*Nucleic Acids Research, 2017* **1**
*doi: 10.1093/nar/gkx854*

# miRandola 2017: a curated knowledge base of non-invasive biomarkers

**Francesco Russo[1,*], Sebastiano Di Bella[2], Federica Vannini[3,†], Gabriele Berti[3,†], Flavia Scoyni[4,†], Helen V. Cook[1], Alberto Santos[1,5], Giovanni Nigita[6], Vincenzo Bonnici[7], Alessandro Laganà[8], Filippo Geraci[9], Alfredo Pulvirenti[10], Rosalba Giugno[7], Federico De Masi[11], Kirstine Belling[1], Lars J. Jensen[1], Søren Brunak[1], Marco Pellegrini[9] and Alfredo Ferro[10]**

[1]Disease Systems Biology, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, 2200, Denmark, [2]Business Unit Oncology, Nerviano Medical Sciences, Milan, 20014, Italy, [3]Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa, 56127, Italy, [4]University of Eastern Finland, Kuopio, 72010, Finland, [5]Clinical Proteomics, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, 2200, Denmark, [6]Department of Cancer Biology and Genetics, The Ohio State University, OH 43210, USA, [7]Department of Computer Science, University of Verona, Verona, 37134, Italy, [8]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029-6574, USA, [9]Institute of Informatics and Telematics (IIT), National Research Council (CNR), Pisa, 56124, Italy, [10]Department of Clinical and Experimental Medicine, University of Catania, Catania, 95125, Italy and [11]Department of Bio and Health Informatics, DTU Bioinformatics, Technical University of Denmark, Lyngby, 2800, Denmark

Received July 6, 2017; Revised August 28, 2017; Editorial Decision September 12, 2017; Accepted September 13, 2017

## ABSTRACT

**miRandola (http://mirandola.iit.cnr.it/) is a database of extracellular non-coding RNAs (ncRNAs) that was initially published in 2012, foreseeing the relevance of ncRNAs as non-invasive biomarkers. An increasing amount of experimental evidence shows that ncRNAs are frequently dysregulated in diseases. Further, ncRNAs have been discovered in different extracellular forms, such as exosomes, which circulate in human body fluids. Thus, miRandola 2017 is an effort to update and collect the accumulating information on extracellular ncRNAs that is spread across scientific publications and different databases. Data are manually curated from 314 articles that describe miRNAs, long non-coding RNAs and circular RNAs. Fourteen organisms are now included in the database, and associations of ncRNAs with 25 drugs, 47 sample types and 197 diseases. miRandola also classifies extracellular RNAs based on their extracellular form: Argonaute2 protein, exosome, microvesicle, microparticle, membrane vesicle, high density lipoprotein and circulating. We also implemented a new web interface to improve the user experience.**

## INTRODUCTION

miRNAs are small non-coding RNAs (ncRNAs) (21–23 nt long) that regulate gene expression at the post-transcriptional level by binding to messenger RNAs (mRNAs) and inhibiting their translation into proteins or by binding to other ncRNAs (1).

First discovered in 1993 in *Caenorhabditis elegans* (2), miRNAs had tremendous impact on the study of gene expression regulation and regulatory networks. Since their discovery as post-transcriptional regulators (3), specific links have been discovered between miRNA and human pathologies (4–7), and further studies indicate the utility of some miRNAs as biomarkers for cancer and other diseases (5). Some miRNA-targeted therapeutics have been tested in clinical trials, including a miRNA mimic of the tumor suppressor miR-34, which reached phase I clinical trials for treating cancer, and antimiRs for miR-122, which reached phase II trials for treating hepatitis (8).

Recently, miRNAs were shown to be present in human body fluids (9), resulting in the potential use of these small RNAs as non-invasive biomarkers (9,10). The great potential of extracellular miRNAs as biomarkers is their high sta-

*To whom correspondence should be addressed. Tel: +45 35325000; Fax: +45 35325001; Email: francesco.russo@cpr.ku.dk
†These authors contributed equally to the paper as third authors.

