Bioinformatics, 36(9), 2020, 2657–2664
doi: 10.1093/bioinformatics/btaa015
Advance Access Publication Date: 13 January 2020
Original Paper

Genome analysis

MDEHT: a multivariate approach for detecting differential expression of microRNA isoform data in RNA-sequencing studies

Md Amanullah[1,†] Mengqian Yu[2,†], Xiwei Sun[2,3,†], Aoran Luo[1], Qing Zhou[1], Liyuan Zhou[2], Ling Hou[1], Wei Wang[3,4], Weiguo Lu[1], Pengyuan Liu[2,4,5,]* and Yan Lu[1,]*

[1]Center for Uterine Cancer Diagnosis & Therapy Research of Zhejiang Province, Women's Reproductive Health Key Laboratory of Zhejiang Province and Department of Gynecologic Oncology, Women's Hospital and Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310029, China, [2]Department of Respiratory Medicine, Sir Run Run Shaw Hospital and Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310016, China, [3]Institute for Advanced Research, Wenzhou Medical University, Wenzhou, Zhejiang 325035, China, [4]Department of Pathology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310029, China and [5]Department of Physiology, Center of Systems Molecular Medicine, Medical College of Wisconsin, Milwaukee, WI 53226, USA

*To whom correspondence should be addressed.

Associate Editor: Peter Robinson

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Received on November 3, 2019; revised on December 15, 2019; editorial decision on January 5, 2020; accepted on January 9, 2020

Abstract

Motivation: miRNA isoforms (isomiRs) are produced from the same arm as the archetype miRNA with a few nucleotides different at 5 and/or 3 termini. These well-conserved isomiRs are functionally important and have contributed to the evolution of miRNA genes. Accurate detection of differential expression of miRNAs can bring new insights into the cellular function of miRNA and a further improvement in miRNA-based diagnostic and prognostic applications. However, very few methods take isomiR variations into account in the analysis of miRNA differential expression.

Results: To overcome this challenge, we developed a novel approach to take advantage of the multidimensional structure of isomiR data from the same miRNAs, termed as a multivariate differential expression by Hotelling's T^2 test (MDEHT). The utilization of the information hidden in isomiRs enables MDEHT to increase the power of identifying differentially expressed miRNAs that are not marginally detectable in univariate testing methods. We conducted rigorous and unbiased comparisons of MDEHT with seven commonly used tools in simulated and real datasets from The Cancer Genome Atlas. Our comprehensive evaluations demonstrated that the MDEHT method was robust among various datasets and outperformed other commonly used tools in terms of Type I error rate, true positive rate and reproducibility.

Availability and implementation: The source code for identifying and quantifying isomiRs and performing miRNA differential expression analysis is available at https://github.com/amanzju/MDEHT.

Contact: yanlu76@zju.edu.cn or pyliu@zju.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

MicroRNAs (miRNAs) are a class of small non-coding RNA molecules with 19–24 nt in length (Lagos-Quintana *et al.*, 2001). miRNAs are critical post-transcriptional regulators of gene expression that act by degrading their RNA targets or by repressing the translation of mRNAs (Ha and Kim, 2014). miRNAs are abundant and highly conserved in organisms and have been estimated to regulate > 50% of genes in the genome (Friedman *et al.*, 2009). With the advent of next-generation sequencing technology, over 2000 miRNAs have been identified in the human genome. miRNAs play a critical role in numerous cellular functions in diverse processes such as cell proliferation, cell death, fat metabolism, hematopoietic differentiation and immunity (Hanna *et al.*, 2019; Wang *et al.*, 2019). Aberrant miRNA expression is associated with many diseases, including cancer (Acunzo *et al.*, 2015; Andrew *et al.*, 2019; Yan *et al.*, 2019).

Most miRNAs are transcribed from DNA sequences into primary miRNAs (pri-miRNAs) and processed into precursor miRNAs (pre-miRNAs) and mature miRNAs (Ha and Kim, 2014; Macfarlane and Murphy, 2010; O'Brien *et al.*, 2018). Specifically, miRNA biogenesis is involved in the following major steps: (i) DNAs are transcribed into pri-miRNA with the processing of RNA polymerase II/III. (ii) The pri-miRNA is cleaved to generate the pre-miRNA by the microprocessor complex comprised of Drosha and DGCR8. (iii) The pre-miRNA is assembled into a complex with the nucleocytoplasmic transport factor Exportin-5 and RanGTP and is subsequently translocated into the cytoplasm. (iv) The cytoplasmic pre-miRNA is processed into miRNA duplex by Dicer. (v) The miRNA duplex liberates the mature miRNA to assemble into the Argonaute family of proteins to form RNA-induced silencing complex.

miRNA isoforms (isomiRs) are miRNA sequences that have variations with respect to the reference sequence (Morin *et al.*, 2008). These sequence variants typically differ from the mature miRNA reference sequences at either their 5′ or 3′ ends. isomiRs are probably generated by the sequential cleavages catalyzed by Drosha and/or Dicer enzymes, although other endonucleases enzymes could also be involved (Neilsen *et al.*, 2012). isomiRs are produced constitutively in human tissues, and their expression depends on tissue type, tissue state, disease subtype, person's sex, population origin and race (Telonis *et al.*, 2017). Many isomiRs are conserved across species and in some cases, differentially expressed according to the tissue or developmental stages (Neilsen *et al.*, 2012).

