Published online 26 May 2016

Nucleic Acids Research, 2016, Vol. 44, No. 14 ***e123***
doi: 10.1093/nar/gkw471

Accurate detection for a wide range of mutation and editing sites of microRNAs from small RNA high-throughput sequencing profiles

Yun Zheng[1,2,*], Bo Ji[1], Renhua Song[3], Shengpeng Wang[1], Ting Li[1], Xiaotuo Zhang[2], Kun Chen[1], Tianqing Li[4] and Jinyan Li[3,*]

[1]Faculty of Life Science and Technology, Kunming University of Science and Technology Kunming, Yunnan 650500, China, [2]Faculty of Information Engineering and Automation, Kunming University of Science and Technology Kunming, Yunnan 650500, China, [3]Advanced Analytics Institute & Centre for Health Technologies, Faculty of Engineering & IT University of Technology Sydney, Australia and [4]Yunnan Key Lab of Primate Biomedicine Research; Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming, Yunnan 650500, China

Received September 30, 2015; Revised May 09, 2016; Accepted May 13, 2016

ABSTRACT

Various types of mutation and editing (M/E) events in microRNAs (miRNAs) can change the stabilities of pre-miRNAs and/or complementarities between miRNAs and their targets. Small RNA (sRNA) high-throughput sequencing (HTS) profiles can contain many mutated and edited miRNAs. Systematic detection of miRNA mutation and editing sites from the huge volume of sRNA HTS profiles is computationally difficult, as high sensitivity and low false positive rate (FPR) are both required. We propose a novel method (named MiRME) for an accurate and fast detection of miRNA M/E sites using a progressive sequence alignment approach which refines sensitivity and improves FPR step-by-step. From 70 sRNA HTS profiles with over 1.3 billion reads, MiRME has detected thousands of statistically significant M/E sites, including 3′-editing sites, 57 A-to-I editing sites (of which 32 are novel), as well as some putative non-canonical editing sites. We demonstrated that a few non-canonical editing sites were not resulted from mutations in genome by integrating the analysis of genome HTS profiles of two human cell lines, suggesting the existence of new editing types to further diversify the functions of miRNAs. Compared with six existing studies or methods, MiRME has shown much superior performance for the identification and visualization of the M/E sites of miRNAs from the ever-increasing sRNA HTS profiles.

INTRODUCTION

MiRNAs can be edited in multiple ways during their biogenesis processes (1–14). An intensively studied editing type is the Adenosine-to-Inosine (A-to-I) editing, which is induced by adenosine deaminase (ADAR) on the double-stranded RNAs (1,12) to convert an adenosine residue into an inosine residue (5,11,12). Inosine residue converted from adenosine in RNA is read as guanosine during reverse transcription for RNA-seq (4,5,11,12,15). A-to-I editing can affect the biogenesis of miRNAs (16–18), and it can also affect the specificity of miRNA target complementarity (19). Another type of editing is the event of adding nucleotides at the 3′ end of mature miRNAs (4,13,20). Generally, uridylation and adenylation induces and prevents the degradation of miRNAs, respectively (21). However, mono-uridylation can increase the expression levels of some miRNAs by facilitating a two-nucleotide overhang for the diver processing (13). Similar to editing, single nucleotide polymorphism (SNP) can affect the function of miRNAs by modulating the transcription of their primary transcripts, processing of pri-miRNAs and pre-miRNAs, maturation, or miRNA-mRNA interactions (22,23). Both the deregulated editing events and the SNPs of miRNAs have been found to lead to severe diseases (24,25).

