different reference spans for each of the two citation markers and therefore needs to create two separate citances (each citance can only have one set of reference spans). The annotator therefore creates two instances of citation texts, resulting in citation text with partial sentences.

  - <u>The following gene-specific sequences were used to generate siRNAs (Dharmacon): siBub3 50-AGCGACUGUGCCAAUUCCA-30; siMad2 50-GGAAGAGUCGGACCACAG-30; siCdc20 50-CGGCAGGACUCCGGGCCGA-30 [**11];**</u> siCdh1 50-UGAGAAGUCUCCCAGUCAG-30 **[11];**
  - <u>The following gene-specific sequences were used to generate siRNAs (Dharmacon):</u> siBub3 50-AGCGACUGUGCCAAUUCCA-30; siMad2 50-GGAAGAGUCGGACCACAG-30; siCdc20 50-CGGCAGGACUCCGGGCCGA-30 [**11**]; <u>siCdh1 50-UGAGAAGUCUCCCAGUCAG-30 **[11];**</u>

[1] In this and the following screenshots, we have abbreviated the citation text for better presentation only. The FULL citation text is the entire sentence "**<u>Co-transfection of miRVec-miR-204 [75] and the Renilla-39 UTR plasmid was in HEK293T cells with TransIT-LT1 Transfection Reagent (Mirus)</u>**."

- Next, create a “citation marker” annotation in the same way. The citation marker should not include parentheses or brackets. Thus, the citation markers in the above examples are **75** or **Voorhoeve 2006**, not [75] or (Voorhoeve 2006).