A fundamental goal of RNA-sequencing (RNA-seq) is to identify expression changes between different biological or disease conditions (Chu *et al.*, 2015). Many bioinformatics tools such as DEseq and edgeR for detecting differential expression from RNA-seq count data have been developed (Anders and Huber, 2010; Auer and Doerge, 2011; Di *et al.*, 2011; Love *et al.*, 2014; Robinson *et al.*, 2010; Smyth, 2004). However, very few tools take isomiRs into account in differential expression analysis. In most analyses, isoforms of a miRNA are treated as the same miRNA and read counts of isomiRs are combined. The utilization of the relationship of isomiR expression may aid in the detection of differential expression of its canonical miRNA. To overcome this challenge, we developed a novel approach to take advantage of the multidimensional structure of isomiR data from the same miRNAs, termed as multivariate differential expression by Hotelling's T^2 test (MDEHT; https://github.com/amanzju/MDEHT). Hotelling's T^2 test is a generalization of the Student's *t*-statistic and is widely used for testing the difference in two multivariate means. To utilize isomiRs, we also developed a computational tool called isomiRseeker to identify isomiRs from the miRNA-sequencing (miRNA-seq) data. We conducted rigorous and unbiased comparisons of MDEHT with seven commonly used tools in simulated and real datasets from The Cancer Genome Atlas (TCGA; https://cancergenome.nih.gov/). Our comprehensive evaluations demonstrated that the newly developed MDEHT method was robust among various datasets and outperformed the other tools in terms of Type I error rate, true positive rate and reproducibility. We also performed *in vitro* cell-based assays for a novel miRNA miR-335-3p in uterine corpus endometrial carcinoma (UCEC) for further validation of MDEHT.

2 Materials and methods

2.1 Hotelling's T^2 statistic

The multivariate Hotelling's T^2 statistic is a generalization of the univariate Student *t*-statistic, proportional to the *F*-distribution that is used in multivariate hypothesis testing (Hotelling, 1931). Suppose we generate read count data of miRNAs from a miRNA-seq study where n_x represents the sample size from the treatment group and n_y represents the sample size from the control group. Suppose that there are a total of L isomiRs in a particular miRNA in both groups (treatment and control). Each miRNA is used as a single variable to construct a T^2 statistic where each miRNA has more than one isoform. LetX_{ik}be the expression level for kth isomiR of a tested miRNA from ith sample in the treatment group andY_{jk}be the expression level for kth isomiR of that miRNA from jth sample in the control group. The expression level vectors for samples i and j are defined as $X_i = (X_{i1}, X_{i2}, \ldots, X_{iL})^T$ and $Y_j = (Y_{j1}, Y_{j2}, \ldots, Y_{jL})^T$ in the treatment and control groups, respectively. The mean expression levels of kth isomiR of the tested miRNA in the treatment and control groups can be expressed as

$$\bar{X}_k = 1/n_x \sum_{i=1}^{n_x} X_{ik} \text{ and } \bar{Y}_k = 1/n_y \sum_{j=1}^{n_y} Y_{jk},$$

respectively. The mean expression level vectors for the tested miRNA in the treatment and control groups can be expressed as $\bar{X} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_L)^T$ and$\bar{Y} = (\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_L)^T$, respectively. The pooled variance–covariance matrix of expression levels for the tested miRNA in both groups is then defined as,

$$\begin{aligned}\sum &= \frac{(n_x-1)\sum_1 + (n_y-1)\sum_2}{n_x+n_y-2}\\ &= \frac{1}{n_x+n_y-2}\left[\sum_{i=1}^{n_x}(X_i-\bar{X})(X_i-\bar{X})^T + \sum_{j=1}^{n_y}(Y_j-\bar{Y})(Y_j-\bar{Y})^T\right],\end{aligned} \tag{1}$$

where $\sum_1$ and $\sum_2$ are the variance-covariance matrix of expression levels for the tested miRNA in treatment and control groups, respectively. Hotelling's T^2 statistic for miRNA differential expression studies is then defined as (Lu *et al.*, 2005),

$$T^2 = \frac{n_x n_y}{n_x+n_y}(\bar{X}-\bar{Y})\Sigma^{-1}(\bar{X}-\bar{Y})^T. \tag{2}$$

Under the null hypothesis that the distributions in both groups are the same, when both have a large sample size, the central limit theorem dictates that,

$$\frac{n_x+n_y-L-1}{L(n_x+n_y-2)}T^2 \tag{3}$$

is asymptotically *F*-distributed with L degrees of freedom for the numerator and $n_x + n_y - L - 1$ for the denominator.

A breaking assumption of Hotelling's T^2 statistic is that if the determinant of the pooled variance–covariance matrix is zero, then Hotelling's T^2 will break. Because an inverse of a square matrix is not possible if the determinant of that matrix is zero, although this condition arises very rarely. We overcame this problem using generalized techniques for inversion of a non-invertible matrix. Specifically, the most well-known Moore-Penrose inverse pseudoinverse technique was applied (Penrose, 1955), implemented using 'ginv' function from R package 'MASS'.

2.2 Identification of isomiRs from miRNA-seq

To utilize isomiRs, we developed a computational tool isomiRseeker to identify isomiRs from miRNA-seq data. Briefly, we downloaded the gene annotations (hg19) and corresponding reference sequences of 2794 mature miRNA in human from miRbase (v21). Then, we built the isomiRs annotation database using these mature miRNA sequences. The start position of isomiRs can be from 5-nt upstream and 5-nt downstream sequence of the 5′ end of mature miRNA, whereas its endpoint ranged from 5-nt upstream to 5-nt downstream sequence of the 3′ end of mature miRNA. As a result, we can generate a total 121 unique sequences (representing 121 isoforms) for a single miRNA. According to this rule, a total of 338 074 (2794 × 121) unique sequences were generated in the isomiRs annotation database.

Next, we downloaded miRNA-seq BAM files of 5743 tumor samples and 546 normal samples among 11 types of cancer from the TCGA data portal (https://portal.gdc.cancer.gov/). These cancer types include bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma LUAD), lung squamous cell carcinoma (LUSC), pancreatic adenocarcinoma (PRAD), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC). Among these 11 cancer types, enough corresponding normal samples ($n > 15$) were available (Supplementary Table S1). Read mapping and quantification of isomiRs were analyzed using our computational tool isomiRseeker. Briefly, the reads that aligned to the reference human genome in the BAM files were first remapped to our isomiR annotation database using bwa (Li and Durbin, 2009), allowing no mismatch per reading. Then, these remapped reads were used to count the number of reads belonging to each of the candidate isomiRs. Currently, our isomiRseeker tool can only handle template isomiRs that are based on known mature miRNAs. However, users are allowed to manually add their non-template isomiR sequences to the isomiRs annotation database so that these non-template isomiRs can be analyzed in the study.