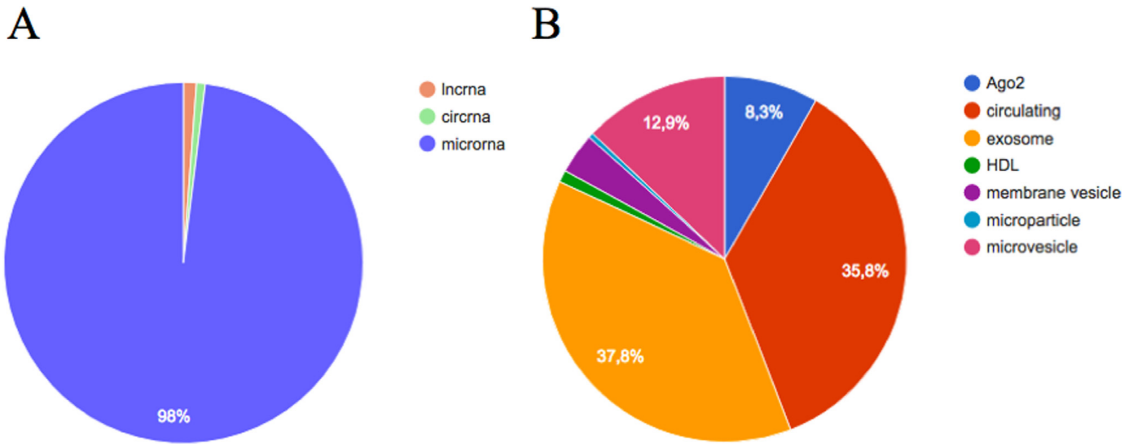



**Figure 1.** Descriptive statistics of the database. (**A**) Number of RNAs across RNA classes; (**B**) Number of RNAs across extracellular RNA forms.

**Table 1.** Comparison between the latest 2017 version and the previous version of miRandola

| | Previous version | miRandola 2017 |
|---|---|---|
| *Papers | 119 | 314 |
| Entries | 2276 | 3283 |
| microRNAs | 590 | 1002 |
| lncRNAs | 0 | 12 |
| CircRNAs | 0 | 8 |
| Extracellular RNA forms | 4 | 7 |
| Drugs | 6 | 25 |
| Organisms | 1 | 14 |
| Sample types | 23 | 47 |
| Visualization tool | No | Yes |
| External data | ExoCarta | ExoCarta and Vesiclepedia |
| **Text-mining-assisted curation | No | Yes |

*See the supplementary table.
**See the manuscript for more details.

bility in plasma, serum, saliva, urine and many other fluids (9–12). This stability is due to the formation of complexes of extracellular miRNAs in membrane-bound vesicles such as exosomes, which offers them protection from RNAses (13,14). Moreover, miRNAs can be found complexed with the Argonaute2 (Ago2) protein, part of the RNA-induced silencing complex (RISC) responsible of the RNA silencing mediated by miRNAs (15,16). miRNAs also complex with high density lipoprotein (HDL) (17), which provides the mechanism for a new pathway for intercellular communication. In fact, miRNAs transported by HDL can be delivered to recipient cells, where they can alter the expression of their targets (17).