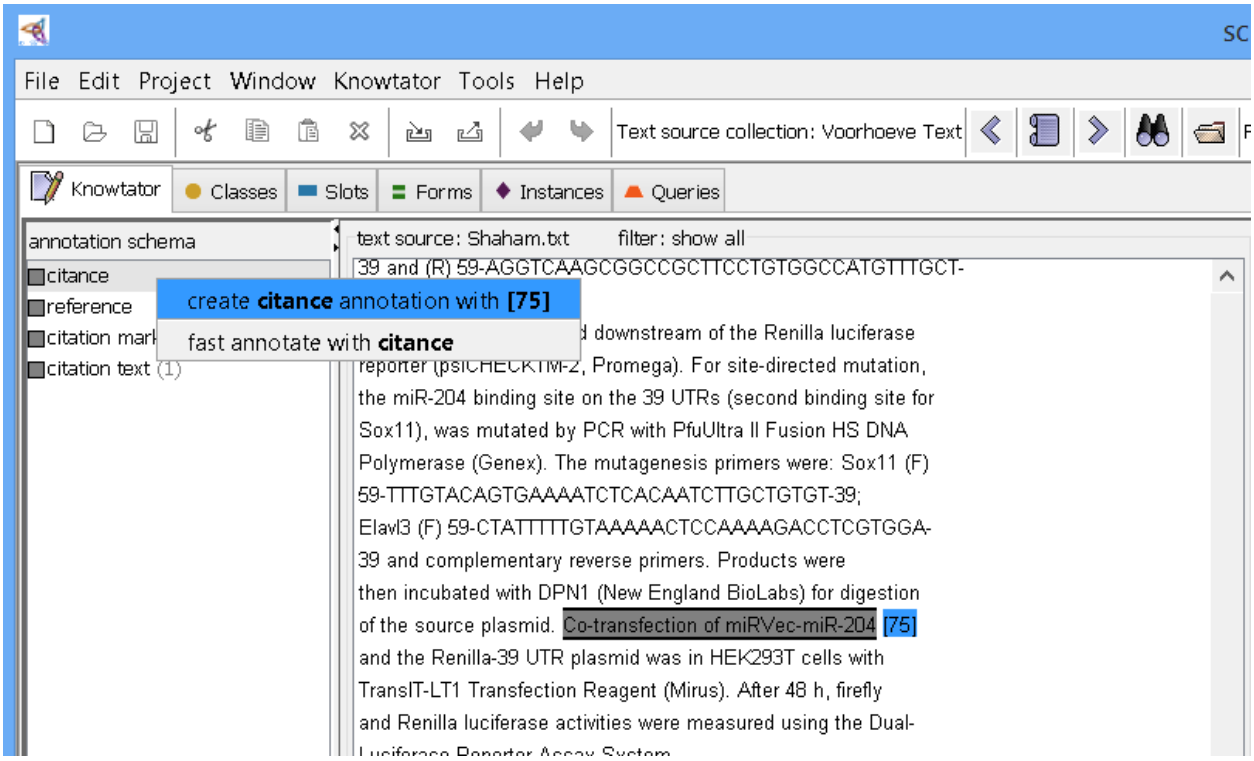


NOTE: in the above screenshot, the square brackets are included in the citation marker. We have since updated the annotation guidelines so that the square brackets should NOT be included in the citation marker span.

If the citation marker is included in a range of citation markers and is not explicitly mentioned in the text, please select the range of citation makers, not including the square brackets (or other parentheses markers). For example:

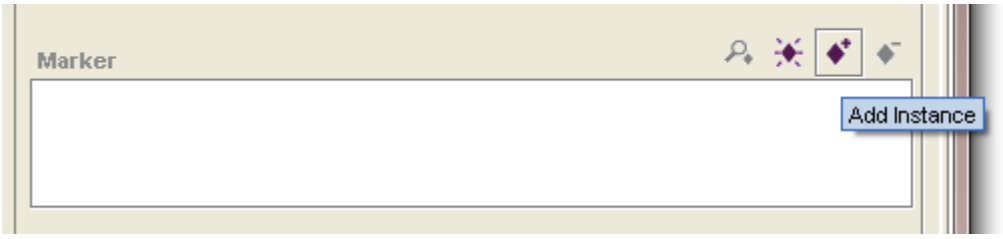
The emerging role of miRNAs in the regulation of fundamental set of cellular mechanisms such as proliferation, apoptosis, development, differentiation and metabolism [9–16] clearly suggests that any aberration in miRNA biogenesis pathway or its activity contributes to the human disease pathogenesis including cancer [17].

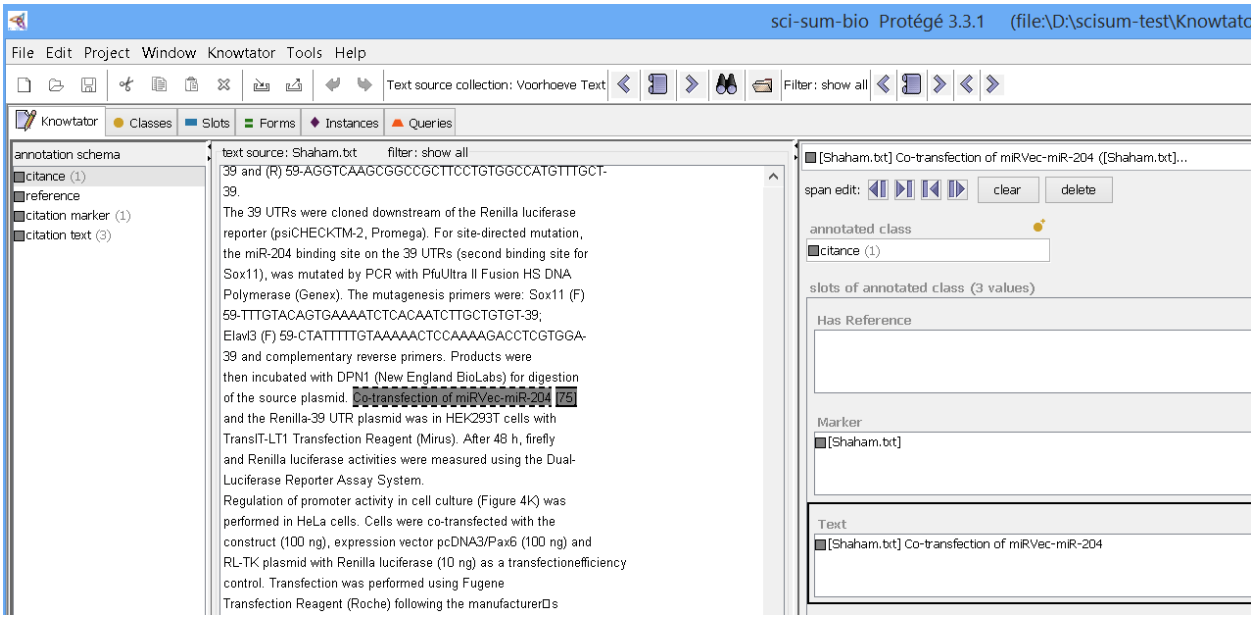
- Now create a “citance” annotation by selecting the node labeled “citance” in the annotation schema