Prior to the differential expression analysis, isomiRs with zero read count in >80% samples were filtered. In the count data matrix, a row represents an isomiR, and a column represents a sample. The expression level of isomiRs was calculated as Reads Per Million (RPM) mapped reads, which has been commonly used in previous miRNA and lncRNA studies (de Rie *et al.*, 2017; Yan *et al.*, 2015). Furthermore, the expression of isomiRs was log2-transformed across samples before subsequent downstream analysis.

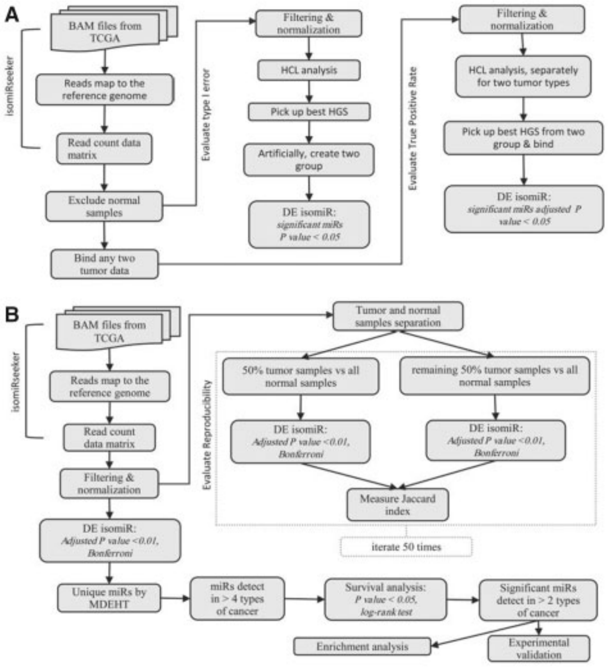
A
BAM files from TCGA
isomiRseeker
Reads map to the reference genome
Read count data matrix
Exclude normal samples
Bind any two tumor data
Evaluate type I error
Filtering & normalization
HCL analysis
Pick up best HGS
Artificially, create two group
DE isomiR: significant miRs P value < 0.05
Evaluate True Positive Rate
Filtering & normalization
HCL analysis, separately for two tumor types
Pick up best HGS from two group & bind
DE isomiR: significant miRs adjusted P value < 0.05
B
BAM files from TCGA
isomiRseeker
Reads map to the reference genome
Read count data matrix
Filtering & normalization
DE isomiR: Adjusted P value <0.01, Bonferroni
Evaluate Reproducibility
Tumor and normal samples separation
50% tumor samples vs all normal samples
remaining 50% tumor samples vs all normal samples
DE isomiR: Adjusted P value <0.01, Bonferroni
DE isomiR: Adjusted P value <0.01, Bonferroni
Measure Jaccard index
iterate 50 times
Unique miRs by MDEHT
miRs detect in > 4 types of cancer
Survival analysis: P value < 0.05, log-rank test
Significant miRs detect in > 2 types of cancer
Enrichment analysis
Experimental validation

Fig. 1. Flowchart for simulation studies and real data analyses. (**A**) Simulation datasets were generated by directly sampling miRNA read count data from TCGA. Two types of simulation datasets were generated, one for evaluating the Type I error rate under the null hypothesis and another for evaluating the true positive rate under the alternative hypothesis. (**B**) Real data from TCGA were used to evaluate the reproducibility of results from statistical methods and to detect DEmiRs in each cancer type. To evaluate the reproducibility of DEmiR results, tumor samples were randomly divided into two groups with equal sample size in each cancer type. Samples from each tumor groups were compared with normal samples from the same cancer types for identifying DEmiRs. To detect DEmiRs, all tumor and normal samples from the same cancer type were analyzed by these statistical methods. Functional enrichment analysis was performed on DEmiRs uniquely detected by the MDEHT method. Eleven types of cancer from TCGA were used in simulation studies and real data analyses. Read mapping and quantification of isomiRs were analyzed using our computational pipeline isomiRseeker. HGS represents the homogeneous sample and HCL represents the hierarchical cluster

2.3 Simulation studies

Simulated datasets that are generated from assumed probabilistic models might not closely recapitulate the complex structure of real RNA-seq data. Instead, we generated simulation datasets by directly sampling miRNA read count data from TCGA datasets. Two types of simulation datasets were generated (Fig. 1A), one for evaluating the Type I error rate and another for evaluating the true positive rate of different statistical tools for miRNA differential expression analysis.

To evaluate the Type I error rate of statistical tools for identifying miRNA differential expression, we generated simulated datasets under the null hypothesis (i.e. two multivariate means are equal) from TCGA tumor samples (Fig. 1A). Briefly, read count data of isomiRs were obtained from miRNA-seq BAM files in TCGA using our computational pipeline isomiRseeker as described above. Lowly expressed isomiR was further filtered if read count of an isoform is zero in > 80% of tumor samples in a cancer type. Then, read count data were transformed into RPM values. Unsupervised hierarchical cluster analysis was applied to the log2-transformed RPM data in each cancer type. Euclidian distance and ward.D2 cluster method were used in the cluster analysis. From the cluster analysis, the best homogeneous samples were identified in each cancer type (Supplementary Fig. S1). Since most cancer types have very limited numbers of normal specimens, only tumor samples were used in the clustering analysis to generate simulation datasets under the null hypothesis. Finally, equal numbers of tumor specimens were sampled without replacement from the best homogeneous samples in each cancer type, forming the treatment and control groups. Subsequently, differential expression analysis of miRNA was performed on the simulated datasets using various statistical tools. Type I error rate was defined as the proportion of miRNAs with P-values <0.05 from a given statistical test. Simulation datasets were generated from each cancer type separately; simulation studies were repeated 100 times in each cancer type.

