# Supplementary Data
# for
# "mirPub: a database for searching microRNA publications"

Section 1 presents mirPub's Web Interface and describes its main functionality. Section 2 discusses supplementary details about capturing the changes in miRNA data. Section 3 summarises interesting statistics related to miRNA publications and compares mirPub's database to other important databases that store miRNA publications. Finally, Section 4 presents a preliminary evaluation on the usefulness of mirPub in literature search for particular miRNAs.

## 1. MirPub's Web Interface

MirPub's Web Interface is freely available at http://www.microrna.gr/mirpub/. Figure 1 illustrates a snapshot of this interface.

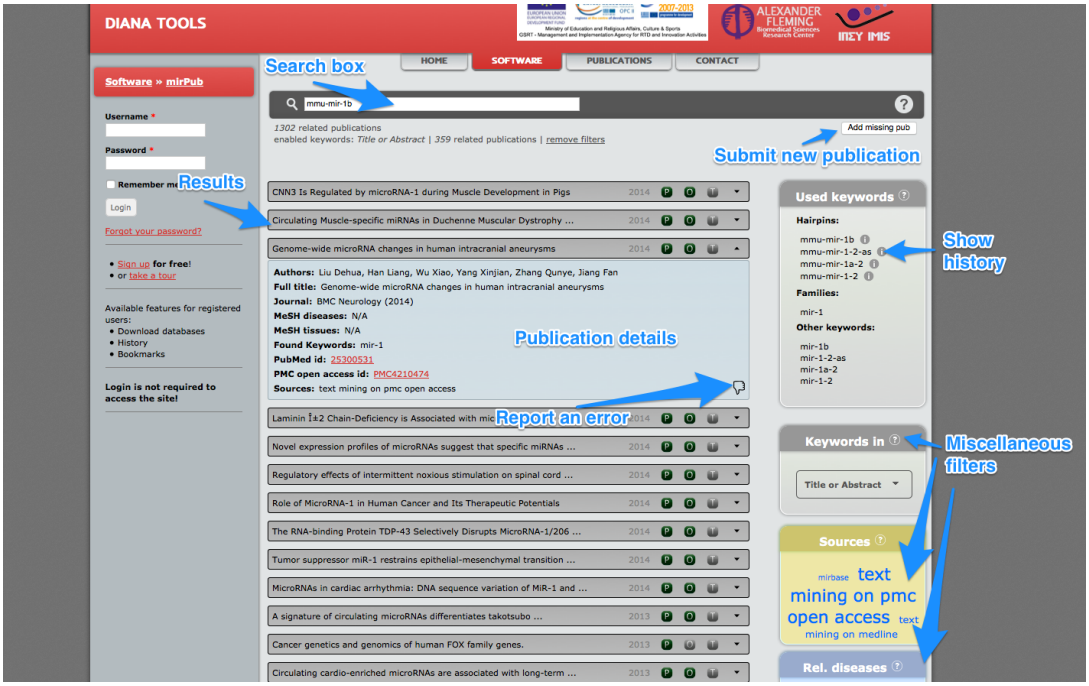


**Figure 1 - MirPub's Web Interface.**

Users provide keywords describing miRNAs in the search box on the top of the interface and press 'Enter'. As a result, relevant articles appear in a list below the search box. Each article is represented by a box that contains its title and its publication date along with three hyperlinks. The first redirects to the abstract of the article in PubMed, the second to its full text (in case it is open access), and the third to the TarBase page that contains the miRNA-to-gene interactions examined in the article. More details about the article (e.g., the list of authors, the Journal in which it was published, etc) can be revealed by clicking on the arrow button at the right of its entry.

MirPub expands the set of user keywords to contain the families of the identified miRNAs and miRNA name variants as well, based on history and modification rules. Therefore, mirPub's results contain also articles which are associated to

these extra keywords (family and variants). Of course, the associations to such keywords is based either on curated data, or on text mining performed on MEDLINE. The complete set of keywords used for each search is displayed in a box at the right of the results, titled "Used keywords", and the user is able to keep only a subset of them filtering out the irrelevant publications (see Figure 1). In particular, this box contains a hyperlink for each one of the terms used during searching for relevant publications. If the user believes that only a subset of these terms is appropriate, the user can keep only the articles which are connected to them by clicking on their hyperlinks in the box. By clicking again on a hyperlink the user deselects the corresponding term.

