dbDEMC 3.0: Functional Exploration of Differentially Expressed miRNAs in Cancers of Human and Model Organisms

3

4 Feng Xu^{1#}, Yifan Wang^{2#}, Yunchao Ling², Chenfen Zhou², Haizhou Wang¹, Andrew
5 E. Teschendorff³, Yi Zhao⁴, Haitao Zhao⁵, Yungang He^{6*}, Guoqing Zhang^{2*}, Zhen
6 Yang^{1*}

7

8 ¹ Center for Medical Research and Innovation of Pudong Hospital, Fudan University,
9 and the Shanghai Key Laboratory of Medical Epigenetics, the International
10 Co-laboratory of Medical Epigenetics and Metabolism, Ministry of Science and
11 Technology, Institutes of Biomedical Sciences, Fudan University, Shanghai 200032,
12 China

13 ² Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology,
14 Shanghai Institute of Nutrition and Health, University of Chinese Academy of
15 Sciences, Chinese Academy of Sciences, Shanghai 200031, China

16 ³ CAS Key Lab of Computational Biology, Shanghai Institute for Nutrition and
17 Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences,
18 Shanghai 200031, China

19 ⁴ Key Laboratory of Intelligent Information Processing, Advanced Computer
20 Research Center, Institute of Computing Technology, Chinese Academy of Sciences,
21 Beijing 100190, China

22 ⁵ Department of Liver Surgery, Peking Union Medical College Hospital, Chinese
23 Academy of Medical Sciences and Peking Union Medical College, Beijing, 100730,
24 China

25 ⁶ Shanghai Fifth People's Hospital, and the Shanghai Key Laboratory of Medical
26 Epigenetics, International Co-laboratory of Medical Epigenetics and Metabolism,
27 Ministry of Science and Technology, Institutes of Biomedical Sciences, Fudan
28 University, Shanghai 200032, China

29

30 # Equal contribution.

31 * Corresponding authors.

32 E-mail: zhenyang@fudan.edu.cn (Yang Z), gqzhang@picb.ac.cn (Zhang GQ),

33 heyungang@fudan.edu.cn (He YG)

34

35 **Running title:** *Xu F et al. / a database for cancer miRNAs*

36

37 Total word counts (from “Introduction” to “Conclusions” or “Materials and
38 methods”): 3084

39 Total figures: 4

40 Total tables: 1

41 Total supplementary figures: 2

42 Total supplementary tables: 5

43 Total number of characters in the article title: 109

44 Total number of characters in the running title: 42

45 Total number of words of keywords: 5

46 Total number of words in Abstract: 191

47

48 **Abstract**

49 microRNAs (miRNAs) are important regulators in gene expression. The deregulation
50 of miRNA expression is widely reported in the transformation from physiological to
51 pathological state of cells. A large amount of differentially expressed miRNAs
52 (DEMs) have been identified in various human cancers by using high-throughput
53 technologies, such as microarray and miRNA-seq. Through mining of published
54 researches with high-throughput experiment information, the database of differentially
55 expressed miRNAs in human cancers (dbDEMC) was constructed with the aim of
56 providing a systematic resource for the storage and query of the DEMs. Here we
57 report an update of the dbDEMC to version 3.0, containing two-fold more data entries
58 than the previous version, now including also data from mouse and rat. The dbDEMC
59 3.0 contains 3,268 unique DEMs in 40 different cancer types. The current datasets for
60 differential expression analysis have expanded to 9 generalized categories. Moreover,
61 the current release integrates functional annotations of DEMs obtained from
62 experimentally validated targets. The annotations can greatly benefit integrative
63 analysis of DEMs. In summary, dbDEMC 3.0 provides a valuable resource for
64 characterizing molecular functions and regulatory mechanisms of DEMs in human
65 cancers. The dbDEMC 3.0 is freely accessible at <https://www.biosino.org/dbDEMC>.