To evaluate the true positive rate of statistical tools for identifying miRNA differential expression, we also generated simulated datasets under an alternative hypothesis (i.e. two multivariate means are unequal) from TCGA tumor samples (Fig. 1A). Processing and filtering the read counts of isomiRs from miRNA-seq BAM files were performed as described above. Then, any two tumor types of read count data were combined into a single data frame. This resulted in a total of 55 combined datasets by considering all the possible combinations of two cancers among 11 cancer types. Similar cluster analysis was performed on the combined datasets. Two distinct clusters were identified, each of which represents homogeneous samples from one of the two tumor types in each of the combined datasets (Supplementary Fig. S2). Finally, in each of combined datasets, equal numbers of tumor specimens were respectively sampled without replacement from the identified two clusters, forming the treatment and control groups. Subsequently, differential expression analysis of miRNA was performed on the simulated datasets using various statistical tools. The true positive rate was defined as the proportion of miRNA with an adjusted P-value < 0.05 from a given statistical test. Adjusted P-values (i.e. false discovery rate) were obtained using the Benjamini–Hochberg method (Benjamini and Hochberg, 1995). Simulation studies were repeated 100 times for any of the two cancer types.

2.4 Real data analysis

Similarly, processing and filtering of read counts of isomiRs from miRNA-seq data in TCGA tumor and normal samples were performed as described above. To evaluate the reproducibility of each statistical method for detecting miRNA differential expression, tumor samples were randomly divided into two groups (A and B) with equal sample size in each cancer type. Since each cancer type has limited normal samples, the normal samples are not split into two groups. Samples from each tumor group were compared with normal samples from the same cancer types for identifying differentially expressed miRNAs (DEmiRs) (false discovery rate < 0.01; Fig. 1B). The Jaccard similarity index from two datasets was calculated for evaluating the reproducibility of each statistical method (Levandowsky and Winter, 1971).

We comprehensively assessed our new MDEHT with other seven methods commonly used for differential expression analysis in simulated and real datasets from TCGA. Specifically, for empirical analysis of digital gene expression in R (edgeR) (Robinson *et al.*, 2010), differential expression analysis for sequence count data (DESeq) (Anders and Huber, 2010), DESeq2 (Love *et al.*, 2014), negative binomial models for RNA-seq data (NBPSeq) (Di *et al.*, 2011) and two-stage Poisson model (TSPM) (Auer and Doerge, 2011), read counts of miRNAs are directly inputted to the analysis, while for Voom (+limma; Smyth, 2004), Vst (+limma; Smyth, 2004) and MDEHT, RPM data of miRNAs are inputted to the analysis. All the analyses were performed with default parameters.

2.5 Cell culture

HEC-1-B cells were obtained from the American Type Culture Collection and were cultured in minimum essential medium (Gibco, USA) supplemented with 10% fetal bovine serum (FBS) (Gibco) and 1% penicillin-streptomycin solution (Gibco). Cells were incubated in a CO_2 incubator (Thermo Fisher Scientific, USA) maintained at 37°C with humidified air and 5% CO_2.

2.6 Tissue specimens

A total of 47 paired UCEC tumor tissues, and the corresponding adjacent non-tumor tissues were obtained at the time of surgery. The diagnosis of all the tissues was confirmed with histopathology, and the TNM Classification of Malignant Tumors (TNM) clinical stages were determined based on the American Joint Committee on Cancer and the Union for International Cancer Control in 2002. The study protocol was reviewed and approved by the Ethics Committees of Women's Hospital of Zhejiang University School of Medicine (Hangzhou, China).

2.7 Oligonucleotide transfection

The miR-335-3p inhibitor was designed and synthesized by GenePharma (Shanghai, China). Cells were transfected in individual wells of 6-well plates with an inhibitor targeting miR-335-3p or negative control (NC) using GeneMute^{TM} reagent (Shanghai, China), according to the manufacturer's instructions. The coding strand of the inhibitor was 5′-GGUCAGGAGCAAUAAUGAAAAA-3′.

2.8 Quantitative real-time PCR analysis

Total RNA was extracted from cells or tissues using the TRIzol® reagent (Invitrogen, USA). For miRNA detection, reverse-transcribed complementary DNA was synthesized with the HiScript®. II first Strand cDNA Synthesis Kit (Vazyme, China) with gene-specific primers for miR-335-3p (Ribobio, China). qPCR analyses were performed with HiScript® II One Step quantitative real-time PCR (qRT-PCR) SYBR Green Kit (Vazyme) and normalized to U6 small nuclear RNA expression. The primers were ordered from Ribobio.

2.9 Western blot analysis

The whole-cell lysates were prepared using RIPA lysis buffer (Beyotime Biotechnology, China) following by centrifugation at 13 000 rpm for 15 min. The protein concentration was determined with a BCA assay (Thermo Fisher Scientific, USA). Equal amounts of proteins were separated on 10% SDS-PAGE, then transferred onto polyvinylidene fluoride membrane, blocked with 5% skim milk and incubated overnight with primary antibody at 4°C. The membranes were then incubated with a suitable secondary antibody conjugated with horseradish peroxidase for 1 h at room temperature, and visualization of hybridization was carried out using a chemiluminescence's reagent. The primary antibodies used are phospho-Wee1 (Ser642), Wee1, Cyclin E2, CDK2, cdc42, p21, p16 and GAPDH obtained from Cell Signaling Technology (Danvers, MA, USA).

2.10 Cell proliferation assay

Cell proliferation was measured with the Cell Counting Kit-8 (CCK-8) (MedChemExpress, USA) following the manufacturer's instructions. In brief, ~3000 cells were placed into each well of 96-well plates after transfection for 24 h, and CCK-8 solution was added after cells attached to the wall (0 h). CCK-8 solution was added into cells every 24 h for 4 days. The absorbance was measured at 450 nm after incubating for 2 h at 37°C.

2.11 *In vitro* migration and invasion assays

For transwell migration assays, 3×10^4 serum-free cells were plated in the top chamber of each insert (Corning, USA) with a non-coated membrane. For invasion assays, 3×10^4 serum-free cells were added to the upper chamber with Matrigel (Corning, USA). For both assay types, 500 μl of medium supplemented with 10% FBS was injected into the lower chambers. After incubating for 16 h, the inserts were fixed in 100% methanol and stained in 0.1% crystal violet. Cells adhering to the lower membrane of the inserts were imaged with a Leica DM4000 microscope (Germany).

2.12 Cell cycle analysis

After transfection for 48 h, the cells were harvested and fixed overnight in 70% ethanol at −20°C. The fixed cells were washed three times with phosphate-buffered saline and stained with propidium iodide (BD Biosciences, USA). DNA contents were measured with a Cytoflex S flow cytometer (Beckman, USA), and the results were analyzed using FlowJo 7.6.1 software.