To help its users to figure out which articles are more relevant to a miRNA, mirPub provides a filter that differentiates the articles based on the position in the manuscript where the miRNA of interest is mentioned. The user can select one of the options provided by the dropdown list which is located in the box titled "Keywords in:" at the right of the user interface. The user can keep only articles that contain a given keyword in their title, or articles that contain the keyword either in their title or abstract, or anywhere in the manuscript. The option "Not specified" retrieves all the articles.

MirPub also provides the timeline of changes of any mature or hairpin miRNA. In particular, each keyword associated to a mature or hairpin miRNA is accompanied by an information button which activates a pop-up window that visualizes the name and sequence changes related to each particular molecule. Figure 2 illustrates the contents of this pop-up for two terms “hsa-mir-29b-1” and “hsa-mir-98”.

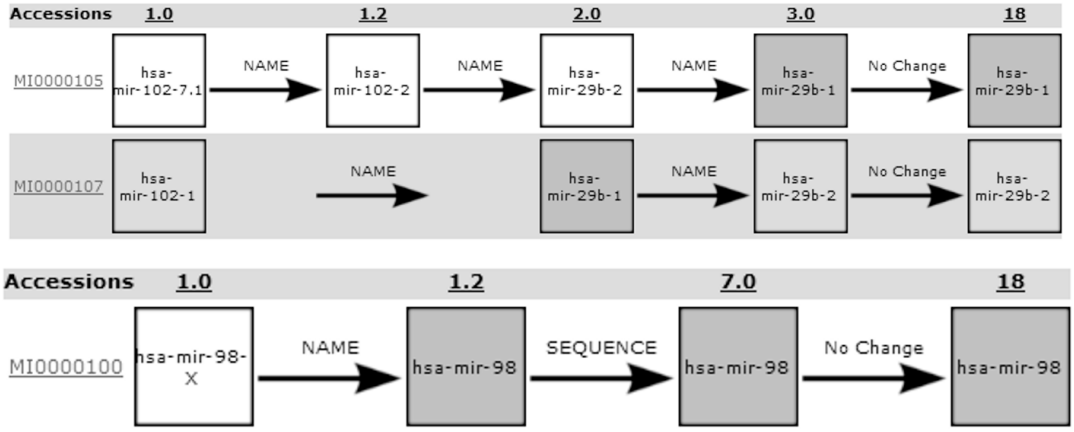


**Figure 2 – The history timelines of term “hsa-mir-29b-1” (involving hairpins MI0000105 and MI0000107) and “hsa-mir-98” (involving hairpin MI0000100).**

Another feature provided by mirPub is a set of tag clouds summarising MeSH diseases, tissues and cells that are relevant to the displayed publications, giving an insight about the role of the queried miRNAs (see Figure 3).

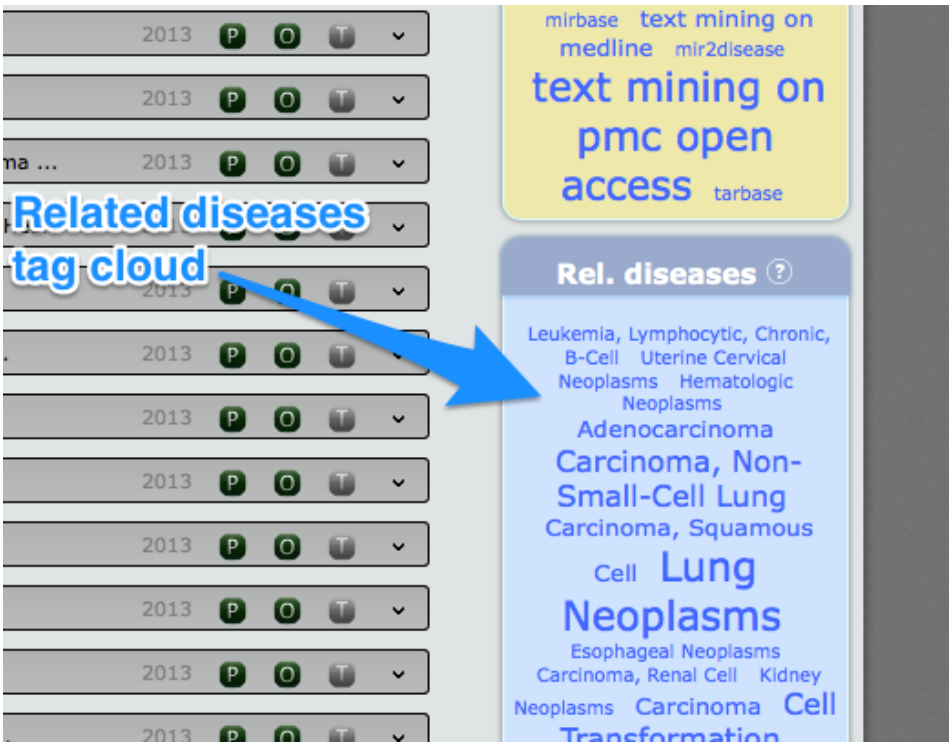


Figure 3 - Snapshot of the tag cloud for related diseases.