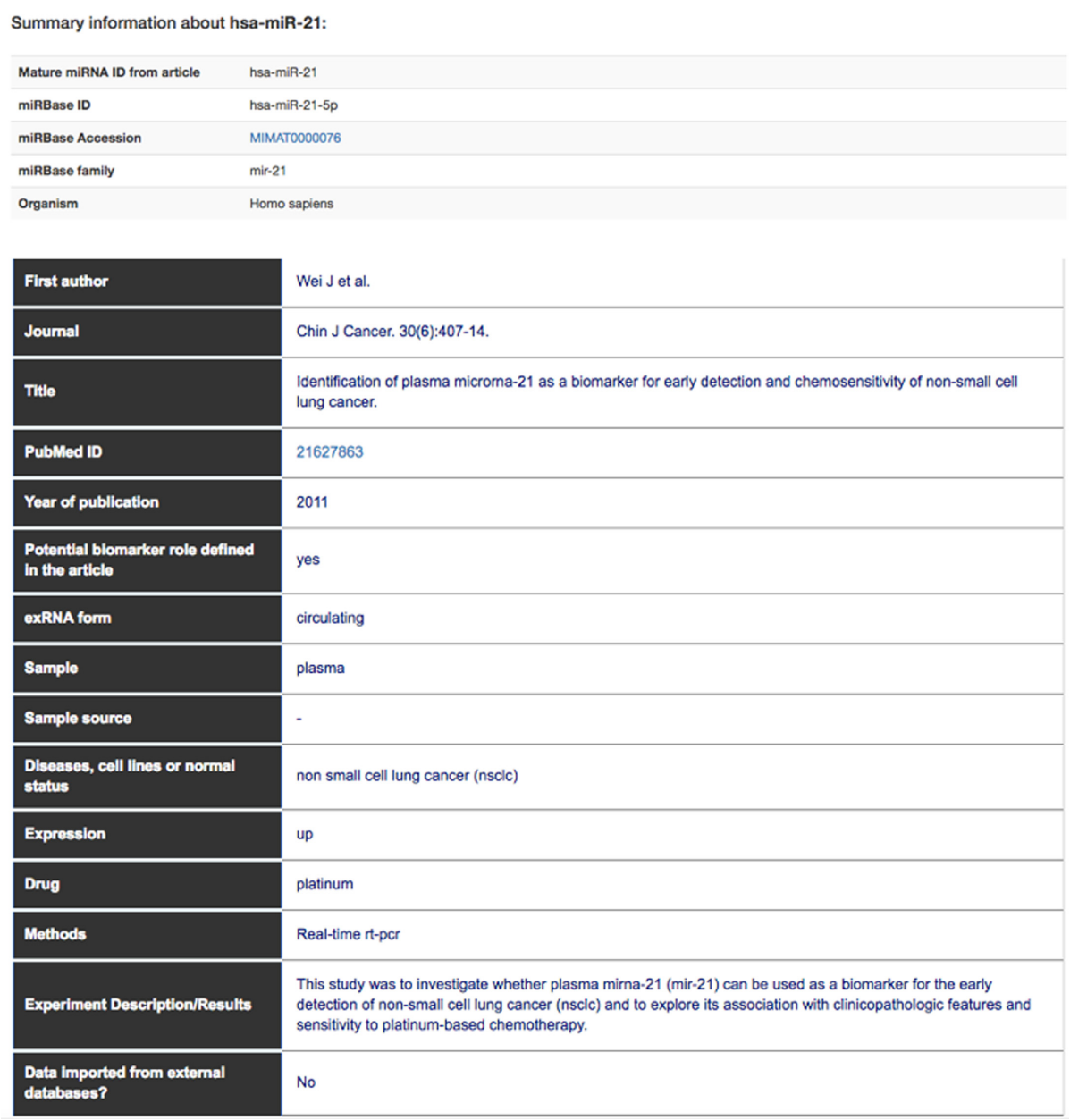
Given the rising interest in miRNAs and more generally other ncRNAs as potential biomarkers, we present an updated version of the miRandola database (11,12), with extensive addition of curated publications, new visualizations and additional functionality in the web interface.

## DATA COLLECTION AND CONTENT

The majority of the data were extracted from the literature in PubMed (https://www.ncbi.nlm.nih.gov/pubmed). This included information regarding: the RNA type, sample type, experimental procedures, associated diseases, extracellular RNA forms and other metadata regarding publications, including a summary of the results of the study. Some articles were collected from two publicly available resources, ExoCarta (18) and Vesiclepedia (19), which are two manually curated databases specialized in collecting information on extracellular vesicles.

The first version of the database made use of human biocurators that searched in PubMed using several keywords such as 'microRNA', 'circulating' and 'extracellular' and then manually extracted the relevant information. For this new version, we introduced text-mining-assisted curation to identify and prioritize papers for manual curation. The text mining approach will help our internal curators to increase the update frequency of the database to at least twice a year.

To identify terms of interest, the text-mining software uses dictionaries of human ncRNAs (20), diseases (21) and keywords that indicate extracellular RNA forms. We performed the text mining on more than 26 million entries in PubMed using the tagger software (22). We scored pairs of these terms by summing scores for all co-occurrences of the terms in the same sentence, paragraph and abstract with decreasing weights. We then normalized these scores, and we calculated the geometric mean of RNA-disease and RNA-circulating scores, which we took as the final score for the combined association between circulating RNA and disease. Scientific articles that contain all three types of terms