With the advanced high-throughput sequencing (HTS) technologies, the whole transcriptomes of small RNAs (sRNAs) have become easily available. The huge number of reads from the sRNA HTS profiles often contain miRNAs that are different from their DNA templates, caused either by editing on RNAs or by mutations in DNAs. Research teams have started recently to explore sRNA HTS profiles for the detection of miRNA editing sites (4,7,8,10–13,18,20,26–29). A serious problem when aligning sRNAs to genome with allowance of mismatches is the cross-

*To whom correspondence should be addressed. Tel: +86 871 65918047; Fax: +86 871 65920570; Email: zhengyun5488@gmail.com
Correspondence may also be addressed Jinyan Li. Tel: +61 2 95149264; Email: jinyan.li@uts.edu.au

mapping problem (6) that may bring many false positive predictions. Some researches proposed some solutions to solve the problem. For examples, Alon *et al*. (11) and Gong *et al*., (27) required reads with unique best hits, i.e. reads cannot be aligned to other places in the genome with the same number of mismatches. However, this requirement is too stringent and omits some edited reads from paralog miRNAs, such as hsa-let-7a-1/-2/-3. On the other hand, it is inefficient to align millions of sequencing reads in sRNA HTS profiles to genome in the computational pipeline of (6). The performances of existing methods are not attractive, demanding new ideas to improve.

We introduce a novel detection method which is accurate and fast for the detection of all types of mutation and editing (M/E) sites of miRNAs from sRNA HTS profiles. Our method is named MiRME (short for detecting **miR**NA **M**utation and **E**diting sites). It also has software components to provide comprehensive analysis on the discovered M/E sites. MiRME is different from the existing approaches at several aspects. First, MiRME has three progressive rounds of sequence alignment steps to reach a high sensitivity without loosing speed. Second, reads mapped to multiple loci in the genome are normalized using the cross-mapping correction method (6) to reduce the number of false positive predictions. Third, MiRME can identify and visualize all types of editing and mutation sites at one system.

We applied MiRME to sRNA HTS profiles of 68 human brain samples and two human cell line samples to evaluate its performance. We successfully re-detected many literature-reported editing sites and found a lot of novel M/E sites. More importantly, by integrating the analysis of genome sequencing profiles of the two human cell lines, we demonstrated that a few non-canonical editing sites were not caused by mutations in genome, suggesting there exist other types of non-3′ end editing in addition to the A-to-I editing in miRNAs. Comprehensive comparisons between MiRME and four existing studies (4,11,26,27) and two methods (28,29) showed that MiRME could identify many novel editing sites from the same data sets and showed much better performance than existing methods. MiRME along with the newly identified M/E sites can serves as a valuable tool and resource to better understand the variations in the small RNA transcriptomes.

MATERIALS AND METHODS

Cell line and sequencing

A human neuroepithelial stem cell line was bred in the NESC medium as reported previously (30). The total DNA of about 10^6 cells were extracted with the Wizard Genomic DNA Purification Kit (Promega) according to the manufacturer's instructions. The integrity of the DNA was checked with Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). The obtained genomic DNA was sequenced using Illumina X Ten sequencer by following the corresponding protocols. The total RNA of about 10^6 cells were extracted with Trizol reagent (Invitrogen) according to the manufacturer's instructions. The integrity of the RNA was checked with an ultraviolet spectrophotometry and 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA). The sRNAs were isolated from the total RNA and were sequenced using Illumina HiSeq4000 sequencer by following the corresponding protocols. The obtained DNA and sRNA sequencing data had been deposited to the NCBI SRA database (http://www.ncbi.nlm.nih.gov/sra) under series accession number SRP068960.

Data sets used

As summarized in Supplementary Table S1, we used 68 published sRNA HTS profiles of 13 different human brain tissues or cell lines, one lymphoblastoid cell line and one neuroepithelial stem cell line to find mutation and editing sites in miRNAs and to evaluate the performance of MiRME. All these data sets were downloaded from the NCBI SRA database. The DNA sequencing profile of lymphoblastoid cell line was downloaded from NCBI SRA database with accession number ERA000005. The unmasked genomic sequence of human (hg19, GRCh37) were downloaded from the UCSC Genome Browser (31). The pre-miRNA sequences and genomic positions in gff3 format were downloaded from the miRBase (release 19) (32).

