1	The ratio of interacting miRNAs' expressions is a robust biomarker						
2	for disease classification in multi-center data						
3	Yonghao Zhang ^{a,#} , Cuidie Ma ^{a,#} , Rui Ding ^b , Han Chen ^c , Lida Xu ^{d,*} , Changyuan Yu ^{a,*}						
4	a, College of Life Science and Technology, Beijing University of Chemical Technology,						
5	Beijing, China						
6	b, Department of Laboratory Medicine, State Key Laboratory of Complex Severe and						
7	Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical						
8	Science and Peking Union Medical College, Beijing, China						
9	c, Shenyang Medical College, Shenyang, China						
10	d, Beijing Glyexo Gene Technology Co., Ltd., Beijing, China						
11	# These authors contributed equally to this work						
12	* Corresponding authors: Lida Xu, <u>lida.xu@hotgen.com.cn</u> ; Changyuan Yu,						
13	Yucy@mail.buct.edu.cn						
14							
15							
16							
17							
18							
19							
20							

21 Abstract

22 Background

Many miRNA-based diagnostic models have been constructed to distinguish diseased individuals. However, due to the inherent differences across different platforms or within multi-center data, the models usually fail in the generalization for medical application.

27 Results

Here, we proposed to use the within-sample expression ratios of related miRNA pairs as markers, by utilizing the internal miRNA: miRNA interactions. The ratio of the expression values between each miRNA pair turned out to be more stable cross multiple data source. Moreover, we adopted the genetic algorithm to solve the curse of dimensions when exploring the features.

33 Conclusions

The application results on three example datasets demonstrated that the expression ratio of interacting miRNA pair is a promising type of biomarker, which is insensitive to batch effects and has better performance in disease classifications.

Keywords: biomarker, miRNA interactions, batch effect, multi-center data, disease
 classifications

39

40 Introductions

MicroRNAs (miRNAs) have emerged as valuable biomarkers for the early diagnosis
of diseases due to their tissue-specific expression profiles and better specificity [1].
However, the expression levels of miRNAs may vary across different platforms or
protocols, which limits the application of diagnostic models. This phenomenon, known

as batch variance, is prevalent in all types of high-throughput biological platforms[2], and
exists commonly in multi-center data[3–6]. The difference in data distribution from
multiple centers is an obstacle to obtaining reliable conclusions in joint analysis, and it
prevents the models learned in one dataset from working in other external datasets[3,7,8].
Thus, effectively handling batch effects in the integration of different datasets is one of
the frontiers in large-scale biological data analysis[9].

Several batch effect correction methods have been developed to facilitate the joint 51 use of multi-center data. For example, the 'ComBat-seq' tool based on the negative 52 binomial regression model was developed specifically for RNA-seq count data[10]; the 53 'removeBatchEffect' function in 'limma' package can be used to correct the data variation 54 caused by the batch effects[11]. However, these correction methods force the data shapes 55 to be transformed artificially, which may introduce false discoveries [12] In contrast, the 56 intrinsic regulatory pathways are not affected by experimental conditions, which makes 57 the relationships between genes have the potential to be a type of normalizer-free and 58 batch-insensitive markers. Under this consideration, we propose the ratio of the 59 expression values between related miRNAs (ERRmiR) as a promising novel form of 60 biomarkers for facilitating aggregation analysis of data from multiple sources. 61

To discover ERRmiR features with biological significance, a miRNA interaction network is needed as prior knowledge. It is widely known that miRNAs not only regulate the expression of mRNAs but also target non-coding RNAs, including long non-coding RNAs and miRNAs[13]. miRNAs can directly bind to the 3'UTR of transcription factors

66	(TF), which can also reversely activate or repress miRNA expressions [14]. For example,
67	miR-181b affects the expression of miR-21 through the transcription factor FOS, a critical
68	signaling protein for glioma progression[15]; miR-660-5p has been reported to control
69	the expression of miR-486-5p via mouse double minute2 (MDM2) and p53 (also known
70	as TP53) in a study of lung cancer[16]. A recent review also summarizes numerous
71	examples of miRNA-TF and TF-miRNA interactions in various cancers, demonstrating
72	the importance of the interaction between miRNA and pluripotent transcription factor in
73	determining the occurrence of human cancers[14]. All these examples provide important
74	clues for understanding the role of the TF-mediated miRNA functional network in tumor
75	regulation.
76	In this study, we constructed a TF-mediated miRNA interaction network using public
77	databases and demonstrated that the ERRmiR features were relatively insensitive to batch
78	effects in multi-center studies. We then adopted a genetic algorithm in the feature
79	screening process to avoid the dimension curse, which had a great capacity for selecting
80	markers with stable performances in developing diagnostic models. Lastly, we used three
81	independent examples involving plasma and tissue samples to illustrate this method and
82	exhibit its effects.

83 Materials and methods

84 **Construction of miRNA interaction network**

85 The TF-mediated miRNA: miRNA interaction network was constructed by 86 combining the data of miRNA-TF and TF-miRNA relationships. If a transcription factor

regulated by miRNA_a was able to regulate miRNA_b, miRNA_a was assumed to be able to influence miRNA_b. Then, they were connected in the miRNA interaction network.