In case users believe that the results contain an article that is irrelevant to their keywords, they can report this by clicking on the "Thumbs down" button that is located in the detailed view of the corresponding box (see Figure 1). A new page is loaded containing a form that collects information about the reason why the user believes that there is no relation between the publication and the miRNA (see Figure 4).

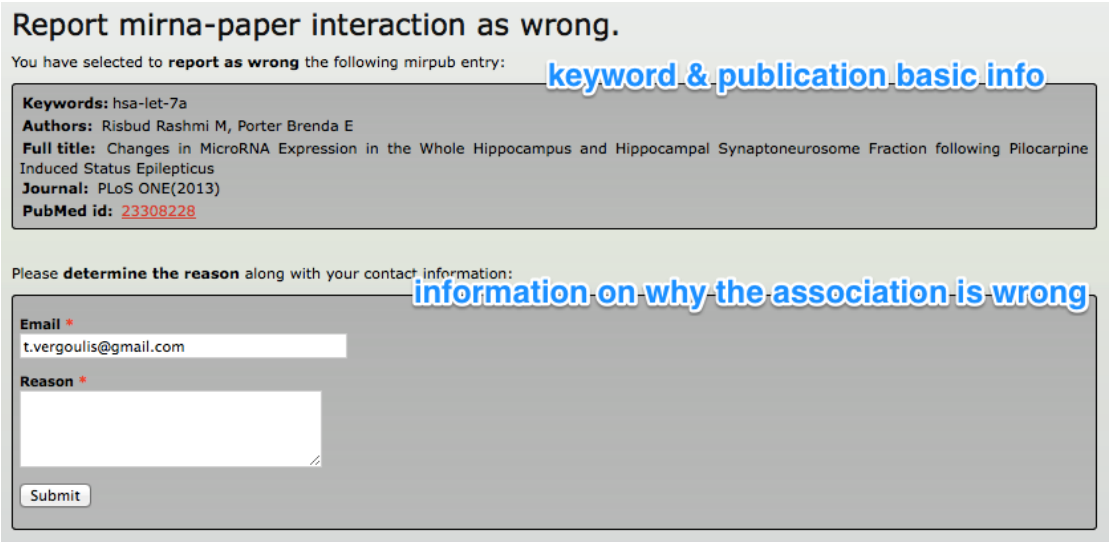


Figure 4 - User Interface for reporting erroneous miRNA-publication associations.

Furthermore, users can also add new miRNA related publications by clicking on the "Add missing pub" button located at the right of mirPub's main page, below the search box (see Figure 1). Then, a page containing a form similar to this described above for the case of reporting erroneous miRNA-to-publication

associations is loaded and the user enters supporting evidence about the miRNA-to-publication association to be added.

Every report for a new submission or error is recorded in mirPub's database, along with reporter's contact information. An expert curator periodically examines the list of reports and decides whether mirPub's data should be modified. If it is necessary, the curator contacts the reporter for details.

## 2. Capturing changes in miRNA data

Since miRNA research is in flux, new publications amend the results of previous work, thus, the content of related databases, like miRBase, should be modified to comply with these changes. As many of these changes involve the name of miRNAs, keeping track of the naming history of a miRNA molecule is important for any researcher searching for publications related to this molecule. This is because knowing older names of the molecule or other molecules having the same name in the past can help the researcher to both expand her search in order to get more results, and remove non-related papers from the result set. mirPub has processed all available miRBase versions to extract useful miRNA data evolution information. The following paragraphs elaborate more on this.

In miRBase, each mature and each hairpin has a name and a sequence, which can be different from version to version, and an accession number, which is fixed and identifies it uniquely through different miRBase versions. All these data, together with metadata information, are stored in versioned files with the file extension .dat, called, from now on, the dat files of miRBase. Dat files follow a format similar to EMBL format, having one record for each distinct hairpin accession. Each record contains the name, the sequence, the produced mature sequences, the related publications, etc., of one particular hairpin. The mature data are stored as sub-entries of the hairpins that produce them.