Preprocessing of small RNA HTS sequencing profiles

Raw reads were filtered to make sure that the first 25 nucleotides of the qualified reads have sequencing scores of 30 or higher. The 3′ adapters were cut for qualified reads. Then, the unique sequences of the remaining reads, i.e. unique reads, were obtained and the counts of unique reads with more than 18 nucleotides were calculated.

The MiRME algorithm

MiRME analyzed the mutation and editing sites in pre-miRNAs by using several inputs, including the sequences and secondary structure of pre-miRNAs (not shown in Figure 1), the alignments of reads to pre-miRNAs generated by BLASTN , the reads mapped to pre-miRNAs, the alignments of reads against genome generated by Bowtie, and the results of the cross-mapping correction method (6). MiRME used a modified Smith–Waterman algorithm to align an sRNA read to a pre-miRNA sequence. Briefly, matched and mismatched nucleotides received rewards of +4 and −3, respectively, in alignment. The affine gap penalty, i.e. the penalty increasing linearly with the length of gap after the initial gap opening penalty, was used for gap opening (−4) and gap extension (−2). The weights of reads were retrieved from the results of the cross-mapping correction method (6) and used to quantity the M/E percentages. Suppose there were m unique sequence covering position i (the ith nucleotide from 5′ end) of the pre-miRNA, then the normalized number of reads mapped to this position, N_i, was calculated with Equation 1

$$N_i = \frac{10^7}{t} \sum_{j}^{m} w_j \times n_j, \quad (1)$$

where t was the total number of reads in the sequencing library, m was the number of unique sequence covered this position, $w$$_j$ was the weight of the jth unique sequence