66
67 **KEYWORDS:** miRNA; Cancer; Differential expression; Model Organisms;
68 Database

69
70

71 **Introduction**

72 Since the first discovery of microRNAs (miRNAs) at the beginning of this century,
 73 this class of small non-coding RNA has received extensive attention [1]. As an
 74 important gene expression regulator acting at the post-transcriptional level, studies
 75 have disclosed the critical role of miRNAs in targeting mRNAs for the degradation or
 76 translational repression [2]. Till now, a total of 2,654 miRNAs have been identified in
 77 the human genome according to the latest version of miRBase database [3].
 78 Significant researches on miRNAs have dramatically expanded our understanding
 79 about gene regulatory network and its roles in physiological and pathological
 80 conditions, such as in broad spectrums of biological processes including cell cycle,
 81 cell proliferation, differentiation, apoptosis and cellular signaling [4, 5]. Owing to the
 82 biological significance of miRNAs, alterations of the miRNAs are linked to the
 83 development of many diseases including the cancer [6]. Differentially expressed
 84 miRNAs (DEMs) are widely reported to hold great value in the diagnosis or prognosis
 85 as well as treatment targeting for cancer research [7]. The potential use of circulating
 86 miRNAs in serum, plasma and other body fluids as non-invasive cancer biomarkers
 87 have also been thoroughly investigated [8].

88 Given the important functions of miRNA in cancer development, several on-line
 89 resources have been built for warehousing information of cancer-related miRNAs,
 90 such as the HMDD [9], miRCancer [10], and OncomiRDB [11]. With the
 91 development of high-throughput techniques such as microarray and miRNA-seq, large
 92 amount of cancer DEMs were identified from miRNA profiling data each year.
 93 However, these valuable data were scattered in the vast literatures and it is of great
 94 necessity to catalogue them in a favorable way, thus to provide integrative tools for
 95 the effective utilization and systematic investigation. With this aim, we developed the
 96 initial database of differentially expressed miRNAs in human cancers (dbDEMC) in
 97 2010 [12] and further updated in 2017 [13]. To our knowledge, dbDEMC is the only
 98 working repository currently available for storing DEMs from de-novo analysis of
 99 high-throughput profiling data in human cancers, characteristic with miRNANome

100 data in various types of cancer. It greatly facilitates the efforts to excavate cancer
101 associated miRNAs and investigate their roles in pathological processes of cancer.
102 While the database could have been much more useful had there been more
103 high-quality data included.

104 In recent years, cancer quantitative miRNA profiling data has been increasing at
105 an unprecedented rate, and given the success of dbDEMC 2.0, this motivates an
106 update of this database. Here we introduce dbDEMC 3.0, a significantly expanded
107 version of this database. This update incorporates a substantial amount of new data.
108 Besides the human data, we have also incorporated the miRNA expression profiling
109 data of mouse and rat. A total of 403 datasets of high-throughput miRNA expression
110 encompassing 40 cancer-types, with the results of 807 differential expression
111 analyses, have been included. The present update is nearly doubling the data amount
112 over the previous version. In addition to the expanded data volume, the content of the
113 database has also been enriched. The new database firstly recorded the experimentally
114 validated DEMs targets and also their enrichment analysis on Gene Ontology (GO)
115 and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. In this way, we
116 provided the functional annotation of DEMs for various cancers. Also, the web
117 interface of the database was refined for a better visualization of the aforementioned
118 data. Taken together, the dbDEMC 3.0 is a comprehensive resource to systematically
119 characterize the function of DEMs in human cancers as well as other model
120 organisms.

121

122 **Data collection and processing**

123 **Data collection**

124 To compile the datasets, we used the keywords ‘cancer’, ‘tumor’, ‘carcinoma’ and
125 ‘neoplasm’, in combination with ‘microRNA’ or ‘miRNA’ to conduct an exhaustive
126 search for the microarray based miRNA expression profiles in Gene Expression
127 Omnibus (GEO) [14] and ArrayExpress [15], and further for miRNA-seq based
128 miRNA expression profiles in Sequence Read Archive (SRA) [16]. While the miRNA

129 profiles of mouse and rat were rapidly accumulating, we also incorporated the miRNA
130 data of the two model organisms in the current update. In addition, we also appended
131 the new miRNA profiling data from TCGA that was released since the last update of
132 dbDEMC 2.0. All the involved data were published as of June 2021. The data records
133 were manually reviewed and evaluated rigorously to guarantee that only high-quality
134 data sets were included. To ensure analysis reliability, we required at least three
135 biological replicates of samples each condition (for both case and control) as usual.

136 **Data processing**

137 For miRNA profiling data sets based on microarray, we used the same protocol as that
138 of dbDEMC 2.0 to identify the DEMs [13]. Briefly, the expression values were
139 logarithmically transformed (base 2) and quantile normalized. Then the limma (Linear
140 Models for Microarray Data) package was applied to select miRNAs whose mean
141 expression level is significantly different between case and control samples with FDR
142 value < 0.05 .