The data of those relationships were collected from several public databases. The 90 microRNA-target interactions validated experimentally were collected from miRTarBase 91 [17], among which 8014 targets were recognized as transcription factors based on the 92 hTFtarget[18] and AnimalTFDB[19] databases. On the other hand, 1266 records of 93 transcription factors regulating precursor miRNAs were obtained from the TransmiR[20] 94 v2.0 database. Combining these two parts of data, a total of 51,770 miRNA:pre-miRNA 95 96 indirect interactions were obtained. Then pre-miRNAs were mapped to mature miRNAs according to the mirbase gff3 file. Finally, the miRNA: miRNA interaction network 97 included 75,507 unique records of the indirect interaction relationships. 98

99 Feature generation

Features were generated by calculating the ratio of expression values between each related miRNA pair in the miRNA interaction network constructed above. miRNAs were filtered based on an expression threshold of 100 to ensure that miRNAs could be detected stably. To avoid the divisor being zero, the denominator was added by one. The feature constructed with the connected pair of miRNA_a and miRNA_b was denoted as ERRmiR(a,b), then the formula was as follows:

106
$$ERRmiR(a,b) = \frac{Expression of miRNA_a}{Expression of miRNA_b + 1}$$

107 Data collection and pre-processing

The data used to display the robustness of ERRmiR features on multi-source data 108 from different library preparation kits were obtained from GSE133719 and GSE141658 109 datasets on Gene Expression Omnibus (GEO)[21] database. Three examples, including 110 SARS-CoV-2-19, RCC, and LUAD projects, were used to verify the method in this study. 111 112 Data of the three projects were collected from the NCI Genomic Data Commons (GDC)[22] database and GEO (detailed in Table1). The miRNA expression matrices in 113 the CPTAC[23]/TCGA[24] database were downloaded using the GDC tool. ExceRpt[25] 114 was used to perform annotation and quantification of the raw data from GEO to obtain 115 116 the expression matrices of miRNAs. For comparison of the results among different 117 datasets within the same project, counts of reads were uniformly converted to RPM (reads 118 per million mapped reads) values. In the SARS-CoV-2 project, the plasma of persons with non-severe symptoms (mild patients and healthy) were used as the controls, and the 119 plasma of those with serious symptoms were used as the disease samples. In the RCC and 120 121 LUAD projects, normal tissues were used as the controls, and primary tumor tissues were used as the disease samples. 122

- 123
- 124 **Table1**: Sample information in detail

Project	Dataset	Number of	Number of	Platform	Source
		control cases	disease cases		
SARS-CoV-2	GSE178246*	272	264	Illumina NextSeq 500	Plasma
	GSE176498	29	16	Illumina NextSeq 550	Plasma
RCC	CPTAC-3-RCC	148	311	Illumina	tissue
	TCGA-KIRP	34	34	Illumina HiSeq 2000	tissue
	GSE109368	12	12	Illumina NextSeq 500	tissue

LUAD	CPTAC-3-LUAD	102	111	Illumina	tissue
	GSE110907	48	48	Illumina HiSeq 2000	tissue
	GSE196633	10	10	Illumina HiSeq 2500	tissue

125

* In this data set, each sample has four pieces of sequencing data, which were treated as four cases.

126 Feature screening and classification modelling

In each project, the dataset with the most samples was divided into a training set and 127 a test set proportionally and randomly according to 0.75:0.25, and the training set was 128 used to perform target screening. Univariate screening of the ERRmiR features was 129 performed based on the foldchange of the mean expression in diseased samples compared 130 to that in the controls and the p-adjust value of t-test between the two groups. The 'sklearn-131 genetic' package was used parallel for 100 times to obtain the optimal subsets of features. 132 The features with higher appearance frequencies in the optimal subsets were selected as 133 targets for the disease. 134 The 'scikit-learn' package was used to build models for disease classifications. 135

overfitting. The trained model was validated on a test set and other external validationdatasets within the same project.

During model training, the learning curves were used to detect whether the estimator was

139

136

Statistical analysis and visualization

The quartile plots of miRNA expression / ERRmiR feature values were drawn by the 'matplotlib' tool. The significance analyses were conducted using 'scipy'. The miRNA network was visualized using 'pyvis' and 'seaborn' tools. In miRNA pathway enrichment analyses, target genes of miRNAs were first identified through the database 'tarbase' using

the 'multMiR' package in R language, then pathway enrichments were performed using'clusterProfiler'.

146 **Results**

147 The schematic of ERRmiR signatures generation and screening

We developed a screening process for ERRmiR signature generation based on 148 machine learning methods (Figure 1). We first constructed a miRNA interaction network 149 by integrating several databases, including miRTarBase, hTFtarget, AnimalTFDB, and 150 151 TransmiR v2.0. The network contained 75,507 unique records of indirect interaction relationships between miRNAs. We then calculated the expression ratios of related 152 miRNA pairs as ERRmiR features. The screening dataset was randomly divided into a 153 154 training set and a test set, and the features were filtered in the training set using univariate analyses such as t-test and the foldchange of the mean expressions between two groups. 155 We used a genetic algorithm to screen the features, and those with higher frequencies in 156 the screening processes were selected as candidate markers. The trained model was 157 validated on the test set within the same screening dataset and evaluated on external 158 159 validation datasets. This approach was suitable for discovering biomarkers for various 160 samples.