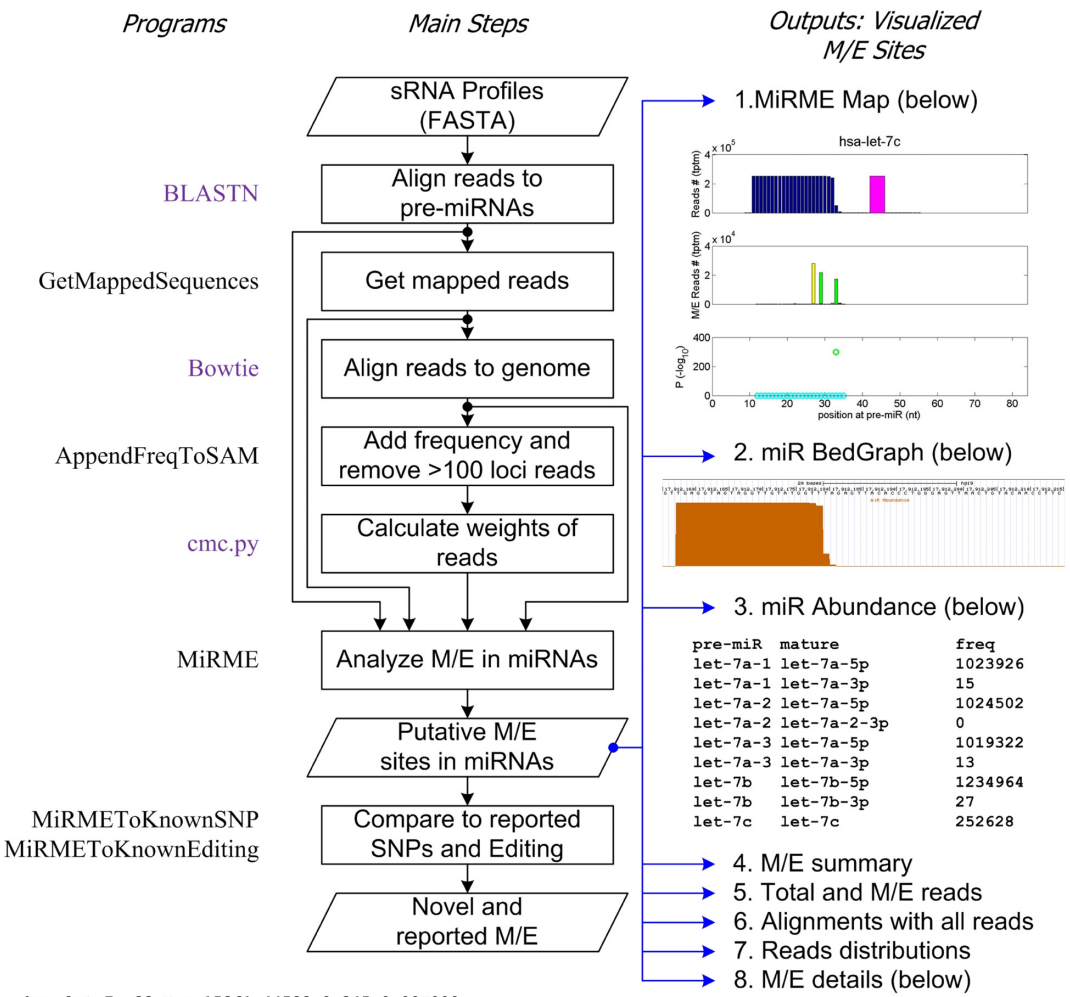

Figure 1. The main steps, corresponding programs and outputs of the MiRME pipeline. The central part lists the steps whose corresponding programs are given on the left. Programs in purple and black are publicly available ones and those developed in this study, respectively. The right and bottom parts pointed by blue lines are outputs of MiRME. Optionally, the pipeline also compares the predicted editing and SNPs in miRNAs to the reported ones to facilitate the discovery of novel editing and/or SNPs in miRNAs.

at the genomic locus, and n_j was the frequency of the jth unique sequence in the library. The number of M/E reads at each position of the pre-miRNA were calculated similar to Equation 1 but only counting the mismatched reads.

MiRME automatically assigned identified M/E sites as one of the following categories: 3′-A, 3′-A before central loop, 3′-U, 3′-U before central loop, 3′-Other, 3′-Other before central loop, A-to-I(G), C-to-U, 5′, Pseudo and Other. If an M/E sites located at −1 to +2 positions of 3′ end of a mature miRNA, it was predicted as a 3′-editing site. If an M/E site located at −1 to −3 positions of 5′ end of a mature miRNA, it was predicted as a 5′-editing site. If a supporting read of a miRNA M/E site had a very small weight (<0.05), calculated by the cross-mapping correction method (6), at the miRNA locus, it was unlikely to be generated from this miRNA, thus was defined as a pseudo edited read. If pseudo edited reads accounted for over 95% of all reads supporting an M/E site, this site was predicted as a Pseudo site. The category of an M/E site was preferentially predicted in the order of Pseudo, A-to-I(G), C-to-U, 3′, 5′ and other if it could be predicted as more than one category.

The computational steps and outputs of the MiRME pipeline

The main steps of the MiRME pipeline were shown in Figure 1. All computational steps of MiRME had been integrated into a whole script whose main program, MiRME, was implemented with the Java programming language. More details of MiRME and its outputs were give in the Supplementary Information. Comprehensive user manual and scripts of the MiRME pipeline, as well as several other auxiliary tools for large-scale analysis, were also given in the Supplementary Information.

P-values of identified mutation and editing sites

The quality of identified mutation and editing sites was evaluated using Equation 2 to excluding the probability of being random sequencing errors.

$$Z = \frac{p_o - p_e}{\sqrt{p_e(1 - p_e)/N}} \quad (2)$$

where p_o was the observed percentage of mutated and/or edited reads, p_e was the expected error rate and N was the number of reads matched to the position of pre-miRNA. Since Z followed a standard normal distribution, P-values of the identified editing or mutation events could be calculated. p_e was related to the score of sequenced nucleotides. For example, a phred score of 20 would lead to a p_e of 1%. Because there could be many mutation and editing sites, the obtained P-values were corrected with the Benjamini–Hochberg correction method (33).