143 For miRNA-seq based profiling data obtained from SRA database, we
144 downloaded the SRA files of raw sequence reads and converted them into FASTQ
145 format using the fastq-dump of SRA Toolkit. Here we included only the data
146 produced on Illumina systems (Genome Analyzer I, II, IIx, HiSeq 1000, HiSeq 2000,
147 HiSeq 2500, HiSeq 4000, NextSeq, MiSeq). The involving miRNA-seq data was
148 analyzed by using QuickMIRSeq toolkit [17]. This toolkit utilizes the Cutadapt to
149 remove sequence adapters and performs quality control [18]. We collected detailed
150 information of DNA adapters of different miRNA-seq libraries from public resources
151 to guarantee that the adapters can be properly trimmed from the raw reads (Table S1)
152 [19]. The clean reads were then aligned to the reference genome by using Bowtie
153 [20], and miRDeep2 was used to obtain count tables of aligned reads for miRNA
154 quantification [21]. The reads count table was further normalized by using
155 limma-voom [22], and DEMs were then identified. For the datasets obtained from
156 TCGA, we directly used the prepared reads count data for further analysis [23].

Experimental validation of DEMs in low-throughput methods (such as real-time PCR and northern blot, etc.) were manually collected from the original papers. These types of information were carefully formatted and integrated into our update.

Functional annotation

For each obtained DEMs set, we collected the experimentally validated targets by using multiMiR [24], which integrate miRNA target data from TarBase [25] and miRTarBase [26]. Then we performed the enrichment analysis of the DEMs targets on GO terms and KEGG pathways by using clusterProfiler package to facilitate the study of context-dependent miRNA functional mechanisms [27]. Enriched GO terms and KEGG pathways were selected where adjusted P value < 0.05. The data collection and curation procedure for dbDEMC 3.0 is shown in **Figure 1**.

Database construction

All the data in dbDEMC 3.0 was managed by using MongoDB. The dynamic web interface was developed using JSP and JavaScript. Data visualization was achieved through the tools of vue, jquery, and Echarts, Elasticsearch was used for search engine. The database was developed by Spring Boot framework. Apache Tomcat web server was used for the http server. All the information in dbDEMC 3.0 is freely available to the public domain through <https://www.biosino.org/dbDEMC>.

miRNA cluster annotation

A miRNA cluster is defined as set of miRNAs which located adjacent genomic regions in the same or opposite orientation and not separated by other transcriptional units. miRNAs within a cluster are thought to be regulated by common factors and involved in same signaling pathways. According to Kabekkodu SP et al., among 1881 precursor miRNAs of human origin annotated in miRBase, 468 of which can be attributed to 153 clusters [28]. Here we obtained these data about miRNA clusters and annotated mature miRNAs by using annotation file from miRBase. Finally, a total of 688 (22.8%) mature miRNAs from 143 clusters were annotated in the human genome. Analysis of homogeneous dysregulation pattern of miRNA clusters in cancer

For systematic study of co-dysregulation pattern of miRNA clusters in human cancers, we considered all miRNAs associated with specific cancer. miRNAs not

187 belonging to any cluster and clusters of which at least half the members are not
 188 associated with any cancer were discarded. To avoid potential bias introduced by
 189 different expression platforms, here we only use the results obtained from TCGA and
 190 checked for experimental design of Cancer vs. Normal for 19 kinds of epithelial
 191 cancers. We finally obtained 106 unique clusters for these cancer types. A cluster was
 192 designated to be homogeneous if at least half of its members show the same direction
 193 of expression pattern (either up- or down-regulated). For each cluster, we computed
 194 the homogeneous fraction as that of co-dysregulation throughout all cancer types
 195 analyzed. A significant P-value for this fraction was calculated as follow: for each
 196 cancer type, the expression of all its associated miRNAs was distributed randomly
 197 within these miRNAs for 10,000 times, keeping the distribution of up- and
 198 down-regulated miRNAs constant for each step. The homogeneous fraction over all
 199 cancers was computed, which yields the P-value as the number of sampled
 200 homogeneous fractions exceeding the original homogeneous fractions divided by
 201 10,000.