161

Figure 1: Overview of the ERRmiR marker discovery process. First, the miRNA network was constructed based on the TF-mediated interactions. Then, the ERRmiR features were calculated between the connected genes in the network as new variables for subsequent process. Finally, target screening and model construction were performed on the screening dataset and verified on the validation dataset.

167 **Construction of a miRNA interaction network**

We constructed a miRNA interaction network based on indirect interactions mediated by transcription factors. The interactions mediated by transcription factors, induce the expression of one miRNA to impact the activation or inhibition of other miRNAs. Take miR-183-5p as an example to show how miRNAs regulate other miRNAs through transcription factors (**Figure 2A**). Here the pentagram-labeled miR-183-5p is a regulatory miRNA, which regulates the square-labeled transcription factors and further affects the round-labeled target miRNAs. The blue linkages represented the interaction of

miR-183-5p acting on the transcription factors, and the pink linkages represented the
effects of transcription factors on other miRNAs. The network contained 75,507 unique
records of indirect interaction relationships between miRNAs. The miRNA interaction
network was visualized, and its degree distribution and several topological characteristics
were presented in Figure 2B and Figure 2C.



Figure 2: Illustration of the miRNA interaction network. (A) TF-mediated miRNA: miRNA indirect interactions. Pentagrams denoted the regulating miRNAs, squares denoted the transcription factors, and rounds denoted the regulated miRNAs. (B) Degree distribution of the miRNA interaction network followed a power-law tail. (C) Topological characteristics of the interaction network.

186

187 Characterization of ERRmiR signatures

To verify the hypothesis that the expression ratios between interacted miRNAs 188 would be more stable across multi-center data, the distribution of ERRmiR values was 189 displayed compared to the distribution of the miRNA expression levels of the same 190 samples (Figure 3). The experiment was about the sequencing data of the peripheral 191 blood CD8+ T cells in triplicate from rheumatoid arthritis (RA) patients and healthy 192 193 controls, by parallel receiving different library construction methods. The Quartile plots showed that the original miRNA expression data generated by different library 194 preparation kits had significant differences on the scale and distributions (Figure 3, upper 195 panel), while the distribution variation of ERRmiR values decreased (Figure 3, lower 196 panel), which demonstrated the potential of ERRmiR features as batch-insensitive 197 markers. We presented three application examples from various sample types and diseases. 198



Figure 3: Quartile plots of miRNA expression (upper panel) and log2-ratios of every two miRNAs (below panel) for each sample. Each plot was represented with the median (a solid point), the 0.25 quartile and the 0.75 quartile of the distribution.

203

199

Example 1: Classification of COVID-19 patients with severe symptoms using plasma samples

The advantage of ERRmiR features was first compared to the miRNA expression values on the dataset of COVID-19 plasma samples. The dataset GSE178246 was divided into a training set and a test set randomly, and the dataset GSE176498 was used as the external independent validation set. According to the protocol, there were 42 ERRmiR targets in total obtained during conducting the genetic algorithm for 100 times on the screening dataset. As shown in **Figure 4A**, the frequency distribution of target appearance was very steep: the highest frequency was up to 60, but there were only three targets with

frequencies greater than 10. We selected the top 3 high-frequency features as markers, 213 and tested them on the validation set. As expected, they were significantly different 214 between the serious and non-serious groups (P<0.05) and showed relatively consistent 215 trends across multiple datasets (Figure 4B). Based on these markers, an SVC model that 216 was established on the training set, showed stable high performances on both the test set 217 and the validation dataset (Figure 4D). To confirm the batch-insensitive nature of the 218 ERRmiR features, the protocol of biomarker selection was also used on the expression 219 matrix of miRNAs directly. As displayed in Figure 4C, the targets screened from the 220 expression matrix of miRNAs lost effectiveness across batches of data, with miR-1224-221 5p even showing opposite regulation trends. Accordingly, the model with miRNA 222 expression values had a high AUC of 0.906 on the test set, but failed on the independent 223 validation set with an AUC of 0.783 (Figure 4E). In addition, the five miRNAs that 224 225 comprised the three ERRmiR markers were used for pathway enrichment, and the top 20 pathways were shown in Figure 4F. Infection pathways of bacteria and viruses, including 226 Salmonella infection and Human papillomavirus infection, were significantly enriched. 227



228



234 Example 2: Diagnostic model of RCC using tissue samples

The strategy of marker discovery was also validated on the dataset of RCC tissue 235 samples. Data from CPTAC-RCC dataset were used for screening targets and building 236 the model. The TCGA-KIRP and GSE109368 datasets were used for external validations. 237 After conducting the genetic algorithm, we obtained 115 targets with the frequency 238 239 distribution shown in Figure 5A. As the same as in Example 1, the top 3 high-frequency ERRmiR features were selected as biomarkers, and presented significant differences 240 between the cancer and control groups (P<0.05) with consistent regulation trends across 241 multiple datasets (Figure 5B). Although part of the miRNAs in ERRmiR markers, such 242 243 as miR-221-3p and miR-221-5p, didn't display significant differential between the two sample groups in all the datasets (Figure 5C). A prediction model using the SVC 244 245 algorithm was established on the screening dataset, and was able to achieve high AUC values on both two independent validation datasets (Figure 5D). The five miRNAs 246 comprising the three ERRmiR markers were significantly enriched in several pathways 247 associated with cancers (Figure 5E). Especially the p53 signaling pathway, and Hippo 248 signaling pathway had been widely reported to be associated with RCC[26,27]. 249



250

Figure 5: ERRmiR markers discovered in the RCC project. (A) Statistics of the frequencies of the ERRmiR features. Violin plots of the top 3 high-frequency ERRmiR features (B) and the miRNAs involved in them (C) among the three independent datasets.