Analyzing selected samples and combining results of different samples

We used the default settings of MiRME (see Supplementary Information) when applying MiRME to the selected sRNA profiles. The criteria used in the analysis include (i) the relative level of editing is at least 5%; (ii) at least 10 reads support the editing event; (iii) the score threshold of sequencing reads is 30; and (iv) a multiple-test corrected P-value of smaller than 0.05. Then, the obtained results of different samples were combined by a separate program in the MiRME package (see Supplementary Information). The identified M/E sites were compared to known SNPs in miRNAs organized in (34) (which was based on the dbSNP v137) and editing sites in miRNAs in the DARNED database (35) and literature (5,10,11,26,27). Finally, the predicted M/E sites that belonged to A-to-I(G), C-to-U and Other were manually examined.

The genome sequencing profile of lymphoblastoid cell line and were aligned to human genome with Bowtie using the following parameters, '-k 10 –best -S -v 1'. The genome sequencing profile of neuroepithelial stem cell line was aligned to human genome with Bowtie2 (36) using the following parameters, '-q –end-to-end -I 0 -X 500 –fr –un unpaired –al aligned –un-conc unconc –al-conc alconc -p 6 –reorder -x'. The obtained SAM files were converted to BAM format and were sorted with samtools (version 1.1) (37). The genome and sRNA sequencing profiles of lymphoblastoid cell line and neuroepithelial stem cell line were loaded into Integrated Genomic Viewer (version 2.3.14) (38) to distinguish the editing sites and SNPs for selected editing/mutation sites.

Target prediction for the original and edited miRNAs

The targets of original and P4 G-to-U edited miRNAs were predicted with the HitSensor algorithm (39). Predicted targets with at least 7 continuous Watson–Crick matches in the seed regions were maintained in the analysis.

GO and pathway analysis for the original and edited miRNAs

The GO term and KEGG pathway enrichment of the targets of the original and edited miRNAs were analyzed with the hypergeometric test (40). The obtained P-values were corrected with the Benjamini–Hochberg correction method (33). Because the GO terms were divided into Biological Process, Cellular Component and Molecular Function, so we conducted enrichment analysis for them, respectively, using the same method.

Comparisons with existing studies for identifying miRNA editing sites

MiRME was applied to the same sRNA HTS profiles used in (11,26). Then, we compared the editing sites predicted in these two studies and the M/E sites predicted by MiRME. We also analyzed two of our selected sRNA HTS profiles, SRR448330 and SRR324686, with two recently published methods Chimira (28) and miTRATA (29), respectively, and compared the results of these two methods with those of MiRME. The results of MiREM were also compared to another two studies (4,27) although these two studies analyzed much more data sets.

The naming of the editing and mutation sites in miRNAs

All identified M/E sites were named by the names of the pre-miRNAs, positions of the sites, the nucleotides from the reference pre-miRNA sequences and the edited/mutated nucleotide at the sites. For example, hsa-mir-376a-1_49_A_g was used to mean an A-to-I editing detected at the position 49 of the hsa-mir-376a-1 precursor, the position of the reference sequence was 'A' and the edited reads had 'g' at this site.

RESULTS

MiRME: A new method to detect a wide range of M/E sites from sRNA HTS profiles

The main steps of MiRME are shown in Figure 1. Briefly, MiRME employs three rounds of progressive sequence alignments to refine the sensitivity and has two important steps to reduce false positive predictions in the systematic detection of all types of M/E events happened to miRNAs. At the first round of sequence alignment, the unique sequencing reads are aligned to pre-miRNAs using BLASTN. This is to achieve a high alignment sensitivity which cannot be achieved by using index-based alignment methods such as Bowtie (36) or SOAP (41) and to avoid low speed incurred by aligning the huge number of unique reads to

the whole genome when using BLASTN. At the second round of sequence alignment, the unique reads mapped to pre-miRNAs are then retrieved from the original sequencing profile and aligned to the genome using Bowtie to check whether they have multiple loci in the genome. After these two rounds of sequence alignments, two steps are followed to reduce false positive predictions. First, the unique reads with too many matched loci (>100) are removed. Second, the remaining unique reads are assessed by the cross-mapping correction method (the cmc.py in Figure 1) (6), which adjusts the weights or percentages of a unique read at each of its genomic loci. Then, at the third round of sequence alignment, the main algorithm MiRME, originated from the Smith–Watermen algorithm, aligns the remaining unique reads to pre-miRNAs again to predict M/E sites. The BLASTN and Bowtie alignment results in previous steps are also used by the major algorithm, MiRME, to achieve fast speed and quantify the M/E levels after the cross-mapping correction, respectively.