- Fill in the “Marker” and “Text” slot values of the citance annotation, by choosing “add instance” on the far right – if you have only created one instance of a citation marker, it will auto-populate, or, it will present a drop-down list of markers for you to choose from. Similarly, for citation text.

❖ You may need to pull down the pane for adding your annotator information in order to see all three slots of the citance: be sure to fill in the marker and the text slot at this time.





Once all of the citances have been created for a given citing paper, it is time to mark the span(s) of text that each citance is referencing. To do so, it is important to read the full text of the reference paper before proceeding. Your job is to identify those passages in the reference text that best reflect what the citation text is talking about. Ask yourself: what parts of the reference paper is the citation text talking about? Where is the evidence for what the citing author has written?

Once you have read the reference paper, you will back go through each of the citances to identify the reference span(s) for the citance. When there are multiple sections of the reference paper that a citance could refer to – we ask that you identify up to 3 such reference spans if appropriate. Wherever possible, please do not limit the reference span(s) to text that appears in the Abstract of the reference paper.

! In the case that you absolutely cannot find ANY text in the paper that is specific to the reference and you feel very strongly that the reference is to the paper as a whole, you may indicate this scenario by marking the title of the reference paper as the reference span. However, we hope that you will endeavor to find specific sections of the reference paper if at all possible.

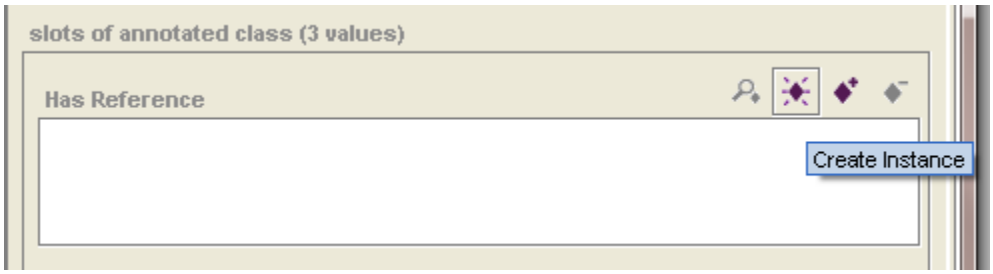
- ❖ Possible best practice may involve printing out the reference paper, making note of the reference spans for each citance, to make the identification of reference spans easier when using Knowtator.

Creating the reference annotation

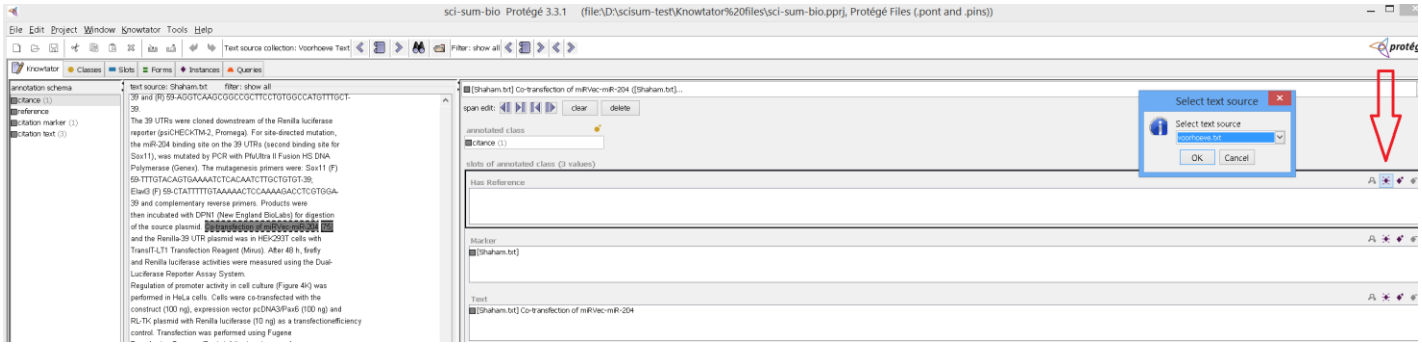
This part is a little bit tricky but you will soon get used to the work flow. Please follow along closely the first time!