(D) ROC curves of the model based on the ERRmiR markers. (E) Pathway enrichmentanalysis of miRNAs in the ERRmiR markers.

256

257 Example 3: Diagnostic model of LUAD using tissue samples

In the LUAD project, the CPTAC-LUAD dataset was used for screening targets and 258 building the model. The GSE110907 and GSE196633 datasets were used for external 259 260 validations. There were 31 targets obtained by conducting the genetic algorithm 100 times, with a relatively flat frequency distribution shown in Figure 6A. Then the top 3 high-261 frequency ERRmiR features were selected, and presented significantly differences 262 263 between the cancer and control groups (P<0.05) with consistent regulation trends across multiple datasets (Figure 6B). The model trained in the screening set, had high AUC 264 values of 0.995 and 0.91 in the GSE110907 and GSE196633 validation sets separately 265 266 (Figure 6C). The five miRNAs comprising the three ERRmiR markers were significantly 267 enriched in the p53 signaling, Cell cycle, PI3K-Akt pathways and so on, which had been 268 widely reported to be associated with LUAD[28-30] (Figure 6D).



Figure 6: Discovery results on the LUAD project. (A) Statistics of the frequencies of
ERRmiR features. (B) Violin plots of ERRmiR features ranked the top 3 by frequency.
(C) ROC curves of the models based on ERRmiR markers. (D) Pathway enrichment
analysis of miRNAs involved in the ERRmiR markers.

274

269

275 Discussion and Conclusions

Using this protocol, we discovered some miRNAs with biological significance in all three examples, reflecting a low discovery rate in the ERRmiR markers. These miRNAs

are disease-related and have been validated in previous studies. The results further demonstrate that the approach of this study is more helpful in implying the pathogenic mechanisms of diseases.

miRNA biomarkers have shown initial success in disease diagnosis and prognosis 281 monitoring [31], but the noncontrollable experimental factors can cause deviation across 282 different batches, making it difficult to use normalization of expression matrices alone for 283 multi-center applications. In this study, we proposed an algorithm based on features 284 285 formed by calculating the expression ratio of interacted miRNAs to remove batch effects. Coordinated with an integrated screening method utilizing the genetic algorithm, the 286 algorithm can distinguish negative samples from positive samples on data from multi-287 sources. We demonstrated the effectiveness of this strategy at tissue and plasma levels 288 with three examples, indicating its capacity for universal usage in developing diagnosis 289 and classification models. 290

However, in previous studies, the lack of considering biological significance has led 291 to improper strategies for construction and screening of expression ratio biomarkers. For 292 example, some studies constructed expression ratio signatures by matching pairs with an 293 upgraded gene and a downgrade gene, which ignores many worthy interactions[32,33]. 294 295 Besides, this DE-dependent method would ignore many worthy interactions, and the construction method by pairing every two genes makes the number of features explode 296 297 extremely, increasing the false discovery rate of targets and bringing tremendous pressure and difficulty to feature screening on small-sample biological data. 298

299 To address these issues, we constructed the expression ratio features based on a prior

19 / 30

knowledge of miRNA interactions, which not only reduces the dimension of features but 300 also helps to discover true relation markers. We included three types of miRNA:miRNA 301 302 interaction (direct interactions, indirect interactions, and global interactions) summarized in a previous review[34] and considered the indirect miRNA interactions mediated by 303 304 transcription factors. We constructed a TF-mediated miRNA interaction network to guide the generation of ERRmiR features, and with new miRNA regulatory relationships being 305 discovered, the interaction network will likely expand, and new markers may gradually 306 be revealed. 307

An efficient screening strategy is crucial to obtain stable biomarkers with excellent 308 performance, especially for high-dimensional data and small sample size. In this study, 309 310 we demonstrated that the expression ratios of miRNA pairs were more stable relative to the expression of individual miRNAs, and we preliminarily excluded low-expressed 311 312 miRNAs to reduce the false discover rate and the dimension in calculating the feature matrix. The screening process comprised univariate analyses and multivariate genetic 313 algorithm, and we repeated the genetic algorithm one hundred times to obtain high-314 frequency features, which were considered to be reliable. 315