By comparing the dat file of a version to the dat file of the next version, all the hairpin and mature changes from one version to the other can be produced. miRBase also provides a diff file for every version, which can be processed to extract the same information. However, available diff files exist only for miRBase versions newer than 3.1. Therefore, all miRNA history data in mirPub are produced by comparing dat files of miRBase versions to eachother (from versions 1.0 to 18). We used the provided diff files only to validate the results of the aforementioned analysis for versions newer than 3.1. Note that, miRBase does not provide accession numbers for matures in versions older than 6.0. Therefore, we capture changes that involve mature data only for miRBase versions newer than 5.1.

After analysing all miRBase files, from version 1.0 to 18, we identified the following change types:

- For mature miRNAs and hairpins:
  - NEW: a novel miRNA has been inserted in the current version of miRBase
  - NAME: a miRNA name has been modified
  - SEQUENCE: a miRNA sequence has been modified

- NAME-SEQUENCE: a miRNA has both its name and its seqeuence been modified
- DELETE: a miRNA has been removed from miRBase, and its accession become obsolete

• Only for hairpins:
- FORWARD: a miRNA accession is replaced by another one

• For mature-hairpin related pairs:
- ADD MATURE-HAIRPIN ASSOC: a mature was found to be produced by a particular hairpin
- REMOVE MATURE-HAIRPIN ASSOC: a mature was found to not be produced by a particular hairpin

We parsed all miRBase files, extracted data related to changes that occur in miRNAs, and stored these data in a relational database.

mirPub users can explore changes in miRNA names, sequences, etc. with a tool that visualizes the timeline of those changes (see Figure 2). This tool is accessible through the information button that accompanies each reference of a miRNA name in mirPub's user interface. Clicking on this button results in rendering a pop-up that contains the history of all matures and hairpins that have been associated with this name for at least one miRBase version.

Finally, an interesting observation is that a miRNA change may often trigger another one. For instance, the insertion of a new hairpin in miRBase is usually followed by the insertion of at least one mature miRNA. In fact, our analysis shows that 86.29% of the NEW mature changes follow a NEW hairpin change.

## 3. Interesting statistics

Currently, mirPub's database contains more than 210,000 distinct <miRNA_keyword,PubMed_id> pairs involving more than 19,800 distinct publications. Table 1 compares mirPub with the most known databases containing publications relevant to miRNAs in terms of the number of papers found to be associated with at least one miRNA. It is clear that mirPub contains the largest number of miRNA related publications compared to other databases.

*Table 1. Comparison of several miRNA publication databases*

| | #papers |
|---|---|
| **mirPub** | 19,839 |
| **miRBase** | 407 |
| **TarBase** | 1,392 |
| **mir2disease** | 519 |
| **miRCancer** | 573 |

Note that mirPub is not supposed to replace the databases included in Table 1. Each of the presented databases has a special role for miRNA researchers (e.g., miRBase collects articles about miRNA discovery, TarBase papers about experimentally validated targets, etc.). In fact, mirPub's main objective is to assist curators of such databases to work more conveniently in keeping their

resources up-to-date. Therefore, the comparison to these databases is presented just to provide an insight about the amount of articles included in mirPub's database.

Another interesting statistics is that, among the 210,317 distinct miRNA-to-publication associations contained in mirPub's updated database, 43,366 of them come from curated data.

Table 2 summarizes some interesting statistics for the publications stored in the mirPub's database. In particular, the maximum and average number of miRNA keywords, MeSH diseases, and MesH tissues & cells, per paper are presented. One paper contains 2,510 distinct miRNA keywords. This is due to the fact that this paper is a large study of miRNA expression (Landgraf et al., 2007) done by sequencing over 250 small RNA libraries from 26 different organ systems and cell types. Thus, it contains a large number of miRNA related keywords. Table 2 also presents the maximum and average number of sources used to retrieve each paper. Note that, mirPub uses 5 data sources: MEDLINE/PubMed, miRBase, TarBase, mir2disease, and user contributed data.