3 Results

3.1 Type I error rate and true positive rate

We generated simulated data under the null hypothesis by directly sampling tumor specimens in TCGA. Primarily, equal numbers of tumor specimens were sampled without replacement from the best

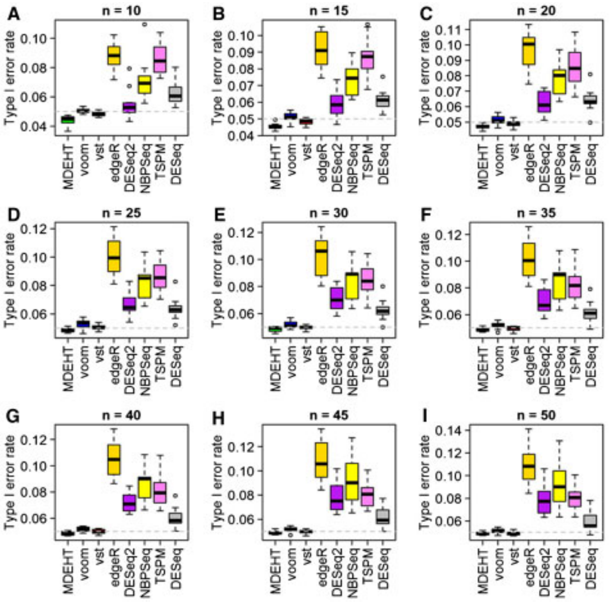

Fig. 2. Type I error rate of different tools for detecting DEmiRs. Simulated data were generated under the null hypothesis by directly sampling tumor specimens from 11 types of cancer datasets in TCGA. Briefly, equal numbers of tumor specimens were sampled without replacement from the best homogeneous samples in each cancer type, forming the treatment and control groups. For each dataset, the Type I error rate was calculated over 100 replicates under different sample sizes

homogeneous samples in each cancer type, forming the treatment and control groups (Fig. 1A). All the statistical methods for miRNA differential expression were performed on these simulated datasets. Type I error rate was calculated over 100 replicates among 11 cancer types (Fig. 2). Our new MDEHT method consistently produced a smaller Type I error rate than the other methods, ~5%, under different sample sizes. MDEHT is slightly conservative when the sample size is small. Both methods based on the Limma statistical package (i.e. Voom and Vst) also yielded a small Type I error rate close to the expected level of 5%. DESeq and DESeq2 slightly inflated the Type I error rate in all scenarios; whereas the other three methods (edgeR, NBPSeq and TSPM) substantially inflated Type I error rate. In most cases, the Type I error rate of the three methods achieved 8–10%, nearly two times higher than the expected level. These simulation results demonstrated that our proposed MDEHT has a better performance in controlling Type I error rate and thus is generally less prone to false positives than the other methods.

Next, we generated simulated data under the alternative hypothesis by directly sampling tumor specimens in TCGA. Equal numbers of tumor specimens were respectively sampled without replacement from two distinct clusters, each of which represents the best homogeneous samples from one of the two tumor types, forming the treatment and control groups (Fig. 1A). The true positive rate was calculated over 100 replicates among any of the two cancer types (Fig. 3). As expected, among all the methods for miRNA differential expression, the true positive rate was increased with increasing sample sizes. In most scenarios, our MDEHT method gave the highest true positive rates under different sample sizes. Voom, Vst, edgeR, DESeq2 and TSPM yielded similar true positive rates, but much lower than the MDEHT. NBPSeq and DESeq performed worst and yielded the lowest true positive rates under different sample sizes. These simulation results demonstrated that the MDEHT method has higher statistical power than the other methods commonly used for detecting miRNA differential expression.

3.2 Reproducibility

Tumor samples were randomly divided into two groups with equal sample size in each cancer type to evaluate the reproducibility of each statistical method for detecting DEmiRs. Samples from each tumor group were compared with normal samples from the same cancer types for identifying DEmiRs (false discovery rate < 0.01; Fig. 1B). The Jaccard similarity index was used to measure the similarity between two sets of DEmiRs detected from the same set of cancer datasets. The Jaccard similarity index was calculated over 50 replicates among 11 cancer types (Fig. 4A–K). Our MDEHT method generally yielded a higher Jaccard similarity index than other methods in most of TCGA datasets and had a much smaller variability in the Jaccard index in each TCGA dataset. In rare scenarios, Voom and Vst had a slightly higher Jaccard index than the MDEHT in UCEC and lung adenocarcinoma, respectively; but these differences are negligible, <1%. TSPM, NBPSeq and DESeq often gave the lowest Jaccard index among all TCGA datasets. According to Jaccard similarity index, our MDEHT ranked first in 9 out of 11 cancer types and ranked second in the other two cancer types. On the other hand, the MDEHT performed best based on the averaged Jaccard index across all types of datasets (Fig. 4L). These data suggested that the MDEHT method generates more reproducible results from miRNA differential expression analysis than those from other methods.

3.3 Identification of DEmiRs in real data datasets

Besides the above simulation studies, we also applied these statistical methods to the analysis of DEmiRs in 11 cancer datasets from TCGA (Fig. 1B). The MDEHT method identified the most extensive list of DEmiRs than other methods in most of these cancer types, whereas the NBPSeq method identified the smallest list of DEmiRs in nearly all cancer types (Supplementary Fig. S3). Next, we took a closer look at the overlapping DEmiRs detected by multiple methods