Using this protocol, we discovered some miRNAs with biological significance in all three examples, reflecting a low discovery rate in the ERRmiR markers. Let-7b-5p, which is in a selected marker for predicting severe COVID-19 in the first example, has been reported to play a role in regulating ACE2 and DPP4 receptors and be significantly downregulated in nasopharyngeal swabs of patients[35]. Meanwhile, miR-21-3p which

is regulated by let-7b-5p, shows an upregulation trend in this project and is consistent 321 322 with the previous experiments of mice infected with SARS-CoV-2[36]. miR-106b-3p and 323 miR-214-5p in the ERRmiR marker that has been selected in the RCC project, are both found to be critical oncogenes in previous studies. The high expression of miR-106b-3p 324 325 may be an important factor in predicting poor prognosis in RCC patients [37,38], and the overexpression of miR-214-5p attenuates cell proliferation and metastasis[39]. In the 326 LUAD project, the pairs containing miR-30a-3p or miR-30c-2-3p have been screened out. 327 The role of the miR-30 family as tumor suppressors has been validated in previous 328 reports[40], especially miR-30c-2-3p is reported to inhibit tumor progression in 329 esophageal squamous cell carcinoma, breast cancer, and hepatocellular carcinoma[41-330 331 43]. miR-9-5p and miR-503-5p which are related with miR-30 in the markers, have also been reported to be associated with cell proliferation, migration, and invasion in non-332 333 small cell lung cancer [44,45]. These miRNAs are disease-related and have been validated in previous studies. The results further demonstrate that the approach of this study is more 334 helpful in implying the pathogenic mechanisms of diseases. 335

336

337

- 338
- 339

340

341

21 / 30

342 **Declarations**

343 Ethical approval and consent to participate

- 344 Ethical approval was not required for this study because we used a public database.
- 345 **Consent to publish**
- 346 Not applicable.

347 Availability of data and materials

- 348 Publicly available datasets were analyzed in this study. This data can be found here:
- 349 https://portal.gdc.cancer.gov/ and http://www.ncbi.nlm.nih.gov/geo/. The dataset
- 350 supporting the conclusions of this article is included within the article and its additional
- 351 file. Source code is also available if required.

352 Competing interests

353 The authors have declared that no competing interest exists.

354 Funding

This work was supported by the National Natural Science Foundation of China (Grant number 82174531).

357 Authors' contributions

- 358 L.X. and C.Y. conceived and supervised the experiments. Y.Z., C.M., and L.X. wrote the
- 359 manuscript. Y.Z., C.M., R.D. and H.C. performed the experiments. All the authors have
- 360 read and approved the final manuscript.

361 Acknowledgements

362 Not applicable.

363 Additional file

- 364 Additional file 1
- 365 Table S1. A collection of ratio features generated based on transcription factor-mediated
- 366 indirect action relationships of miRNAs.

367 Additional file 2

- Table S2. Values of the features in Figure 4B.
- 369 Table S3. Values of the features in Figure 4C.
- Table S4. Values of the features in Figure 5B.
- Table S5. Values of the features in Figure 5C.
- Table S6. Values of the features in Figure 6B.

373 **References**

- 374 [1] X. Chen, Y. Ba, L. Ma, X. Cai, Y. Yin, K. Wang, J. Guo, Y. Zhang, J. Chen, X. Guo, Q. Li, X.
- 375 Li, W. Wang, Y. Zhang, J. Wang, X. Jiang, Y. Xiang, C. Xu, P. Zheng, J. Zhang, R. Li, H. Zhang, X.
- 376 Shang, T. Gong, G. Ning, J. Wang, K. Zen, J. Zhang, C.-Y. Zhang, Characterization of microRNAs in
- 377 serum: a novel class of biomarkers for diagnosis of cancer and other diseases, Cell Res. 18 (2008)
- 378 997–1006. https://doi.org/10.1038/cr.2008.282.
- 379 [2] C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D.Y. Weiss-Solís, R.
- 380 Duque, H. Bersini, A. Nowé, Batch effect removal methods for microarray gene expression data
- integration: a survey, Brief. Bioinform. 14 (2013) 469–490. https://doi.org/10.1093/bib/bbs037.

- 382 [3] J.T. Leek, R.B. Scharpf, H.C. Bravo, D. Simcha, B. Langmead, W.E. Johnson, D. Geman, K.
- 383 Baggerly, R.A. Irizarry, Tackling the widespread and critical impact of batch effects in high-throughput
- 384 data, Nat. Rev. Genet. 11 (2010) 733–739. https://doi.org/10.1038/nrg2825.
- 385 [4] F. Heinicke, X. Zhong, M. Zucknick, J. Breidenbach, A.Y.M. Sundaram, S. T. Flåm, M. Leithaug,
- 386 M. Dalland, A. Farmer, J.M. Henderson, M.A. Hussong, P. Moll, L. Nguyen, A. McNulty, J.M. Shaffer,
- 387 S. Shore, H.K. Yip, J. Vitkovska, S. Rayner, B.A. Lie, G.D. Gilfillan, Systematic assessment of
- 388 commercially available low-input miRNA library preparation kits, RNA Biol. 17 (2020) 75-86.
- 389 https://doi.org/10.1080/15476286.2019.1667741.
- 390 [5] F. Heinicke, X. Zhong, M. Zucknick, J. Breidenbach, A.Y.M. Sundaram, S. T. Flåm, M. Leithaug,
- 391 M. Dalland, S. Rayner, B.A. Lie, G.D. Gilfillan, An extension to: Systematic assessment of
- 392 commercially available low-input miRNA library preparation kits, RNA Biol. 17 (2020) 1284–1292.
- 393 https://doi.org/10.1080/15476286.2020.1761081.
- 394 [6] S. Ibing, B.E. Michels, M. Mosdzien, H.R. Meyer, L. Feuerbach, C. Körner, On the impact of
- batch effect correction in TCGA isomiR expression data, NAR Cancer. 3 (2021) zcab007.
 https://doi.org/10.1093/narcan/zcab007.
- S. Whalen, J. Schreiber, W.S. Noble, K.S. Pollard, Navigating the pitfalls of applying machine
 learning in genomics, Nat. Rev. Genet. 23 (2022) 169–181. https://doi.org/10.1038/s41576-02100434-9.
- 400 [8] Y. Zhang, P. Patil, W.E. Johnson, G. Parmigiani, Robustifying genomic classifiers to batch
 401 effects via ensemble learning, Bioinformatics. 37 (2021) 1521–1527.
 402 https://doi.org/10.1093/bioinformatics/btaa986.