Summary information about **hsa-miR-21:**

| | |
|---|---|
| **Mature miRNA ID from article** | hsa-miR-21 |
| **miRBase ID** | hsa-miR-21-5p |
| **miRBase Accession** | MIMAT0000076 |
| **miRBase family** | mir-21 |
| **Organism** | Homo sapiens |

| | |
|---|---|
| **First author** | Wei J et al. |
| **Journal** | Chin J Cancer. 30(6):407-14. |
| **Title** | Identification of plasma microrna-21 as a biomarker for early detection and chemosensitivity of non-small cell lung cancer. |
| **PubMed ID** | 21627863 |
| **Year of publication** | 2011 |
| **Potential biomarker role defined in the article** | yes |
| **exRNA form** | circulating |
| **Sample** | plasma |
| **Sample source** | - |
| **Diseases, cell lines or normal status** | non small cell lung cancer (nsclc) |
| **Expression** | up |
| **Drug** | platinum |
| **Methods** | Real-time rt-pcr |
| **Experiment Description/Results** | This study was to investigate whether plasma mirna-21 (mir-21) can be used as a biomarker for the early detection of non-small cell lung cancer (nsclc) and to explore its association with clinicopathologic features and sensitivity to platinum-based chemotherapy. |
| **Data imported from external databases?** | No |

**Figure 2.** Results table containing the core information of the database. In this example, we report one result for *hsa-miR-21*.

were given the same score as the triple (RNA-extracellular form-disease). Scores were then used to rank articles and facilitate the manual curation.

Altogether, the collected data consisted of 314 articles (see Supplementary Table and website), a notably higher number of papers compared with the first version of the database ($n = 89$) (12) and the previous short update ($n = 119$) (11). For details on the database content see Table 1.

The database aims to present a comprehensive list of all known extracellular ncRNAs. Still, the majority of studies included in this version of miRandola focus on extracellular miRNAs, since they are the most investigated type of ncRNA. Although, in this new version, we started to collect information on two new RNA classes, namely long ncRNAs (lncRNAs) of more than 200 nt, and circular RNAs (circRNAs), which are transcripts that form a continuous loop. These RNA classes represent a small portion of the entries in the database (counts are reported in Table 1 and Figure 1A) and future updates will introduce additional information. Figure 1B shows the variety of extracellular forms that the ncRNAs are found complexed with, including Ago2, exosomes and HDL. More than 35% of RNAs have been annotated only as 'circulating', indicating that authors did not specify whether the RNA was complexed with known extracellular forms (Figure 1B).

Recently, some well-known extracellular lncRNAs have been used as potential non-invasive biomarkers. Probably the most famous example is *PCA3* (also known as *DD3*), which is highly overexpressed in most types of prostate cancer cells and detectable in urine (23). This new non-invasive