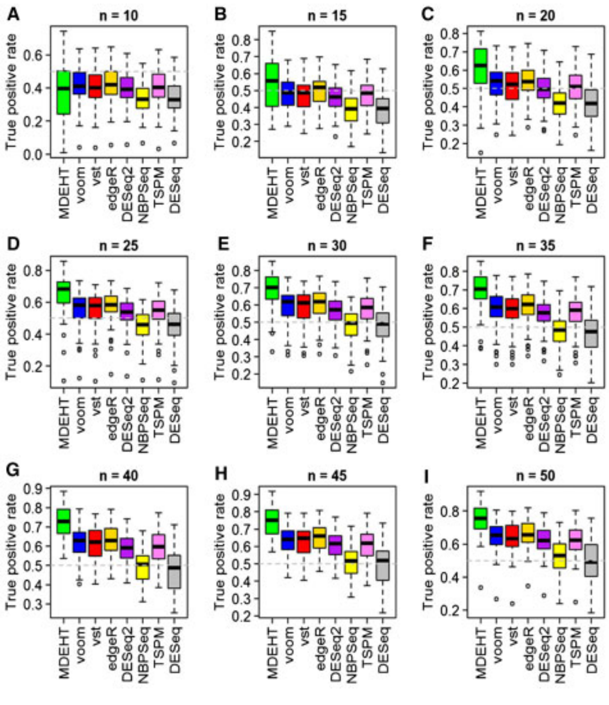

Fig. 3. True positive rate of different tools for detecting DEmiRs. Simulated data were generated under the alternative hypothesis by directly sampling tumor specimens in any of the two combined TCGA datasets. Briefly, equal numbers of tumor specimens were respectively sampled without replacement from two distinct clusters, each of which represents the best homogeneous samples from one of the two tumor types, forming the treatment and control groups. For each combination of any of the two cancer types, the true positive rate was calculated over 100 replicates under different sample sizes

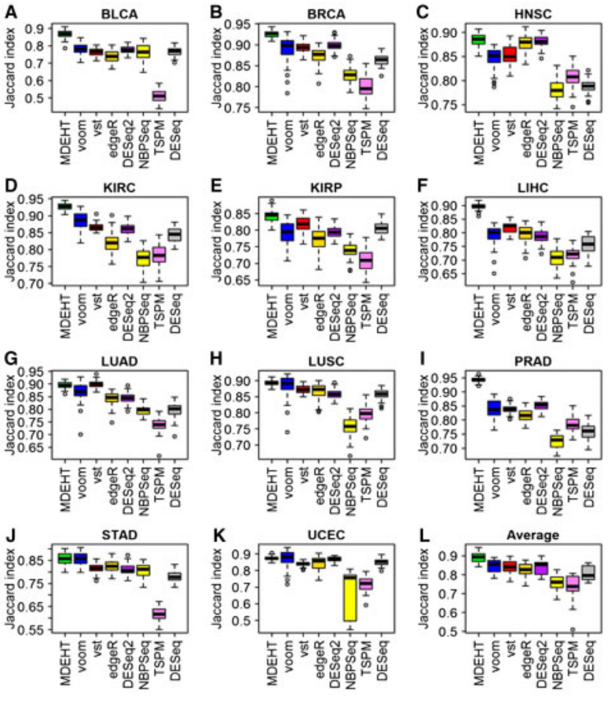

Fig. 4. Jaccard similarity index of different tools for detecting DEmiRs. The Jaccard similarity index measures the similarity between two sets of DEmiRs detected from the same cancer datasets. In each cancer type, tumor samples were randomly divided into two groups with equal sample size. Samples from each tumor group were compared with normal samples from the same cancer types for identifying DEmiRs using these statistical methods. The Jaccard similarity index was calculated over 50 replicates in each cancer type. Higher Jaccard similarity index indicates better reproducibility of statistical methods for detecting DEmiRs

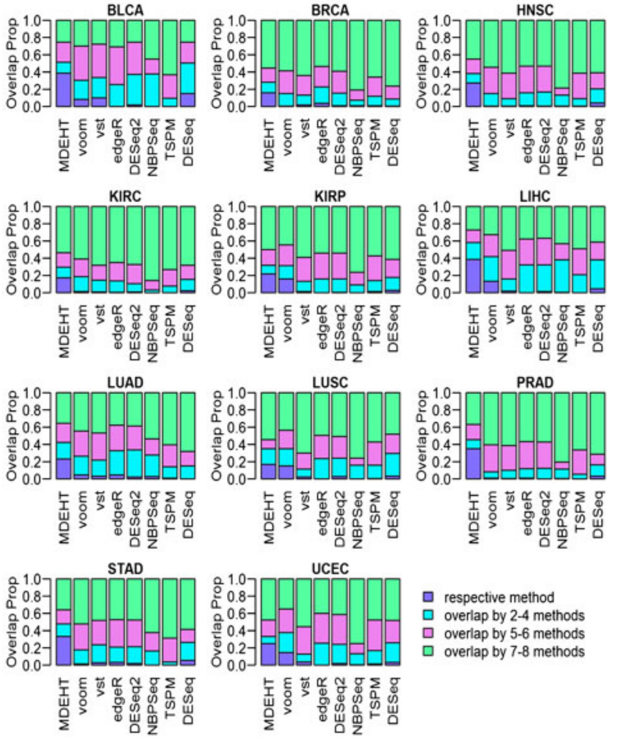

Fig. 5. Proportions of detected DEmiRs for each method by consensus among methods. DEmiRs were divided into four categories based on their overlap among different methods: uniquely detected by one method, overlapped by two to four methods, overlapped by five to six methods and overlapped by seven to eight methods

in the same datasets. These DEmiRs were classified into four categories: DEmiRs uniquely detected by a method, DEmiRs overlapped by two to four methods, overlapped by five to six methods and overlapped by seven to eight methods. The proportion of these overlapped DEmiRs was shown in Figure 5. Overall, the MDEHT method detected a more significant proportion of unique DEmiRs that are not identified by other methods in TCGA cancer datasets, which may also explain why the MDEHT method detected the most extensive list of DEmiRs in most cancer types.

3.4 Functional enrichment analysis of novel DEmiRs

From the list of DEmiRs uniquely detected by the MDEHT method, 59 of these DEmiRs were detected in at least five cancer types (Supplementary Fig. S4 and Supplementary Table S2). The subsequent analysis is focused on these 59 novel miRNAs to find out their functionality. First, we collected clinical data of patients' samples from TCGA and implemented univariate Cox proportional regression analysis to assess the association of these novel miRNAs with the patients' overall survival (Supplementary Table S3). As a result, 14 miRNAs were significantly associated with clinical outcome of patients in at least three types of cancer (Supplementary Table S4). Then, we performed an enrichment analysis of targets that are regulated by these survival-associated miRNAs using the DIANA online tools (Vlachos *et al.*, 2015). A total of 682 genes were predicted to be targeted by these 14 survival-associated miRNAs. There are 39 significant pathways involved in these survival-associated miRNAs, many of which are cancer-related pathways such as mitogen-activated protein kinase (MAPK), Wnt, PI3K-AKT, mTOR and TGF-beta signaling (Supplementary Fig. S5). It is worth noting that many pathways enriched for miR-335-3p targets are the most significant among all the detected pathways. These data suggested that these novel DEmiRs that were uniquely detected by the MDEHT method are potentially involved in tumor development and progression.