- 403 [9] W.W.B. Goh, W. Wang, L. Wong, Why Batch Effects Matter in Omics Data, and How to Avoid
- 404 Them, Trends Biotechnol. 35 (2017) 498–507. https://doi.org/10.1016/j.tibtech.2017.02.012.
- 405 [10] Z. Y, P. G, J. We, ComBat-seq: batch effect adjustment for RNA-seq count data, NAR Genomics
- 406 Bioinforma. 2 (2020). https://doi.org/10.1093/nargab/lqaa078.
- 407 [11] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, limma powers
- 408 differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43
- 409 (2015) e47–e47. https://doi.org/10.1093/nar/gkv007.
- 410 [12] V. Nygaard, E.A. Rødland, E. Hovig, Methods that remove batch effects while retaining group
- 411 differences may lead to exaggerated confidence in downstream analyses, Biostat. Oxf. Engl. 17 (2016)
- 412 29–39. https://doi.org/10.1093/biostatistics/kxv027.
- 413 [13] M. Hill, N. Tran, MicroRNAs Regulating MicroRNAs in Cancer, Trends Cancer. 4 (2018) 465-
- 414 468. https://doi.org/10.1016/j.trecan.2018.05.002.
- 415 [14] R. Vishnubalaji, H. Shaath, M. Al-Alwan, E.M. Abdelalim, N.M. Alajez, Reciprocal interplays
- 416 between MicroRNAs and pluripotency transcription factors in dictating stemness features in human
- 417 cancers, Semin. Cancer Biol. 87 (2022) 1–16. https://doi.org/10.1016/j.semcancer.2022.10.007.
- 418 [15] T. Tao, Y. Wang, H. Luo, L. Yao, L. Wang, J. Wang, W. Yan, J. Zhang, H. Wang, Y. Shi, Y. Yin,
- 419 T. Jiang, C. Kang, N. Liu, Y. You, Involvement of FOS-mediated miR-181b/miR-21 signalling in the
- 420 progression of malignant gliomas, Eur. J. Cancer Oxf. Engl. 1990. 49 (2013) 3055-3063.
- 421 https://doi.org/10.1016/j.ejca.2013.05.010.
- 422 [16] C. Borzi, L. Calzolari, G. Centonze, M. Milione, G. Sozzi, O. Fortunato, mir-660-p53-mir-486
- 423 Network: A New Key Regulatory Pathway in Lung Tumorigenesis, Int. J. Mol. Sci. 18 (2017) 222.

- 424 https://doi.org/10.3390/ijms18010222.
- 425 [17] H.-Y. Huang, Y.-C.-D. Lin, J. Li, K.-Y. Huang, S. Shrestha, H.-C. Hong, Y. Tang, Y.-G. Chen,
- 426 C.-N. Jin, Y. Yu, J.-T. Xu, Y.-M. Li, X.-X. Cai, Z.-Y. Zhou, X.-H. Chen, Y.-Y. Pei, L. Hu, J.-J. Su, S.-
- 427 D. Cui, F. Wang, Y.-Y. Xie, S.-Y. Ding, M.-F. Luo, C.-H. Chou, N.-W. Chang, K.-W. Chen, Y.-H.
- 428 Cheng, X.-H. Wan, W.-L. Hsu, T.-Y. Lee, F.-X. Wei, H.-D. Huang, miRTarBase 2020: updates to the
- 429 experimentally validated microRNA-target interaction database, Nucleic Acids Res. 48 (2020) D148-
- 430 D154. https://doi.org/10.1093/nar/gkz896.
- 431 [18] Q. Zhang, W. Liu, H.-M. Zhang, G.-Y. Xie, Y.-R. Miao, M. Xia, A.-Y. Guo, hTFtarget: A
- 432 Comprehensive Database for Regulations of Human Transcription Factors and Their Targets,
- 433 Genomics Proteomics Bioinformatics. 18 (2020) 120–128. https://doi.org/10.1016/j.gpb.2019.09.006.
- 434 [19] H. Hu, Y.-R. Miao, L.-H. Jia, Q.-Y. Yu, Q. Zhang, A.-Y. Guo, AnimalTFDB 3.0: a comprehensive
- 435 resource for annotation and prediction of animal transcription factors, Nucleic Acids Res. 47 (2019)
- 436 D33–D38. https://doi.org/10.1093/nar/gky822.
- 437 [20] Z. Tong, Q. Cui, J. Wang, Y. Zhou, TransmiR v2.0: an updated transcription factor-microRNA
- 438 regulation database, Nucleic Acids Res. 47 (2019) D253–D258. https://doi.org/10.1093/nar/gky1023.
- 439 [21] E. Clough, T. Barrett, The Gene Expression Omnibus Database, Methods Mol. Biol. Clifton NJ.
- 440 1418 (2016) 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5.
- 441 [22] M.A. Jensen, V. Ferretti, R.L. Grossman, L.M. Staudt, The NCI Genomic Data Commons as an
- 442 engine for precision medicine, Blood. 130 (2017) 453–459. https://doi.org/10.1182/blood-2017-03443 735654.
- 444 [23] N.J. Edwards, M. Oberti, R.R. Thangudu, S. Cai, P.B. McGarvey, S. Jacob, S. Madhavan, K.A.