202 In order to check whether clustered miRNAs are more enriched in cancer
 203 development compared to single miRNA, we calculated an enrichment score of
 204 log-odds (LOD) score for each cancer type:

$$205 \quad \text{LOD} = \log_2\left(\left(\frac{X_c}{X_c + Y_c}\right) / \left(\frac{X_{all}}{X_{all} + Y_{all}}\right)\right)$$

206 Where X_c and Y_c denote the number of cluster miRNAs and non-cluster
 207 miRNAs for each cancer type, and X_{all} denotes the number of the cluster miRNAs
 208 and denotes the number of the miRNAs not contained in any cluster. Here we take
 209 into account all known human miRNAs annotated in the human genome, thus
 210 designate as the 688 clustered miRNAs and 900 non-clustered miRNAs. In this case, a
 211 positive LOD score indicates enrichment for cluster miRNAs compared to non-cluster
 212 miRNAs in a specific cancer.

213

214 **Implementation and results**

215 Database content

216 In the current release of dbDEMC, the data of miRNA transcriptome of total 46,388
217 samples from 403 studies of human, mouse or rat were collected from public
218 resources (Table S2). These profiles are derived from 149 subtypes of 40 different
219 cancers and many different cell lines (Table S3). We then performed a systematic
220 analysis on each dataset, and yielded a total of 807 experiments for differential
221 expression analysis. dbDEMC 3.0 now hosts a total of 3,268 differentially expressed
222 miRNAs, among them, 2,584 are specific to human. A total of 160,799 miRNA
223 variations related to cancers were deposited in our database. The detailed information
224 about number of miRNAs, cancer types, datasets, and experiments for different
225 species was presented in **Table 1**.

226 **Figure 2A** depicts the number of DEMs for each type of human cancer. The
227 breast cancer presents a large number of DEMs that 1,833 up-regulated and 1,988
228 down-regulated DEMs had been identified. The number of DEMs from mouse and rat
229 can be found in Figure S1. Whereas Figure 2B demonstrates the number of DEMs
230 validated by low-throughput methods across major cancers, the brain cancer,
231 colorectal cancer and breast cancer are top ranked cancer types. Figure 2C shows the
232 percentages of experiments for top ranked cancers. The breast cancer accounted 15%
233 of the total experiments and ranked the first of the list. It was followed by colorectal
234 cancer and lung cancer. Whereas the 9 different comparisons, cancer samples vs.
235 normal controls constitutes about half of the total experiments, and followed by
236 comparisons of high-grade versus low-grade cancer samples (Figure 2D). Overall, the
237 sizes of analysis experiments and related literatures in dbDEMC 3.0 have a two-fold
238 increment by comparing with the previous version (Figure S2).

239 New features

240 In the dbDEMC 2.0, we assigned the different experimental designs to 7 different
241 classes: Cancer vs. Normal; High grade vs. Low grade; Metastasis vs. Primary cancer;
242 Subtype1 vs. Subtype2; Poor outcome vs. Good outcome; Blood sample of patients
243 vs. Blood sample of normal controls, and also Treatment vs. Non-treatment. In recent
244 years, many studies disclosed exosomes and microvesicles act as cell communication

agents, where miRNA is the most important molecular in exosomes and microvesicles for regulating cancer progression [29]. In addition, circulating miRNAs were also widely found in body fluids and represent a gold mine of noninvasive biomarkers in cancer [30]. In this update version, we thus added these two classes of experimental design: Exosome sample from patients vs. Exosome sample from control, and Body fluid from patients vs. Body fluid from control. Moreover, for each DEMs set, targets of miRNAs and enrichment information of the target genes in the KEGG pathways and GO terms were deposited in the dbDEMC 3.0, which makes it possible for inspecting functional mechanisms behind a set of miRNAs.

Newly designed web interface

The web interface of dbDEMC 3.0 has been significantly refined and improved, allowing better use of the data. The Search page permits users to perform a quick search and extract summarized information of a DEM list across cancer types. Users can also specify the cancer type, experimental design or platform to select the interested experiments (**Figure 3A–C**). After filtering the experimental results, users can select interested experiments. Detailed information of DEM related experiment, which includes description with the up-regulated and down-regulated miRNAs, can be accessed. In the functional chart section, heatmap of the DMEs expression, miRNA-target regulatory network for top ranked DEMs, as well as the bubble chart for miRNA targets enriched KEGG and GO terms were presented (Figure 3D). Using a single miRNA query, summary information of the interested miRNA can be retrieved, including the general description of the interested miRNA and differential expression summary heatmap which decided by number of experiments shows up expression or down expression. In addition, summary statistics tables for both high-throughput data analysis and low-throughput validation data were also displayed (Figure 3E).