Eight different outputs are produced by MiRME to quantify and visualize all the detected M/E sites. Particularly as shown in Figure 1, a three-panel figure, called MiRME map, is used to display all M/E sites in a pre-miRNA. The upper panel reports the total number of reads mapped to each nucleotide of a miRNA precursor, the central panel lists the numbers of M/E reads and the *P*-values of these M/E events are plotted on the lower panel. More details about the parameters and outputs of MiRME are described in Supplementary Information.

Overall summary of the detected M/E sites

MiRME was applied to 70 sRNA HTS profiles (68 brain tissues, 1 lymphoblastoid cell line and 1 neuroepithelial stem cell line), containing more than 1.3 billion raw reads (Supplementary Table S1). From the 68 brain data sets, we detected a total of 45253 M/E sites each supported by at least 1 normalized sequencing read (tags per ten million (TPTM) sequencing reads). Of these, 3214 from 533 pre-miRNAs are significant M/E sites that are supported by at least 10 TPTM and have multiple test corrected *P*-values smaller than 0.05 (Figure 2 and Supplementary Table S2). Of them, 50 M/E sites locate in seed regions, i.e. the first to eighth nucleotide from 5′ end of mature miRNAs (Supplementary Table S2). The largest categories of these 3214 significant editing sites are the 3′-A and 3′-U editing types, consisting of 31.5% and 29.3%, respectively (Figure 2A). 3′-Other, i.e. 3′-C and 3′-G, covers 11.8%. There are 647 or 20.1% special editing sites, where the 5′ ends of mature miRNAs have additional nucleotides, named as 5′-editing. The remaining M/E sites include 57 canonical A-to-I sites, 17 C-to-U sites, 95 Other editing sites, 18 SNPs and 45 Pseudo editing sites (those caused by reads mapped to multiple genomic loci, as defined in Materials and Methods). The A-to-I, C-to-U and 95 other editing sites are further classified as shown in Figure 2B. These results suggest that there indeed exist all the 12 possible editing events due to nucleotide substitutions, and that A-to-I is the largest editing type (Figure 2B). Furthermore, there could be insertion and deletion events in miRNAs as well (see the last row and column of Figure 2B, respectively).

We closely examined the number of significant editing events (except the Pseudo sites) in pre-miRNAs (Figure 2C). It can be seen that some miRNAs can be edited at different positions and can be edited by substitution/addition of different nucleotides during their maturation, but most editing events happened at the 3′ end (the green bars in Figure 2C). Some miRNAs also have a few editing events at 5′ end (the orange bars in Figure 2C). Each miRNA only has 1 or 2 central editing sites in most cases (the blue bars in Figure 2C).

We note that there are several 3′-editing happened at the end of reads mapped to the central regions of pre-miRNAs (Supplementary Figure S1). Since the 3′-editing events have been intensively studied and characterized in the literature (4,7,20,42,43), our detailed analysis is focused on non-3′ editing types and SNPs.