- Select the citance annotation if it isn't already selected (i.e., you are annotating the citing paper). Remember to scroll the window down just a bit to see the TEXT field under the Marker field.

- Select the “create instance” button for the “has reference” slot

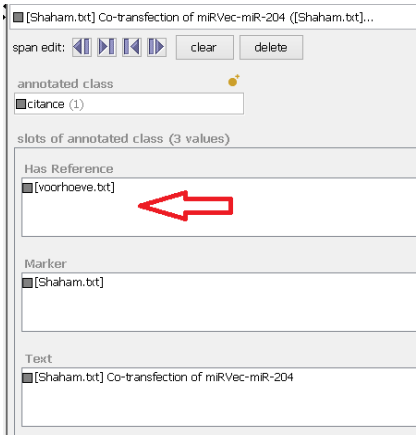


- Choose the reference paper for your document set from the drop down that pops up when selecting “create instance”

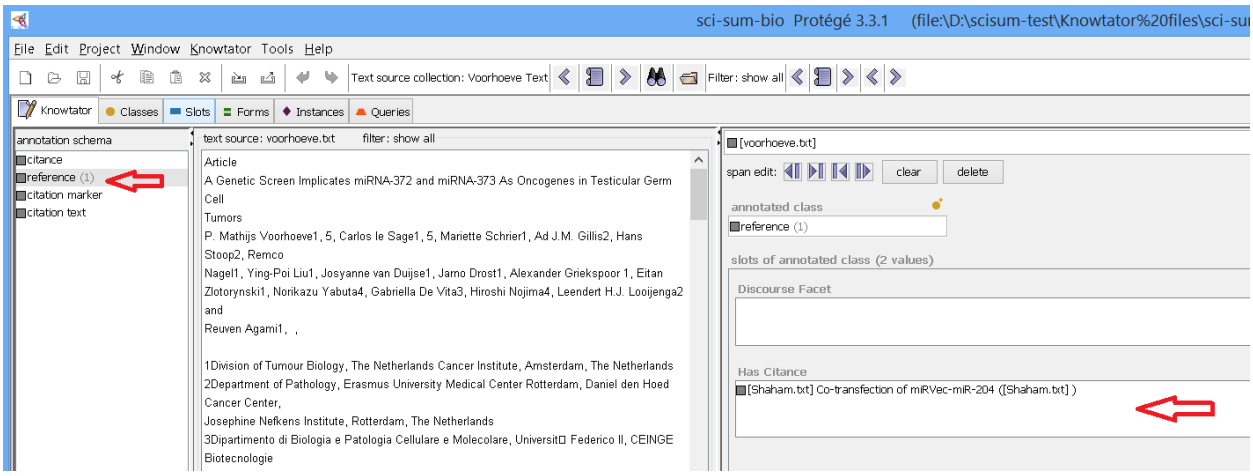


- At this point you have created a reference annotation that has no span that is correctly linked to the “Has Reference” slot of the citance annotation. Now we need to go to the reference annotation and add a text span to it.