Analysis tools

miRBase is the central reference database for miRNA annotation by assigning names and unique gene IDs for novel miRNAs. During its development, some miRNA definition and annotation may have been changed. This leads to the inconsistency of

the miRNA IDs from different data sets, which derived from different miRBase versions and making it difficult for comparing research results for integrative analysis. To solve this problem, we provide a “ID convertor” in our database, by which users could convert miRBase old version IDs to the latest version (v22.0) for the three species of human, mouse, and rat. In addition, other analyzing tools including BLAST and meta-profiling, which are used for sequence similarity search of unknown miRNAs and identify the confident cancer related miRNAs in pan-cancer wide, are also available in dbDEMC v3.0. For the meta-profiling study, the vote-counting approach was used to calculate the consistent score of differential expression for meta-analysis [31]. Common miRNAs identified in multiple cancer types with similar differential expression pattern may suggest they could have similar regulatory mechanisms and play important roles in cancer development.

miRNA clusters are significantly overrepresented in cancers

A large proportion of miRNAs localized as conserved clusters in the genome and present similar expression pattern across tissues. It is critical to understand whether miRNA clusters present similar differential expression across cancers and correlate with the similar pathobiology. Here we obtained human miRNA cluster annotation from public resources, which includes 22.8% (688/2588) of mature miRNAs appear as 143 clusters of at least two members within (Table S4). We systematically analyzed the homogeneity of expression patterns within miRNA clusters. We excluded those clusters having less than half of all miRNAs annotated from the results of TCGA, which leads to 106 remaining clusters. The clusters are denoted as exhibiting a homogeneous expression pattern if annotated miRNA members are either up- or down-regulated (see Data processing). In total, cancer-associated clusters revealed homogeneous expression patterns for 74% of all annotated cancer, which confirms the hypothesis of co-regulation pattern of miRNA clusters in cancer. For example, the cluster of miR-142-5p, miR-142-3p, miR-4736 presents a consistent differential expression pattern in 91% (11/12) of the cancers analyzed (Table S5). A null model by randomly linking miRNA expression patterns (permutation 10,000 times within each cancer) indicated 52 clusters (49%, P-value < 0.05) showed a

significantly higher homogeneity pattern in all 19 kinds of cancer compared to that expected by chance (Table S5). These clusters exhibit a homogeneous expression pattern in at least 78.5% of all these types of cancer.

To further investigate the association of miRNA clusters with different kinds of cancer, we estimated the enrichment of miRNA clusters in cancer-associated miRNAs by using a LOD score. We found enrichment for all 19 kinds of cancer (**Figure 4**). Within these 19 kinds of cancer, miRNAs located in clusters are, on average, 1.56 times (LOD = 0.65) enriched compared to random. In summary, our analyses show a significant enrichment of clustered miRNAs in cancers compare to the single miRNA members, which demonstrated that the different miRNAs within cluster act synergistically in cancer development.

Discussion

Over the last decade, a large amount of miRNA transcriptome profiles of various cancers have been generated. Many studies have performed miRNA transcriptome analysis to explore the underlying molecular mechanisms of miRNA genes in cancer development [32, 33]. This progress motivated a novel release of dbDEMC to keep track of the latest published data. Along with this, we curate these data and provide a platform to facilitate the study of miRNA-cancer associations. For dbDEMC 3.0, it not only contains more miRNA-cancer associations, we also extended our database to the species of mouse and rat, which benefits those studies characterizing the miRNA functional machinery in cancer using the model organisms. Beyond the rapid increase of data amount, our database now offers many new features and powerful tools for the downstream analysis of DEMs, such as the integrated target identification and functional enrichment analysis for miRNA-regulated biological processes.

One of the key questions of differential expression analysis of miRNAs is which cancer types are regulated by a particular miRNA (miRNA-centric view), or conversely, which miRNAs may involve in a given type of cancer (cancer-centric view). Our databases support both miRNA- and cancer-centric investigations, users

334 are able to search miRNAs to determine the spectrum of cancer types that it involved,
 335 or to find candidate miRNA list based on reference of the miRBase which link to
 336 individual type of cancer. Although studies have indicated false positive and negative
 337 records may exist in miRBase, researchers need to be cautious about resources based
 338 on references from miRBase [34, 35]. Overall, we expect that dbDEMC 3.0 could
 339 serve as a valuable resource with comprehensive data amount and data analysis tools
 340 to facilitate the study of DEMs in cancers. In the future, we will continue to make
 341 improvements to the functional characterizations as well as integrate more
 342 heterogeneous data for the flexible analysis of miRNA functions in various cancers.
 343 We believe that the development of dbDEMC database can help accelerate the
 344 integration between miRNANome and cancer studies.