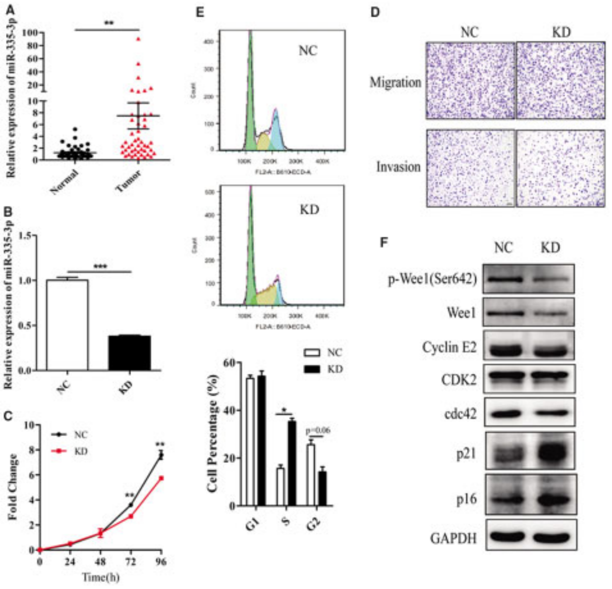

Fig. 6. miR-335-3p promotes cell proliferation by affecting cell cycle in UCEC. (**A**) qRT-PCR analyses of miR-335-3p expression level in 47 UCEC tumor tissues and their matched adjacent non-tumor tissues. (**B**) Relative expression changes of miR-335-3p in HEC-1-B cells transfected with miR-335-3p inhibitor or NC. (**C**) Cell proliferation in HEC-1-B cells transfected with miR-335-3p inhibitor or NC assessed with CCK-8 assay. (**D**) Migration and invasion assays following knock-down of miR-335-3p in HEC-1-B cells. (**E**) Cell cycle distribution in miR-335-3p knockdown HEC-1-B cells detected with flow cytometry. (**F**) Western blot analyses of protein makers related to the cell cycle in HEC-1-B cells transfected with miR-335-3p inhibitor or NC. GAPDH was used as control. Error bars represent the standard deviation of three independent experiments. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ using a two-sided Student's *t*-test

3.5 Experimental validation of a novel DEmiR

miR-335-3p is one of the survival-associated miRNAs and has been rarely studied. miR-335-3p is upregulated in UCEC, LUSC and KIRC tumor tissues. Its upregulation is highly predictive of poor prognosis in cancer patients (Supplementary Fig. S6 and Supplementary Table S4). Therefore, we also performed *in vitro* cell-based assays for this novel miRNA in UCEC for further validation of MDEHT. We first performed qRT-PCR to examine the miR-335-3p expression in 47 cases of UCEC tissues and their matched adjacent non-tumor tissues. Our qRT-PCR data showed that the miR-335-3p levels were significantly upregulated in UCEC tumor tissues in comparison to their adjacent normal tissues (Fig. 6A). To further investigate the potential role of miR-335-3p in UCEC, HEC-1-B cells were transfected with a miR-335-3p inhibitor to knock down its expression, and the expression level of miR-335-3p was decreased after the transfection when compared with the NC group (Fig. 6B). As a result, the knockdown of miR-335-3p significantly suppressed HEC-1-B cell proliferation (Fig. 6C), whereas the ability of cell migration and invasion was barely affected (Fig. 6D), suggesting that the cancer-related roles of miR-335-3p might be partly restricted to the regulation of cancer cell proliferation. As cell cycle regulation is essential for cell growth, we further examined the cell cycle distribution using flow cytometry. As shown in Figure 6E, downregulation of miR-335-3p caused a significant cell cycle arrest at the S stage. Meanwhile, the expression levels of cell cycle-related proteins, such as phospho-Wee1 (Ser642), Wee1, Cyclin E2, cdc42, p21 and p16, were significantly affected when miR-335-3p was knockdown (Fig. 6F). Our data demonstrated that miR-335-3p could play a role in promoting cell proliferation via enhancing cell cycle progression in UCEC.

4 Discussion

isomiRs are produced from the same arm as the archetype miRNA with a few nucleotides different at 5 and/or 3 termini. These well-conserved isomiRs are of functional importance and have contributed to the evolution of miRNA genes (Tan *et al.*, 2014). However, very few methods consider isomiR variations in the analysis of miRNA differential expression (Anders and Huber, 2010; Auer and Doerge, 2011; Di *et al.*, 2011; Love *et al.*, 2014; Robinson *et al.*, 2010; Smyth, 2004). In most analyses, isoforms of a miRNA are usually not distinguished and read counts of isomiRs are combined into one miRNA. Therefore, there is a pressing need to develop new methods that account for the multidimensional structure of isomiRs in the differential expression analysis.

In this study, we proposed a new statistical framework, MDEHT, for analyzing isomiRs data. The utilization of the information hidden in isomiRs enables MDEHT to increase the power of identifying DEmiRs that are not marginally detectable in univariate testing methods. We conducted rigorous and unbiased comparisons of MDEHT with seven commonly used tools in both simulated and real datasets from TCGA. To closely recapitulate the complex structure of real RNA-seq data, we generated simulation data by directly sampling miRNA read count data from TCGA rather than from assumed probabilistic distributions such as Poisson distribution and negative binomial distribution. These comparisons revealed that the MDEHT method performs much better in controlling the Type I error rate, substantially increases statistical power and generally yields higher reproducible results in differential expression analysis of miRNAs. It is worth noting that the main usage of MDEHT is to detect miRNAs that are differentially expressed between different biological or disease conditions. Of course, once a DEmiR is identified by MDEHT, subsequent differential expression analysis can be applied to its individual isomiRs. Focusing on DEmiRs detected by MDEHT, multiple testing problems are much less serious than univariate testing of all individual isomiRs.