*Table 2. Some mirPub statistics*

| | Max. | Avg. |
|---|---|---|
| **miRNA keywords/paper** | 2,510 | 10.6012 |
| **diseases/paper** | 8 | 1.86 |
| **tissues & cells/paper** | 8 | 1.58 |
| **sources/paper** | 5 | 1.2232 |

## 4. Evaluating literature retrieval

The strength of mirPub lies in the large set of miRNA-to-publication associations it contains. The most of these associations cannot be retrieved by PubMed itself because it hasn't access to all the curated data mirPub has, and it cannot exploit the variants and the history of the miRNA names.

To evaluate mirPub's effectiveness against PubMed in retrieving miRNA literature we examined the results retrieved given a query set of 50 randomly selected miRNA names. The complete set of miRNA names used for this experiments can be found in the caption of Table 3. We configured the search engines of both mirPub and PubMed to retrieve, for the same time period, only results that contain exact occurrences of the search terms. In mirPub this can be easily done by selecting the appropriate keywords from the ``Used keywords'' filter (see Figure 1). Note that for PubMed we additionally exploit the miRNA-to-paper associations contained in the NCBI gene database since these data are also accessible through the PubMed result page. Because of our experimental setting (i.e., "exact" queries) most differences in mirPub's and PubMed's results are expected due to the larger amount of curated data contained in mirPub's database.

The results are presented in Table 3. Consider that, under our scenario, both mirPub and PubMed have 100% precision. This means that all the retrieved

papers are related to user's search. The actual size of recall cannot be measured, since the entire set of related publications is unknown. However, the fact that mirPub returns more results than PubMed allows us to conclude that mirPub has higher recall. In particular, based on the number of retrieved papers, mirPub is expected to have recall that is more than twofold PubMed's recall.

*Table 3. Evaluation of mirPub's effectiveness in miRNA literature retrieval. MiRNA names used for this experiment were: api-mir-315, der-mir-9a, dps-mir-2517a-4, zma-MIR397b, pma-mir-135a, mml-mir-133c, gma-MIR4372, fru-mir-190, osa-MIR395t, aca-mir-5467, aga-mir-282, tcc-MIR172a, gga-mir-3539, hsa-mir-576, bmo-mir-3314, gga-mir-1397, mtr-MIR5267l, tgu-mir-2996-2, gso-MIR3533b, hsa-mir-432, dse-mir-7, hsa-mir-9-1, hsa-mir-4476, dre-mir-196b, bma-mir-133, hsa-mir-200a, dre-mir-734, bta-mir-2478, bmo-mir-2778a-2, mmu-mir-181d, ssl-MIR395, hsa-let-7b, hsa-mir-429, ppy-mir-518d, dre-mir-206-1, osa-MIR1438, dsi-mir-987-2, ppc-mir-2238c, ptr-mir-759, dme-mir-277, ppy-mir-216b, vun-MIR164, mmu-mir-105, dpu-mir-252b, tae-MIR1121, bdi-MIR156d, mml-mir-490, cin-mir-4114, ptr-mir-26a-1, bmo-mir-3414*

| | #papers |
|---|---|
| **mirPub** | 382 |
| **PubMed** | 176 |

We also evaluated how knowing miRNA data history improves retrieval effectiveness. In particular, we asked an expert (a senior PhD student performing research in the field of miRNAs for the last 3 years) to use mirPub in order to search for 20 miRNAs. For the needs of the experiment we selected miRNAs having significant history. First we asked him to configure mirPub's filters in the way he believes that he will get the most relevant to his search results. Then, we asked him to examine the history of all the matures and hairpins in the "Used keywords" box and, then, reconfigure the filters. Finally, he should provide judgements for the relevance of the retrieved results. Table 4 summarizes the results of this experiment.

*Table 4. Evaluation of the improvement in literature search by the use of miRNA history knowledge*

| | #papers |
|---|---|
| **retrieved w/o ev. knowl.** | 90 |
| **retrieved with ev. knowl.** | 143 |
| **relevant** | 129 |

It is evident that being informed about the history of miRNAs can be very useful in miRNA literature search. In particular, searching without knowing miRNA history failed to retrieve the 30.23% of the relevant papers found otherwise. Moreover, the evolution knowledge helps the user to fix errors that rise during his search (e.g., the use of old keywords that refer to modified sequences) preserving high levels of precision (approx. 90.2% for our experiment).

## 5. References

Landgraf,P. et al. (2007) A mammalian microRNA expression atlas based on small RNA library sequencing, Nucleic Acids Res, **129(7)**, 1401-1414.

Xie,B. et al. (2013) MirCancer: a microRNA-cancer association database constructed by text mining literature, *Bioinformatics*, **29(5)**, 638-644.