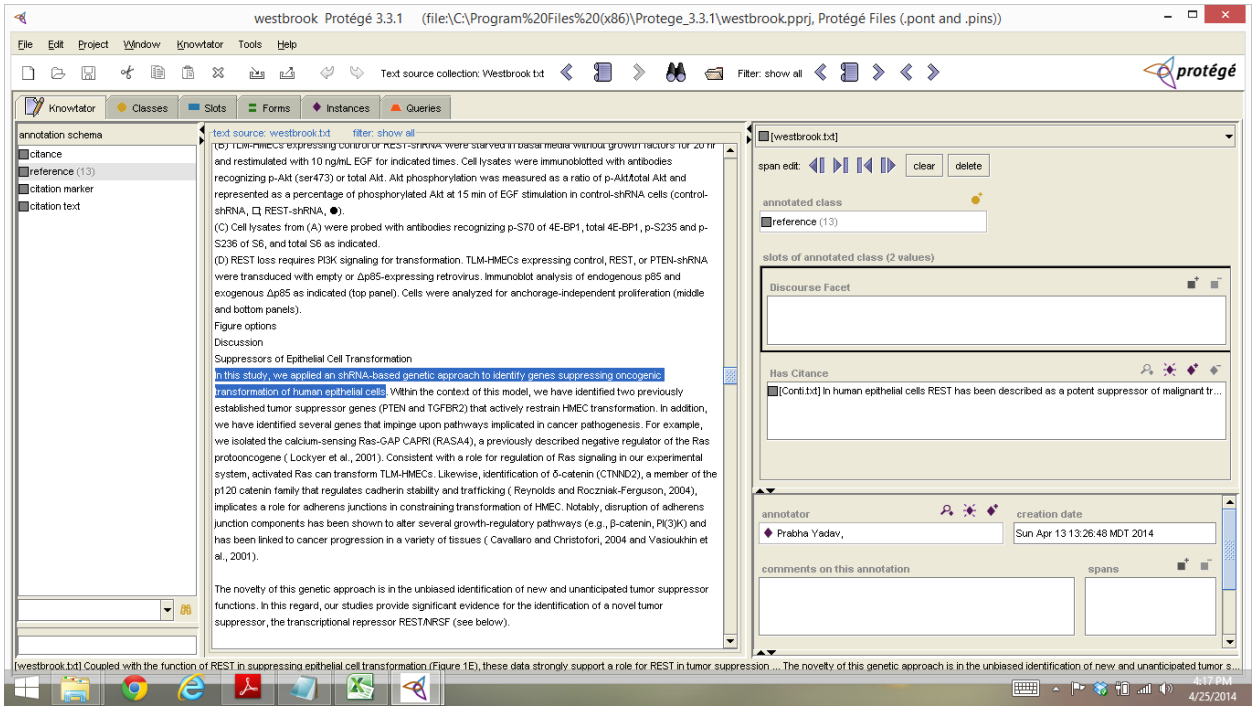
- Double click on the reference annotation filling the “Has Reference” slot, which now refocuses Knowtator to the reference paper.



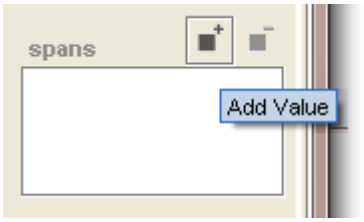
❖ Knowtator is now refocused on the reference paper, with one reference mention already created, with a slot HasCitance, which is the citance that you just created.