5′-editing sites

We detected more than six hundred 5′-editing sites on 349 pre-miRNAs (Supplementary Table S3). Most of these 5′-editing events happen at the −1 or −2 position of the mature miRNAs (Supplementary Figure S2A). C is the dominant nucleotide (89.3%) in these 5′-editing sites, followed by A and U which constitute 7.3% and 3.4%, respectively (Supplementary Figure S2B). This prevalence differs from the A and U preferences in the 3′-editing events (Supplementary Figure S2C). This means that a cytosine is added to 5′ end of mature miRNA in most 5′-editing events. Some of these 5′ editing sites have significant editing levels in many samples (Supplementary Figure S2D), suggesting that these changes are not random events. However, most of these 5′-editing sites are only detected in a few samples (Supplementary Figure S2E). A typical 5′-editing site is given in Supplementary Figures S2F to S2H. Two sites, the −1 and −2 positions of hsa-let-7f-2-5p, have significant 5′-editing events in one of the superior frontal gyrus of the brain samples (GSM450607).

By making use of genome sequencing profiles, 19 5′-editing sites were further examined to exclude the possibility that the variations were originated from mutations in genome (see the last two columns of Supplementary Table S3 and Figure S3). Three (hsa-mir-21_7_G_c, hsa-mir-26a-1_9_G_c and hsa-mir-26a-2_13_A_c) of these 19 sites are detected in both of the two cell lines with genome sequencing profiles used in this study. For examples, some of reads aligned to hsa-mir-26a-1 have additional cytosines at −1 position of the hsa-miR-26a-5p shown in Supplementary Figure S3A and S3E, which is not caused by mutation in genome (Supplementary Figure S3A). hsa-mir-130a_54_G_a is an event of adding adenosines to 5′ end of hsa-miR-130a-3p (Supplementary Figure S3B and S3F). Two other 5′-editing sites at the −1 positions of hsa-miR-21-5p and hsa-miR-155-5p shown in Supplementary Figure S3C, S3D, S3G and S3H are also adding of cytosines. There are no mutations at genomic positions of these three sites as well (Supplementary Figure S3B to S3D).

A-to-I editing sites

Fifty seven significant A-to-I editing sites have been detected (Supplementary Table S4 and Figure 3A), of which

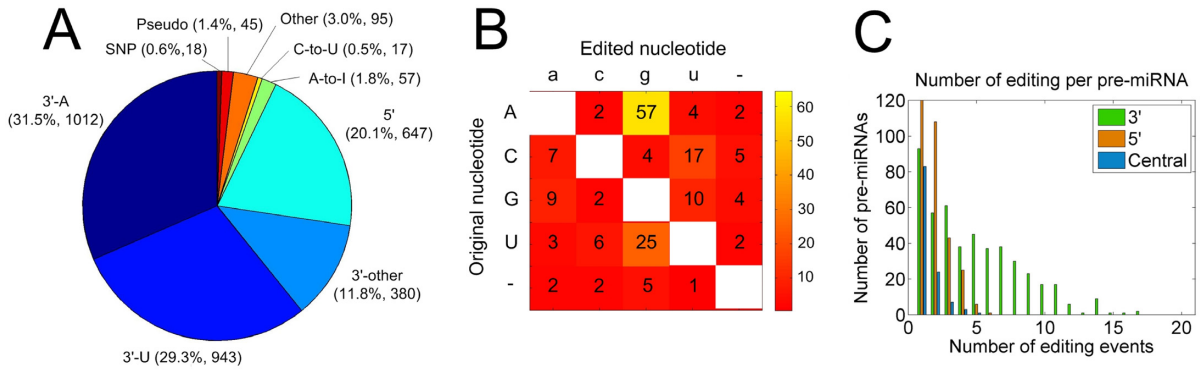

Figure 2. The number of significant M/E sites in miRNAs and their categories in the analyzed sRNA libraries. (**A**) The categories of significant M/E sites in miRNAs. (**B**) The numbers of different types of editing events that do not happen at the 5′ or 3′ end of mature miRNAs. (**C**) The distribution of the numbers of pre-miRNAs with different numbers of 5′-, 3′-editing and Central editing sites, i.e. editing sites that do not happen at the ends of mature miRNAs.

33 are newly identified (marked with a star in Figure 3A). The 57 identified A-to-I editing sites show a weak preference of U and G immediately before and after the editing sites, respectively (Figure 3B), consistent with the UAG motif reported previously (5,11).