345

346

347 **Data availability**

348 dbDEMC v3.0 is freely accessible at <https://www.biosino.org/dbDEMC/>.

349

350 **CRedit author statement**

351 **Feng Xu:** Data curation, Formal analysis. **Yifan Wang:** Methodology, Validation.

352 **Yunchao Ling:** Visualization. **Chenfen Zhou:** Data curation. **Haizhou Wang:** Data

353 curation. **Andrew E. Teschendorff:** Methodology, Software. **Yi Zhao:** Investigation,

354 Software. **Haitao Zhao:** Validation. **Yungang He:** Supervision, Investigation.

355 **Guoqing Zhang:** Methodology, Supervision. **Zhen Yang:** Conceptualization,

356 Supervision, Writing- Original draft. All authors read and approved the final

357 manuscript.

358

359 **Competing interest**

360 The authors declare that they have no competing interests.

361

362 **Acknowledgments**

363 This work is supported by National Natural Science Foundation of China [91959106,

364 31871255, 91731310, and 81827901]; Strategic Priority Research Program of the

365 Chinese Academy of Sciences [XDB38030100]; Shanghai Municipal Science and

366 Technology [2017SHZDZX01].

367

368 **ORCID**

369 0000-0001-5185-2164 (Feng Xu)

370 0000-0003-4256-2696 (Yifan Wang)

371 0000-0002-5438-0753 (Yunchao Ling)

372 0000-0002-0294-7308 (Chenfen Zhou)

373 0000-0002-3376-386X (Haizhou Wang)

374 0000-0001-7410-6527 (Andrew E. Teschendorff)

375 0000-0001-6046-8420 (Yi Zhao)
376 0000-0002-3444-8044 (Haitao Zhao)
377 0000-0002-2931-2871 (Yungang He)
378 0000-0001-8827-7546 (Guoqing Zhang)
379 0000-0002-5647-9976 (Zhen Yang)
380

References

- [1] Ambros V. microRNAs: tiny regulators with great potential. *Cell* 2001;107:823-6.
- [2] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;116:281-97.
- [3] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2019;47:D155-D62.
- [4] Shivdasani RA. MicroRNAs: regulators of gene expression and cell differentiation. *Blood* 2006;108:3646-53.
- [5] Hwang HW, Mendell JT. MicroRNAs in cell proliferation, cell death, and tumorigenesis. *Br J Cancer* 2007;96 Suppl:R40-4.
- [6] Esquela-Kerscher A, Slack FJ. Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer* 2006;6:259-69.
- [7] Wang J, Chen J, Sen S. MicroRNA as Biomarkers and Diagnostics. *J Cell Physiol* 2016;231:25-30.
- [8] Wang H, Peng R, Wang J, Qin Z, Xue L. Circulating microRNAs as potential cancer biomarkers: the advantage and disadvantage. *Clin Epigenetics* 2018;10:59.
- [9] Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, et al. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res* 2019;47:D1013-D7.
- [10] Xie B, Ding Q, Han H, Wu D. miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics* 2013;29:638-44.
- [11] Wang D, Gu J, Wang T, Ding Z. OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs. *Bioinformatics* 2014;30:2237-8.
- [12] Yang Z, Ren F, Liu C, He S, Sun G, Gao Q, et al. dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics* 2010;11 Suppl 4:S5.

408 [13] Yang Z, Wu L, Wang A, Tang W, Zhao Y, Zhao H, et al. dbDEMC 2.0: updated
409 database of differentially expressed miRNAs in human cancers. *Nucleic Acids Res*
410 2017;45:D812-D8.

411 [14] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al.
412 NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*
413 2013;41:D991-5.

414 [15] Athar A, Fullgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress
415 update - from bulk to single-cell expression data. *Nucleic Acids Res*
416 2019;47:D711-D5.

417 [16] Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence
418 Database C. The Sequence Read Archive: explosive growth of sequencing data.
419 *Nucleic Acids Res* 2012;40:D54-6.

420 [17] Zhao S, Gordon W, Du S, Zhang C, He W, Xi L, et al. QuickMIRSeq: a pipeline
421 for quick and accurate quantification of both known miRNAs and isomiRs by jointly
422 processing multiple samples from microRNA sequencing. *BMC Bioinformatics*
423 2017;18:180.