In real data analysis, the MDEHT method identified the most extensive list of DEmiRs than other methods in most cancer types. MDEHT tended to identify more unique DEmiRs than other methods in most of the 11 TCGA datasets. Among these unique DEmiRs, 59 were detected in at least 5 cancer types. Subsequent survival analysis of these 59 DEmiRs revealed that 14 were associated with clinical outcome of patients in at least 3 cancer types. Targets of these novel DEmiRs are significantly enriched in many cancer-related pathways, implying their important regulatory roles in cancer development and progression. Furthermore, we also performed *in vitro* cell-based assays for the novel miRNA miR-335-3p in UCEC in further validation of MDEHT. MiR-335-3p has rarely been studied, and its role in cancer remains elusive. For the first time, our preliminary data demonstrated that miR-335-3p play a role in promoting cell proliferation via enhancing cell cycle progression in UCEC. Further investigations are required to elucidate the molecular mechanisms of this novel miRNA in regulating cell cycle and the potential role of its isomiRs in cervical carcinogenesis and progression.

Several caveats in MDEHT should be mentioned. First, the MDEHT method cannot handle intricate experimental designs. Only a two-group differential test is considered in MDEHT. Further investigations are required to extend our current framework to accommodate for simultaneously comparing means for multiple dependent variables across two or more groups. Second, the small sample size may not warrant multivariate normal distribution of the data, in which permutation tests using the Hoteling's T^2 statistic could be necessary for the MDEHT method.

In summary, the newly developed MDEHT method accounts for the multidimensional structure of isomiRs in the differential expression analysis of miRNA. Our comprehensive evaluations demonstrated that the MDEHT method was robust in various datasets and outperformed the other commonly used tools in terms of Type I error rate, true positive rate and reproducibility.

Acknowledgements

We thank TCGA for providing access to small RNA-seq and patients' clinical dataset. We thank the Bioinformatics Core Facility at Zhejiang University School of Medicine for providing computing capacity and anonymous reviewers for reading and commenting on the article.

Author's contribution

Y.L., P.L and W.L. designed the study. M.A. developed the algorithm and performed the data analysis. X.S. developed the algorithm to extract isomiRs data from BAM files. M.Y., A.L. and W.W. conducted *in vitro* cell assays. L.H. and L.Z. collected clinical samples and provided technical support for high performance computation. Y.L., P.L. and M.A. wrote the article. All of the authors discussed and commented the study.

Funding

This work was supported in part by National Key Research and Development Program of China [2019YFC1315700 and 2016YFA0501800], the Key Program of Zhejiang Provincial Natural Science Foundation of China [LZ20H160001], Medical Health Science and Technology Key Project of Zhejiang Provincial Health Commission [WKJ-ZJ-2007 and 2017211914], and National Natural Science Foundation of China [81772766, 81871864 and 81572256].

Conflict of Interest: none declared.

References

Acunzo,M. *et al.* (2015) MicroRNA and cancer–a brief overview. *Adv. Biol. Regul.*, **57**, 1–9.

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Andrew,A.S. *et al.* (2019) MicroRNA dysregulation and non-muscle-invasive bladder cancer prognosis. *Cancer Epidemiol. Biomarkers Prev.*, **28**, 782–788.

Auer,P.L. and Doerge,R. (2011) A two-stage Poisson model for testing RNA-seq data. *Stat. Appl. Genet. Mol.*, **10**, 26.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 12.

Chu,C. *et al.* (2015) deGPS is a powerful tool for detecting differential expression in RNA-sequencing studies. *BMC Genomics*, **16**, 455.

de Rie,D. *et al.* (2017) An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.*, **35**, 872–878.

Di,Y. *et al.* (2011) The NBP negative binomial model for assessing differential gene expression from RNA-seq. *Stat. Appl. Genet. Mol.*, **10**, 28.

Friedman,R.C. *et al.* (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.

Ha,M. and Kim,V.N. (2014) Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.*, **15**, 509–524.

Hanna,J. *et al.* (2019) The potential for microRNA therapeutics and clinical research. *Front. Genet.*, **10**, 478.

Hotelling,H. (1931) The generalization of student's ratio. *Ann. Math. Statist*, **2**, 360–378.

Lagos-Quintana,M. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.

Levandowsky,M. and Winter,D. (1971) Distance between sets. *Nature*, **234**, 34–35.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Lu,Y. *et al.* (2005) Hotelling's T^2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics,* **21**, 3105–3113.

Macfarlane,L.A. and Murphy,P.R. (2010) MicroRNA: biogenesis, function and role in cancer. *Curr. Genomics*, **11**, 537–561.

Morin,R.D. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, **18**, 610–621.

Neilsen,C.T. *et al.* (2012) isomiRs–the overlooked repertoire in the dynamic microRNAome. *Trends Genet.*, **28**, 544–549.

O'Brien,J. *et al.* (2018) Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Front. Endocrinol. (Lausanne)*, **9**, 402.

Penrose,R. (1955) A generalized inverse for matrices. *Math. Proc. Camb. Philos. Soc.*, **51**, 406–413.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.

Tan,G.C. *et al.* (2014) 5' isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res.*, **42**, 9424–9435.

Telonis,A.G. *et al.* (2017) Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. *Nucleic Acids Res.*, **45**, 2973–2985.

Vlachos,I.S. *et al.* (2015) DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res.*, **43**, W460–W466.

Wang,X. *et al.* (2019) Identification of prognostic markers for hepatocellular carcinoma based on miRNA expression profiles. *Life Sci.*, **232**, 116596.

Yan,H.Z. *et al.* (2019) The expression and clinical significance of miRNA-99a and miRNA-224 in non-small cell lung cancer. *Eur. Rev. Med. Pharmacol. Sci.*, **23**, 1545–1552.

Yan,X. *et al.* (2015) Comprehensive genomic characterization of long non-coding RNAs across human cancers. *Cancer Cell*, **28**, 529–540.