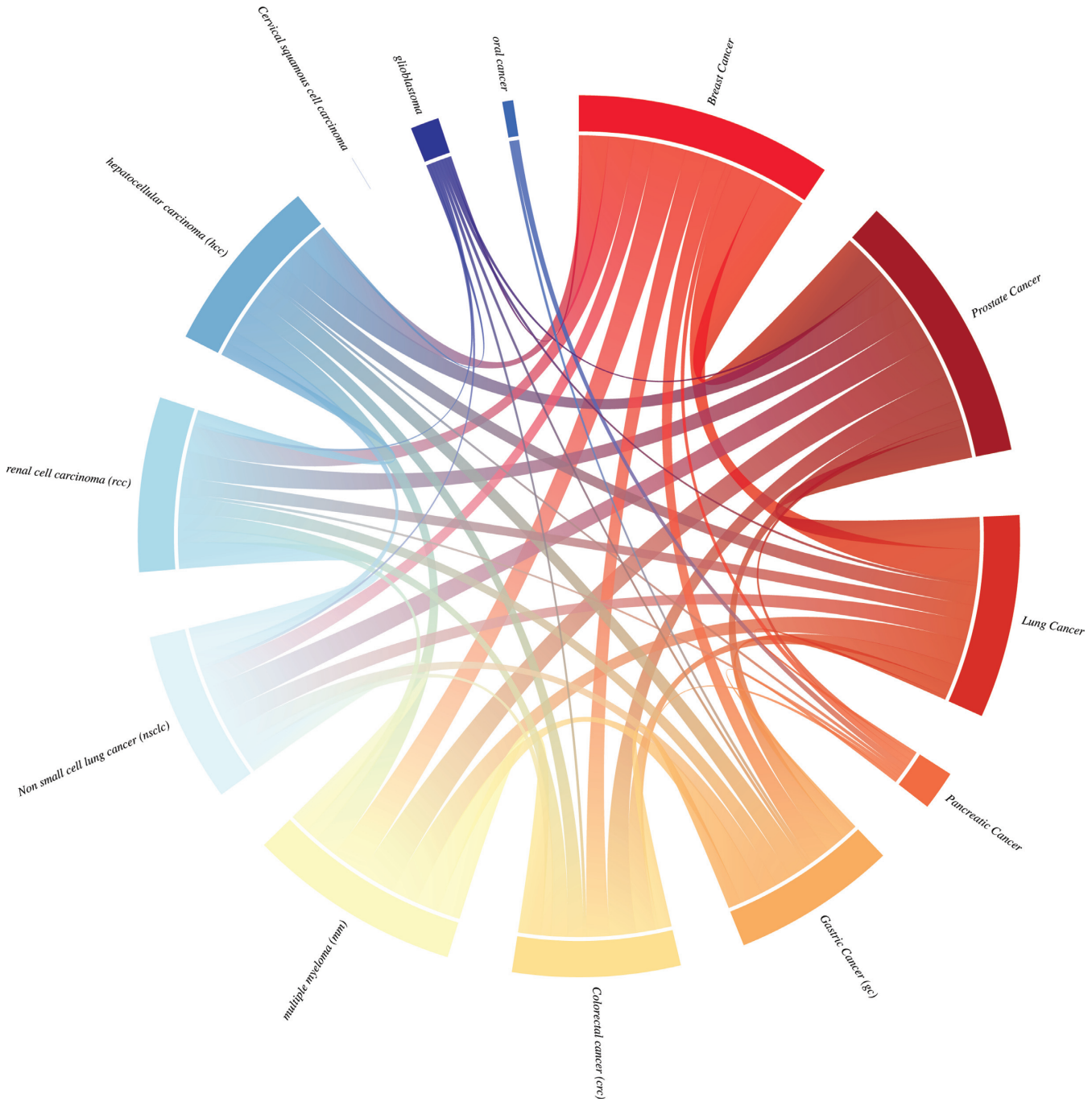



**Figure 3.** Circos plot of the most representative tumor types in the database. The plot shows how many RNAs are shared by different tumors.

biomarker shows great potential to improve patient care by reducing the number of biopsies (23).

The function of circRNAs is largely unknown, but some studies have shown that they are able to act as a natural 'sponge', by binding and down-regulating miRNAs (24,25). Since circRNAs are stable molecules (26), they have been proposed as novel non-invasive biomarker candidates (26).

## DATABASE DEVELOPMENT AND WEB INTERFACE

Data were collected and are maintained in a MySQL database running on an Apache server. The redesigned web interface was implemented using PHP and JavaScript (via the libraries AngularJS and D3.js). Furthermore, Bootstrap is used as front-end framework for faster and easier web development, allowing compatibility with web browsers. In this new version of the database we implemented new functionalities to explore and visualize data making the website more dynamic.

Starting from the home page of the database (see Supplementary Figure S1A), users can quickly search for the name of the RNA of interest by typing it into a search bar. We have implemented an autocomplete function in order to facilitate the search (see Supplementary Figure S1B). After clicking on the RNA of interest, users will have an overview of the extracellular forms in which the RNA has been found, and are able to click on the specific term to browse the results.

The user can browse by the following data types (see Supplementary Figure S2A): 'miRNAs', 'lncRNAs', 'circRNAs', 'Diseases', 'exRNA forms', 'Samples', 'Drugs' and 'Organisms'. For instance, after clicking on 'miRNAs', a summary table will be shown (see Supplementary Figure S2B) with miRNA identifiers as reported in literature, but also with the official miRNA identifiers annotated in the last version of the miRNA registry miRBase (27).

Each table can be filtered on a term of interest (see Supplementary Figure S2B), and can also be sorted. After clicking on a specific RNA of interest (see Supplementary Figure S2C and B), a results table is displayed (Figure 2).