- Now, select a span of text that you want for the reference annotation

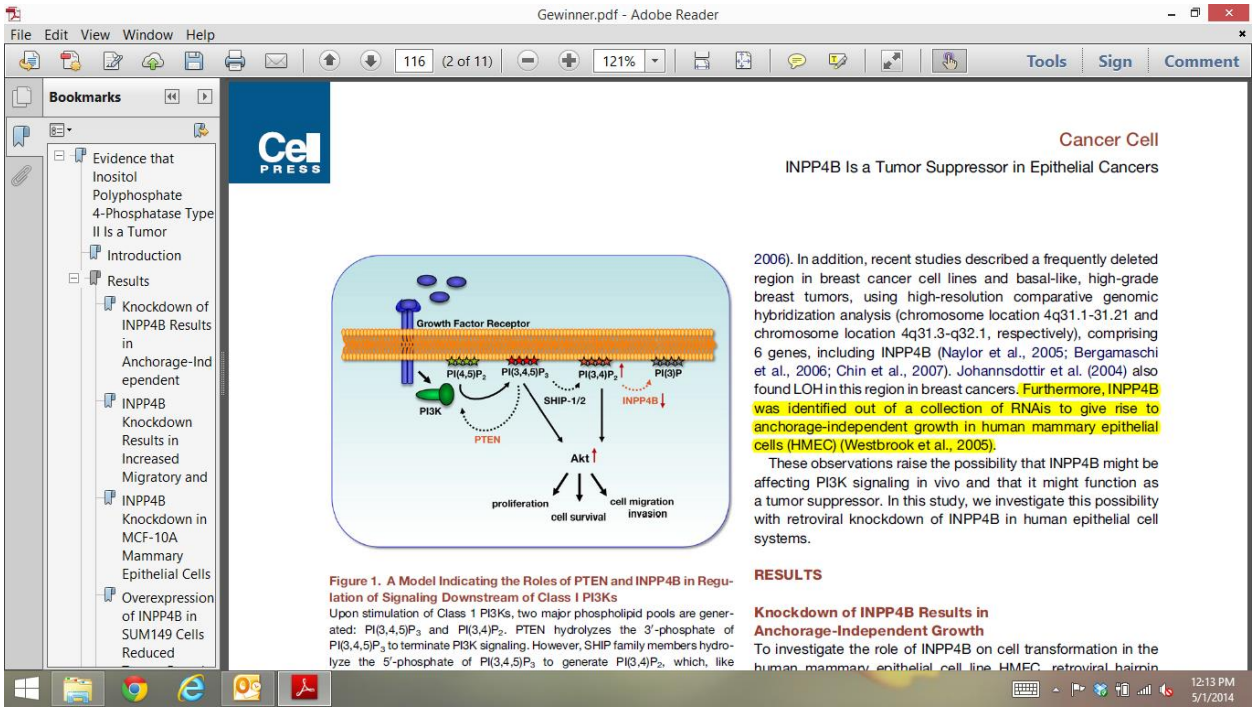


- Now click on the “add span” button near the lower right hand corner of the screen. There may be more than one reference span for a citance, to do so, you simply add another reference span using the same method as the first one. We do not recommend more than 3 reference spans, however, so if there are multiple spans, please select the 3 most informative.