424 [18] Chen C, Khaleel SS, Huang H, Wu CH. Software for pre-processing Illumina
425 next-generation sequencing short read sequences. *Source Code Biol Med* 2014;9:8.

426 [19] Zhong X, Heinicke F, Lie BA, Rayner S. Accurate Adapter Information Is
427 Crucial for Reproducibility and Reusability in Small RNA Seq Studies. *Noncoding*
428 *RNA* 2019;5.

429 [20] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient
430 alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10:R25.

431 [21] Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2
432 accurately identifies known and hundreds of novel microRNA genes in seven animal
433 clades. *Nucleic Acids Res* 2012;40:37-52.

434 [22] Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear
435 model analysis tools for RNA-seq read counts. *Genome Biol* 2014;15:R29.

436 [23] Chu A, Robertson G, Brooks D, Mungall AJ, Birol I, Coope R, et al. Large-scale
437 profiling of microRNAs for The Cancer Genome Atlas. *Nucleic Acids Res*
438 2016;44:e3.

439 [24] Ru Y, Kechris KJ, Tabakoff B, Hoffman P, Radcliffe RA, Bowler R, et al. The
440 multiMiR R package and database: integration of microRNA-target interactions along
441 with their disease and drug associations. *Nucleic Acids Res* 2014;42:e133.

442 [25] Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S,
443 Kanellos I, et al. DIANA-TarBase v8: a decade-long collection of experimentally
444 supported miRNA-gene interactions. *Nucleic Acids Res* 2018;46:D239-D45.

445 [26] Huang HY, Lin YC, Li J, Huang KY, Shrestha S, Hong HC, et al. miRTarBase
446 2020: updates to the experimentally validated microRNA-target interaction database.
447 *Nucleic Acids Res* 2020;48:D148-D54.

448 [27] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing
449 biological themes among gene clusters. *OMICS* 2012;16:284-7.

450 [28] Kabekkodu SP, Shukla V, Varghese VK, J DS, Chakrabarty S, Satyamoorthy K.
451 Clustered miRNAs and their role in biological functions and diseases. *Biol Rev Camb*
452 *Philos Soc* 2018;93:1955-86.

453 [29] Sun Z, Shi K, Yang S, Liu J, Zhou Q, Wang G, et al. Effect of exosomal miRNA
454 on cancer biology and clinical applications. *Mol Cancer* 2018;17:147.

455 [30] Cortez MA, Bueso-Ramos C, Ferdin J, Lopez-Berestein G, Sood AK, Calin GA.
456 MicroRNAs in body fluids--the mix of hormones and biomarkers. *Nat Rev Clin*
457 *Oncol* 2011;8:467-77.

458 [31] Guan P, Yin Z, Li X, Wu W, Zhou B. Meta-analysis of human lung cancer
459 microRNA expression profiling studies comparing cancer tissues with normal tissues.
460 *J Exp Clin Cancer Res* 2012;31:54.

461 [32] Ueda T, Volinia S, Okumura H, Shimizu M, Taccioli C, Rossi S, et al. Relation
462 between microRNA expression and progression and prognosis of gastric cancer: a
463 microRNA expression analysis. *Lancet Oncol* 2010;11:136-46.

[33] Iorio MV, Croce CM. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Mol Med* 2012;4:143-59.

[34] Fromm B, Keller A, Yang X, Friedlander MR, Peterson KJ, Griffiths-Jones S. Quo vadis microRNAs? *Trends Genet* 2020;36:461-3.

[35] Alles J, Fehlmann T, Fischer U, Backes C, Galata V, Minet M, et al. An estimate of the total number of true human miRNAs. *Nucleic Acids Res* 2019;47:3353-64.

471

472

473 **Figure legends**

474 **Figure 1 Schematic illustration of the data collection and architecture of the**
475 **dbDEMC 3.0**

476

477 **Figure 2 Statistics of data content in dbDEMC 3.0 for humans**

478 **A.** Number of DEMs from each cancer types identified by high-throughput methods.

479 **B.** Number of DEMs from major cancer types identified by low-throughput methods.

480 **C.** The percentage of experiments for major cancer types. **D.** The percentage of
481 experiments in seven types of experimental design.