The results table contains annotations specific to the selected RNA and other details such as publication identifier, reporting title, publication year, first author and journal. We show associations to diseases, sample types, the extracellular RNA type, the RNA expression level and the drug used in the experiment. We also report the methods used to verify the expression or other relevant techniques, and a short description of the results. The results table contains a field called 'Potential biomarker role defined in the article', indicating whether the selected RNA has a potential role as a biomarker, as stated in the published article.

Users can also use the 'Search' section (see Supplementary Figure S3) to search for pairs of terms such as '*hsa-miR-21*' and 'non-small cell lung cancer'.

All the data in miRandola are available in the 'Download' section of the database.

## VISUALIZATION

In this new version of miRandola, we introduce a visualization to show RNA-disease co-occurrences extracted from literature, and a circos plot (Figure 3) that shows how many RNAs are shared between the most representative tumor types in our database. This plot reveals that most tumors share several extracellular RNAs, with the exception of 'Cervical squamous cell carcinoma' for which we have no evidence of RNAs that are shared with any of the other tumor types. This common signature can be used to help identify common non-invasive biomarkers for many types of cancer.

## FUTURE PERSPECTIVE

When the field of extracellular RNAs was still new, miRandola was started as a small project and it soon became a successful database due to the work of a few biocurators and developers. In this new version, the involvement of additional biocurators and the introduction of assisted curation using text mining both contributed to the collection of many more curated articles, and on an ongoing basis, improves our ability to update the database regularly. We emphasize manual curation as an indispensable step, and define it as the fingerprint of miRandola. For this reason, we will evaluate the participation of the scientific community in the curation process in future updates. We aim to develop a new online tool to achieve this goal, giving users the possibility to curate articles that have been identified and prioritized by text mining. After this step, our internal curators will verify the information introduced by the scientific community.

The final goal of miRandola is to be a reference database for all non-invasive biomarkers, and future updates will consider other data such as extracellular DNA, giving a comprehensive panel of disease-specific biomarkers.

## DATA AVAILABILITY

miRandola is available at http://mirandola.iit.cnr.it/.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We would like to thank the editor and the anonymous reviewers for their constructive suggestions and comments. Following the suggestions, we improved the manuscript and the database.

## FUNDING

Novo Nordisk Foundation [NNF14CC0001 to F.R., H.V.C., A.S., K.B., L.J.J., S.B.]. Funding for open access charge: Novo Nordisk Foundation [NNF14CC0001].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.
2. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
3. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
4. Calin,G.A., Dumitru,C.D., Shimizu,M., Bichi,R., Zupo,S., Noch,E., Aldler,H., Rattan,S., Keating,M., Rai,K. *et al.* (2002) Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 15524–15529.
5. Calin,G.A., Ferracin,M., Cimmino,A., Di Leva,G., Shimizu,M., Wojcik,S.E., Iorio,M.V., Visone,R., Sever,N.I., Fabbri,M. *et al.* (2005) A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N. Engl. J. Med.*, **353**, 1793–1801.
6. Calin,G.A., Sevignani,C., Dumitru,C.D., Hyslop,T., Noch,E., Yendamuri,S., Shimizu,M., Rattan,S., Bullrich,F., Negrini,M. *et al.* (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 2999–3004.
7. Lagana,A., Russo,F., Sismeiro,C., Giugno,R., Pulvirenti,A. and Ferro,A. (2010) Variability in the incidence of miRNAs and genes in fragile sites and the role of repeats and CpG islands in the distribution of genetic material. *PLoS One*, **5**, e11166.
8. Rupaimoole,R. and Slack,F.J. (2017) MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.*, **16**, 203–222.