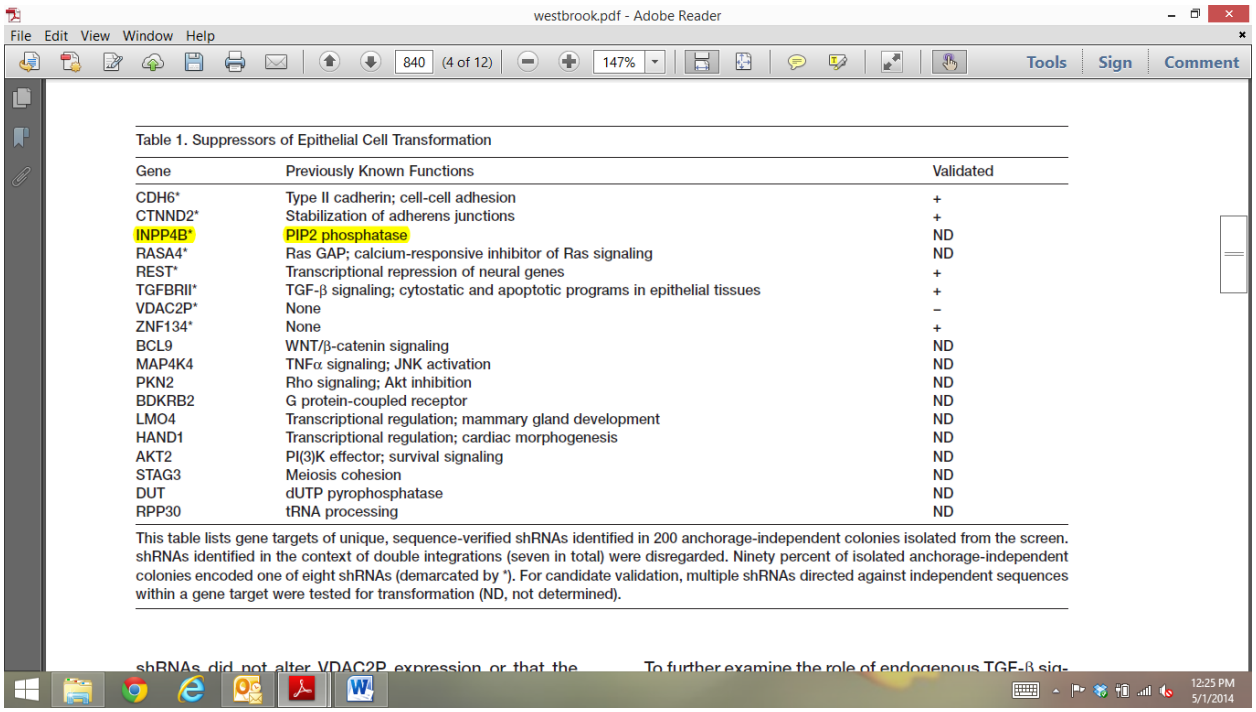


- **Tables as possible reference spans**: if you feel that a table is one of the best reference spans for a citing text, please consider the following:
  - If the table is described in the text, and that text would make a good reference span, please choose that text instead of the table itself.
  - If the text description is not specific enough to be a good reference span for the citing text, please consider whether the caption of the table would make a good reference span.
  - Only if neither the text nor the caption are specific enough to be a good reference span for the citing text, then select either the specific rows or the whole table as the reference span.

  In the citation below, Gewinner has referred to the table in Westbrook et al, 2005, but neither is the table described in the text nor is the caption on the table informative enough about how Westbrook 2005 is being cited. The citation text is highlighted below-



This citation in the Gewinner text refers to the following table in Westbrook Paper:



westbrook.pdf - Adobe Reader

Table 1. Suppressors of Epithelial Cell Transformation

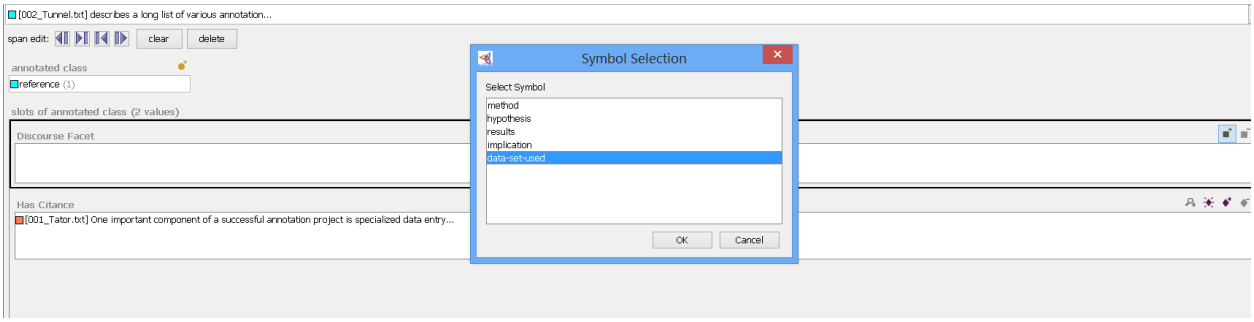
Gene	Previously Known Functions	Validated
CDH6*	Type II cadherin; cell-cell adhesion	+
CTNND2*	Stabilization of adherens junctions	+
INPP4B*	PIP2 phosphatase	ND
RASA4*	Ras GAP; calcium-responsive inhibitor of Ras signaling	ND
REST*	Transcriptional repression of neural genes	+
TGFBRII*	TGF-β signaling; cytostatic and apoptotic programs in epithelial tissues	+
VDAC2P*	None	–
ZNF134*	None	+
BCL9	WNT/β-catenin signaling	ND
MAP4K4	TNFα signaling; JNK activation	ND
PKN2	Rho signaling; Akt inhibition	ND
BDKRB2	G protein-coupled receptor	ND
LMO4	Transcriptional regulation; mammary gland development	ND
HAND1	Transcriptional regulation; cardiac morphogenesis	ND
AKT2	PI(3)K effector; survival signaling	ND
STAG3	Meiosis cohesion	ND
DUT	dUTP pyrophosphatase	ND
RPP30	tRNA processing	ND

This table lists gene targets of unique, sequence-verified shRNAs identified in 200 anchorage-independent colonies isolated from the screen. shRNAs identified in the context of double integrations (seven in total) were disregarded. Ninety percent of isolated anchorage-independent colonies encoded one of eight shRNAs (demarcated by *). For candidate validation, multiple shRNAs directed against independent sequences within a gene target were tested for transformation (ND, not determined).