- 445 Ketchum, The CPTAC Data Portal: A Resource for Cancer Proteomics Research, J. Proteome Res. 14
- 446 (2015) 2707–2713. https://doi.org/10.1021/pr501254j.
- 447 [24] C. Hutter, J.C. Zenklusen, The Cancer Genome Atlas: Creating Lasting Value beyond Its Data,
- 448 Cell. 173 (2018) 283–285. https://doi.org/10.1016/j.cell.2018.03.042.
- 449 [25] J. Rozowsky, R.R. Kitchen, J.J. Park, T.R. Galeev, J. Diao, J. Warrell, W. Thistlethwaite, S.L.
- 450 Subramanian, A. Milosavljevic, M. Gerstein, exceRpt: A Comprehensive Analytic Platform for
- 451 Extracellular RNA Profiling, Cell Syst. 8 (2019) 352-357.e3.
- 452 https://doi.org/10.1016/j.cels.2019.03.004.
- 453 [26] K.V. Gurova, J.E. Hill, O.V. Razorenova, P.M. Chumakov, A.V. Gudkov, p53 pathway in renal
- 454 cell carcinoma is repressed by a dominant mechanism, Cancer Res. 64 (2004) 1951-1958.
- 455 https://doi.org/10.1158/0008-5472.can-03-1541.
- 456 [27] Y. Guan, Z. Gong, T. Xiao, Z. Li, Knockdown of miR-572 suppresses cell proliferation and
- 457 promotes apoptosis in renal cell carcinoma cells by targeting the NF2/Hippo signaling pathway, Int. J.
- 458 Clin. Exp. Pathol. 11 (2018) 5705–5714.
- 459 [28] X. Huang, L. Jiang, S. Lu, M. Yuan, H. Lin, B. Li, Z. Wen, Y. Zhong, Overexpression of
- 460 ERCC6L correlates with poor prognosis and confers malignant phenotypes of lung adenocarcinoma,
- 461 Oncol. Rep. 48 (2022) 131. https://doi.org/10.3892/or.2022.8342.
- 462 [29] H. Tang, J. Liu, J. Huang, GMFG (glia maturation factor gamma) inhibits lung cancer growth
- 463 by activating p53 signaling pathway, Bioengineered. 13 (2022) 9284–9293.
- 464 https://doi.org/10.1080/21655979.2022.2049958.
- 465 [30] N. Zhang, S. Cao, R. Sun, Y. Wang, L. Liu, W. Wang, X. Meng, Signal peptidase 21 suppresses

- 466 cell proliferation, migration, and invasion via the PTEN-PI3K/Akt signaling pathway in lung
- 467 adenocarcinoma, PeerJ. 10 (2022) e14206. https://doi.org/10.7717/peerj.14206.
- 468 [31] J. Inoue, J. Inazawa, Cancer-associated miRNAs and their therapeutic potential, J. Hum. Genet.
- 469 66 (2021) 937–945. https://doi.org/10.1038/s10038-021-00938-6.
- 470 [32] J. Zhang, G.S. Raju, D.W. Chang, S.-H. Lin, Z. Chen, X. Wu, Global and targeted circulating
- 471 microRNA profiling of colorectal adenoma and colorectal cancer, Cancer. 124 (2018) 785-796.
- 472 https://doi.org/10.1002/cncr.31062.
- 473 [33] A.E. Szafranska, T.S. Davison, J. John, T. Cannon, B. Sipos, A. Maghnouj, E. Labourier, S.A.
- 474 Hahn, MicroRNA expression alterations are linked to tumorigenesis and non-neoplastic processes in
- 475 pancreatic ductal adenocarcinoma, Oncogene. 26 (2007) 4442-4452.
- 476 https://doi.org/10.1038/sj.onc.1210228.
- 477 [34] M. Hill, N. Tran, miRNA interplay: mechanisms and consequences in cancer, Dis. Model. Mech.
- 478 14 (2021) dmm047662. https://doi.org/10.1242/dmm.047662.
- 479 [35] A. Latini, C. Vancheri, F. Amati, E. Morini, S. Grelli, C. Matteucci, V. Petrone, V.L. Colona, M.
- 480 Murdocca, M. Andreoni, V. Malagnino, M. Raponi, D. Cocciadiferro, A. Novelli, P. Borgiani, G.
- 481 Novelli, Expression analysis of miRNA hsa-let7b-5p in naso-oropharyngeal swabs of COVID-19
- 482 patients supports its role in regulating ACE2 and DPP4 receptors, J. Cell. Mol. Med. 26 (2022) 4940-
- 483 4948. https://doi.org/10.1111/jcmm.17492.
- 484 [36] S. Nersisyan, N. Engibaryan, A. Gorbonos, K. Kirdey, A. Makhonin, A. Tonevitsky, Potential
- 485 role of cellular miRNAs in coronavirus-host interplay, PeerJ. 8 (2020) e9994.
- 486 https://doi.org/10.7717/peerj.9994.