9. Mitchell,P.S., Parkin,R.K., Kroh,E.M., Fritz,B.R., Wyman,S.K., Pogosova-Agadjanyan,E.L., Peterson,A., Noteboom,J., O'Briant,K.C., Allen,A. *et al.* (2008) Circulating microRNAs as stable blood-based markers for cancer detection. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 10513–10518.
10. Weber,J.A., Baxter,D.H., Zhang,S., Huang,D.Y., Huang,K.H., Lee,M.J., Galas,D.J. and Wang,K. (2010) The microRNA spectrum in 12 body fluids. *Clin. Chem.*, **56**, 1733–1741.
11. Russo,F., Di Bella,S., Bonnici,V., Lagana,A., Rainaldi,G., Pellegrini,M., Pulvirenti,A., Giugno,R. and Ferro,A. (2014) A knowledge base for the discovery of function, diagnostic potential and drug effects on cellular and extracellular miRNAs. *BMC Genomics*, **15**(Suppl. 3), S4.
12. Russo,F., Di Bella,S., Nigita,G., Macca,V., Lagana,A., Giugno,R., Pulvirenti,A. and Ferro,A. (2012) miRandola: extracellular circulating microRNAs database. *PLoS One*, **7**, e47786.
13. Zhang,J., Li,S., Li,L., Li,M., Guo,C., Yao,J. and Mi,S. (2015) Exosome and exosomal microRNA: trafficking, sorting, and function. *Genomics Proteomics Bioinformatics*, **13**, 17–24.
14. Huang,X., Yuan,T., Tschannen,M., Sun,Z., Jacob,H., Du,M., Liang,M., Dittmar,R.L., Liu,Y., Liang,M. *et al.* (2013) Characterization of human plasma-derived exosomal RNAs by deep sequencing. *BMC Genomics*, **14**, 319.
15. Arroyo,J.D., Chevillet,J.R., Kroh,E.M., Ruf,I.K., Pritchard,C.C., Gibson,D.F., Mitchell,P.S., Bennett,C.F., Pogosova-Agadjanyan,E.L., Stirewalt,D.L. *et al.* (2011) Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 5003–5008.
16. Turchinovich,A., Weiz,L., Langheinz,A. and Burwinkel,B. (2011) Characterization of extracellular circulating microRNA. *Nucleic Acids Res.*, **39**, 7223–7233.
17. Vickers,K.C., Palmisano,B.T., Shoucri,B.M., Shamburek,R.D. and Remaley,A.T. (2011) MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. *Nat. Cell Biol.*, **13**, 423–433.
18. Keerthikumar,S., Chisanga,D., Ariyaratne,D., Al Saffar,H., Anand,S., Zhao,K., Samuel,M., Pathan,M., Jois,M., Chilamkurti,N. *et al.* (2016) ExoCarta: a web-based compendium of exosomal cargo. *J. Mol. Biol.*, **428**, 688–692.
19. Kalra,H., Simpson,R.J., Ji,H., Aikawa,E., Altevogt,P., Askenase,P., Bond,V.C., Borras,F.E., Breakefield,X., Budnik,V. *et al.* (2012) Vesiclepedia: a compendium for extracellular vesicles with continuous community annotation. *PLoS Biol.*, **10**, e1001450.
20. Junge,A., Refsgaard,J.C., Garde,C., Pan,X., Santos,A., Alkan,F., Anthon,C., von Mering,C., Workman,C.T., Jensen,L.J. *et al.* (2017) RAIN: RNA-protein Association and Interaction Networks. *Database (Oxford)*, **2017**, doi:10.1093/database/baw167.
21. Pletscher-Frankild,S., Palleja,A., Tsafou,K., Binder,J.X. and Jensen,L.J. (2015) DISEASES: text mining and data integration of disease-gene associations. *Methods*, **74**, 83–89.
22. Pafilis,E., Frankild,S.P., Fanini,L., Faulwetter,S., Pavloudi,C., Vasileiadou,A., Arvanitidis,C. and Jensen,L.J. (2013) The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS One*, **8**, e65390.
23. Hessels,D., Klein Gunnewiek,J.M., van Oort,I., Karthaus,H.F., van Leenders,G.J., van Balken,B., Kiemeney,L.A., Witjes,J.A. and Schalken,J.A. (2003) DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. *Eur Urol.*, **44**, 8–15.
24. Memczak,S., Jens,M., Elefsinioti,A., Torti,F., Krueger,J., Rybak,A., Maier,L., Mackowiak,S.D., Gregersen,L.H., Munschauer,M. *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, **495**, 333–338.
25. Hansen,T.B., Jensen,T.I., Clausen,B.H., Bramsen,J.B., Finsen,B., Damgaard,C.K. and Kjems,J. (2013) Natural RNA circles function as efficient microRNA sponges. *Nature*, **495**, 384–388.
26. Memczak,S., Papavasileiou,P., Peters,O. and Rajewsky,N. (2015) Identification and Characterization of Circular RNAs As a New Class of Putative Biomarkers in Human Blood. *PLoS One*, **10**, e0141214.
27. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.