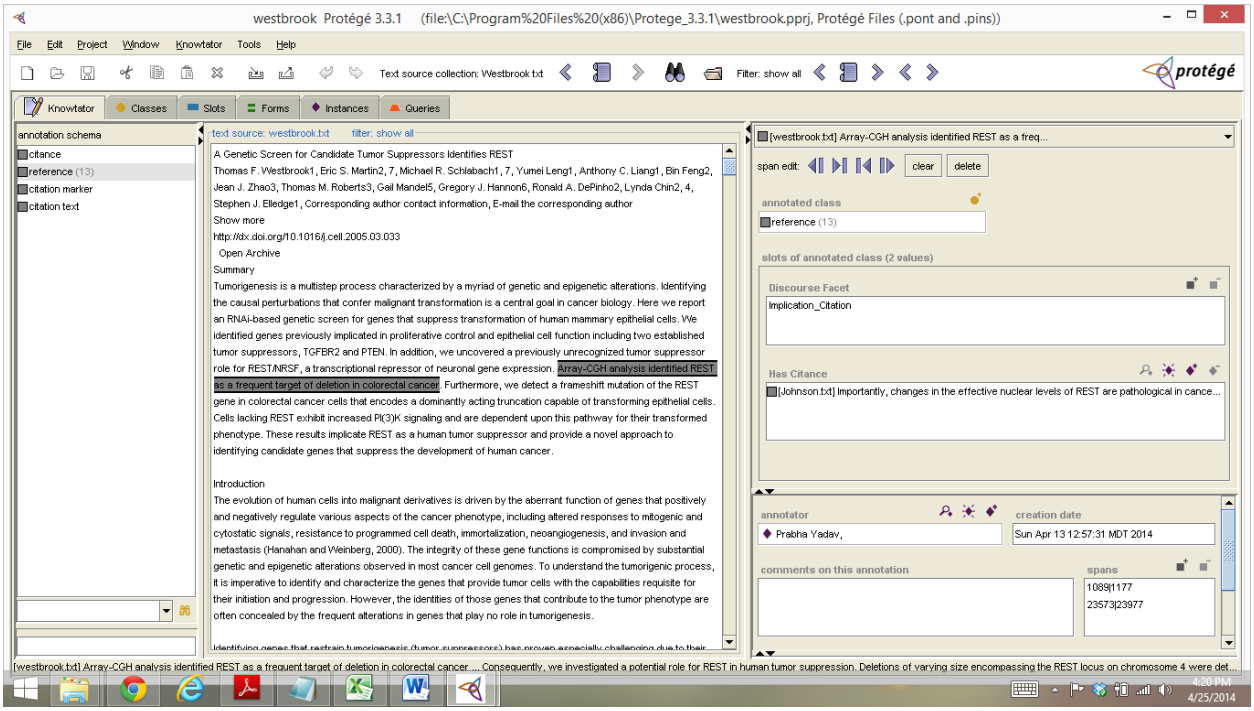
The table row is the only mention of INPP4B in the Westbrook Paper.

- Finally, for each reference, you must consider what part of the scientific discourse the text is part of. You do that by navigating to the discourse facet, clicking on “add value” and then selecting from the drop-down box of discourse facets.

  We intend for these discourse facets to be self-explanatory. We are asking you here to identify what part of the scientific argument the selected text is part of. Please note that we are not asking you to identify in which section of the paper the text occurs in! You can rely on your scientific training to determine whether the text discussed is a method, a hypothesis, an implication, a result, or part of general discussion; please select the most specific of these classes.



- That's it! The reference annotation is done and you can navigate back to the citance annotation by double clicking on it



Final Steps:

Writing the summary:

a. Now that you have a really good understanding of the reference paper, write a 250 word summary (NO MORE THAN 250 WORDS) of the reference paper, taking into consideration not only the abstract for the reference paper itself, but now also those parts of the reference paper that were mentioned by the citing papers and how the citing papers discussed the reference paper. This summary will highlight the merits of the paper and the discussion that references the paper subsequently.

b. Please save your summary as a txt file with the task-id, followed by "_summ" as the name of the doc.