487 [37] Y. Li, D. Chen, Z. Su, Y. Li, J. Liu, L. Jin, M. Shi, Z. Jiang, Z. Qi, Y. Gui, S. Yang, X. Mao, X.

- 488 Wu, Y. Lai, MicroRNA-106b functions as an oncogene in renal cell carcinoma by affecting cell 489 proliferation, migration apoptosis, Mol. Med. Rep. 13 (2016) 1420-1426. and 490 https://doi.org/10.3892/mmr.2015.4656. [38] K. Liu, X. Pan, X. Peng, C. Zhang, H. Li, X. Guan, W. Xu, J. Xu, L. Zhao, T. Wang, Y. Lai, 491 492 Associations of high expression of miR-106b-5p detected from FFPE sample with poor prognosis of 493 RCC patients, Pathol. Res. Pract. 215 (2019) 152391. https://doi.org/10.1016/j.prp.2019.03.019. 494 [39] R. Guo, B. Zou, Y. Liang, J. Bian, J. Xu, Q. Zhou, C. Zhang, T. Chen, M. Yang, H. Wang, F. Pei, 495 Z. Xu, LncRNA RCAT1 promotes tumor progression and metastasis via miR-214-5p/E2F2 axis in 496 renal cell carcinoma, Cell Death Dis. 12 (2021) 1-14. https://doi.org/10.1038/s41419-021-03955-7. 497 [40] A.D. Saleh, H. Cheng, S.E. Martin, H. Si, P. Ormanoglu, S. Carlson, P.E. Clavijo, X. Yang, R. 498 Das, S. Cornelius, J. Couper, D. Chepeha, L. Danilova, T.M. Harris, M.B. Prystowsky, G.J. Childs, 499 R.V. Smith, A.G. Robertson, S.J.M. Jones, A.D. Cherniack, S.S. Kim, A. Rait, K.F. Pirollo, E.H. Chang, 500 Z. Chen, C. Van Waes, Integrated Genomic and Functional microRNA Analysis Identifies miR-30-5p 501 as a Tumor Suppressor and Potential Therapeutic Nanomedicine in Head and Neck Cancer, Clin. 502 Cancer Res. Off. J. Am. Assoc. Cancer Res. 25 (2019) 2860-2873. https://doi.org/10.1158/1078-503 0432.CCR-18-0716. 504
- [41] T. Ma, Y. Zhao, Q. Lu, Y. Lu, Z. Liu, T. Xue, Y. Shao, MicroRNA-30c functions as a tumor
- 505 suppressor via targeting SNAI1 in esophageal squamous cell carcinoma, Biomed. Pharmacother.
- 506 Biomedecine Pharmacother. 98 (2018) 680-686. https://doi.org/10.1016/j.biopha.2017.12.095.
- 507 [42] H.-D. Zhang, L.-H. Jiang, J.-C. Hou, S.-Y. Zhou, S.-L. Zhong, L.-P. Zhu, D.-D. Wang, S.-J. Yang,

- 508 Y.-J. He, C.-F. Mao, Y. Yang, J.-Y. Wang, Q. Zhang, H.-Z. Xu, D.-D. Yu, J.-H. Zhao, J.-H. Tang, Z.-L.
- 509 Ji, Circular RNA hsa circ 0072995 promotes breast cancer cell migration and invasion through
- 510 sponge for miR-30c-2-3p, Epigenomics. 10 (2018) 1229–1242. https://doi.org/10.2217/epi-2018-0002.
- 511 [43] J. Zhang, M. Cai, D. Jiang, L. Xu, Upregulated LncRNA-CCAT1 promotes hepatocellular
- 512 carcinoma progression by functioning as miR-30c-2-3p sponge, Cell Biochem. Funct. 37 (2019) 84–
- 513 92. https://doi.org/10.1002/cbf.3375.
- 514 [44] Y. Sun, L. Li, S. Xing, Y. Pan, Y. Shi, L. Zhang, Q. Shen, miR-503-3p induces apoptosis of lung
- 515 cancer cells by regulating p21 and CDK4 expression, Cancer Biomark. Sect. Dis. Markers. 20 (2017)
- 516 597–608. https://doi.org/10.3233/CBM-170585.
- 517 [45] K. Zhu, J. Lin, S. Chen, Q. Xu, miR-9-5p Promotes Lung Adenocarcinoma Cell Proliferation,
- 518 Migration and Invasion by Targeting ID4, Technol. Cancer Res. Treat. 20 (2021) 15330338211048592.
- 519 https://doi.org/10.1177/15330338211048592.
- 520