482

483 **Figure 3 Web interface of dbDEMC 3.0**

484 **A.** Search page. miRNAs can be searched via miRBase IDs, or filter experiments with

485 interested conditions. **B.** Filtering result page of experiments. **C.** Search result page

486 with example miRNAs. **D.** Experiment page. The page summarizes the description of

487 the experiments and associated differentially expressed miRNA list, functional chart

488 for expression heatmap, regularly network and miRNA targets enriched KEGG

489 pathways and GO terms are also depicted. **E.** miRNA page. This page mainly consists

490 of four sections: miRNA Summary, Expression Profile and Expression Detail and

491 Validation.

492

493 **Figure 4 miRNA cluster enrichment for 19 kinds of cancer**

494 For each cancer type, the log-odds (LOD) score is plotted. There is an enrichment of
495 miRNA cluster members for all 19 kinds of cancer (100%, P-value < 1e-4). Within
496 these 19 types of cancer, miRNAs located in clusters are, on average, 1.56 times
497 (LOD = 0.65) enriched compared to random.

498

499 **Tables**

500 **Table 1 Summary of the data content of the current release of dbDEMC**

501

502

503 **Supplementary material**

504 **Figure S1 Number of differentially expressed miRNAs identified by**
505 **high-throughput methods for each cancer type. A. Number of DEMs for mouse; B.**
506 **Number of DEMs for rat.**

507

508 **Figure S2 Increasing number of experiments for each cancer type.** The number
509 of experiments for each cancer type in dbDEMC v3.0 and v2.0 are depicted.

510

511 **Table S1 Adapters for miRNA-seq kits for the Illumina platform.**

512

513 **Table S2 The table below lists datasets collected from public resources,**
514 **including the GEO, ArrayExpress, SRA and TCGA.** The Source Data ID, PubMed
515 ID, Species, Cancer Type, platform ID and the total number of samples for each
516 dataset were listed.

517

518 **Table S3 Cancer types and associated subtypes and cell line names covered in**
519 **dbDEMC 3.0.**

520

521 **Table S4 All miRNA clusters and associated miRNA IDs identified.** A total of
 522 143 miRNA clusters and associated 688 human mature miRNAs were obtained in the
 523 human genome.

524

525 **Table S5 miRNA clusters showing a homogenous expression pattern in all 19**
 526 **kinds of cancer. T:** number of cancers for which each cluster showing a
 527 homogeneous expression pattern; **F:** number of diseases for which each cluster shows
 528 no homogeneous expression pattern; **Homogeneous-fraction:** number of diseases for
 529 which each cluster shows a homogeneous expression pattern for each miRNA cluster
 530 as defined; **P-value:** estimate by randomly linking miRNA-expression patterns 10,000
 531 times within each cancer.

532

Collecting miRNA Profiling Data From Public Resources

GEO, ArrayExpress, SRA, TCGA

Data Preprocessing

microarray

- Probe mapping
- Data imputation
- Quantile normalization

miRNA-seq

- Adapter trimming
- Reads mapping
- miRNA quantification

Data Analysis

- Experiment condition annotation
- Sample curation
- Differential expression analysis
- ID conversion
- Target identification
- Regulatory network
- Functional annotation

Low-throughput
validation data

dbDEMC
v3.0

presentation

Search & Browse

- miRNA identifier
- Cancer types
- Platforms
- Experimental design

Visualization

- Differential expression miRNAs
- Expression heatmap
- Regulatory network
- Gene Ontology terms
- KEGG terms
- miRNA expression summarization

Analysis Tools

- Blast
- Id convertor
- Meta-profiling

miRBase NCBI Ensembl HUGO MGI RGD TarBase miRTarBase KEGG GO

Sample Info Clinical Info Tumor Stage Tumor Grade Metastasis
Treatment Vital Status Overall Survival

[Home](#) [Search](#) [Browse](#) [Blast](#) [ID Converter](#) [Meta-profiling](#) [Statistics](#) [Download](#) [Help](#)

D

Differentially Expressed miRNA list

Functional Charts

E

Expression Chart

C

Expression Detail

Expression Detail

Validation Information

Cancer type	tumor subtype or cell line	Design	platform	status	PubMed ID
colon cancer	N/A	cancer vs normal	QRT-PCR	up	18230780
brain cancer	meningioma	cancer vs normal	QRT-PCR	up	19703993
lung cancer	squamous cell carcinoma	cancer vs normal	QRT-PCR	up	19584273
colon cancer	colon adenocarcinoma/tumor/colon adenomas	cancer vs normal	QRT-PCR	up	18230780

Status DOWN UP