Different A-to-I editing sites have different number of samples where A-to-I editing events happen at significant levels (Figure 3C). Five A-to-I editing sites happen in 40 or more samples of the 68 samples. As shown in Figure 3A, three widely reported sites, hsa-mir-376a-1_49_A_g, hsa-mir-376a-2_55_A_g and hsa-mir-376c_48_A_g (5,19), have high editing levels in most samples of embryonal tumor (ET), glioma (GLI), prefrontal cortex of Alzheimer's disease patients (early Alzheimer's disease (EAD) and late Alzheimer's disease (LAD)) and superior frontal gyrus (SFG). As an example, hsa-mir-497_83_A_g (Figures 3D and F) occurs in 24 of the 68 samples and had also been detected in colon tissues (43). On the other hand, 25 sites happen in only five or less samples, likely due to their sporadic appearances in different tissues. For example, hsa-mir-3176_74_A_g (Figures 3E and G) only happens (or showed increased editing levels) in U118A2, a cell line with transfected ADAR2, suggesting that ADAR2 may contribute specifically to A-to-I editing of some miRNAs. It had been postulated previously that ADAR2 can contribute to most of A-to-I editing events (18). Consistent with this, in addition to hsa-mir-3176_74_A_g, six other editing sites (hsa-mir-24-2_18_A_g, hsa-mir-27a_10_A_g, hsa-mir-301a_70_A_g, hsa-mir-378a_58_A_g, hsa-mir-421_61_A_g and hsa-mir-455_32_A_g) have significant levels only in U118A2 and/or U82A2 (see Figure 3A).

Clustered A-to-I editing sites reported in the coding genes of (44) also occur in several miRNAs, including hsa-mir-376a-1, hsa-mir-376a-2, hsa-mir-378a, hsa-mir-381, hsa-mir-497 and hsa-mir-3676.

By integrating the analysis of the genome sequencing profiles in lymphoblastoid cell line, two known (hsa-let-7c_27_A_g and hsa-mir-378c_31_A_g) and one novel (hsa-mir-3609_76_A_g) A-to-I editing sites are proved not to be mutations at their corresponding genomic positions (Supplementary Figure S4A to S4C, S4F). Two novel A-to-I editing sites (hsa-mir-625_58_A_g and hsa-mir-378a_58_A_g shown in Supplementary Figure S4D and S4E, respectively) are also examined in lymphoblastoid cell line to exclude the possibility that these variations are caused by mutations in genome, although their editing levels are insignificant. Similarly, four novel (hsa-mir-181a-2_59_A_g, hsa-mir-181a-1_44_A_g, hsa-mir-381_55_A_g and hsa-mir-130b_71_A_g) and one known (hsa-mir-381_52_A_g) A-to-I editing sites are proved to happen at low levels in neuroepithelial stem cell line and are not mutations in genome (Supplementary Figure S5).

Other types of miRNA editing sites

There potentially exist other types of editing in miRNAs. As shown in Figure 2B, there are 25 U-to-G and 17 C-to-U events (listed in Supplementary Table S5 and Table S6, respectively).

Most of the U-to-G editing events happen in the superior frontal gyrus of brain samples (SFG in Supplementary Figure S6A). There is a clear preference of G immediately before and after the identified U-to-G editing sites (Supplementary Figure S6D). hsa-mir-485_21_U_g is an example of U-to-G editing sites in the SFG samples as shown in Supplementary Figure S5B and S5F. hsa-mir-1260a_22_U_g (Supplementary Figures S6C and S6G) happened at 100% or nearly 100% editing level in the other tissues or cell lines except SFG (Supplementary Figure S6A). After examining the scores of the raw reads that carry hsa-mir-1260a_22_U_g (Supplementary Figure S6E), it is clear that this editing site is not caused by low quality reads or nucleotides. hsa-mir-1260a_22_U_g is also detected in the neuroepithelial stem cell line and will be discussed in the following. Four U-to-G editing sites in hsa-miR-181a-1/-2 had been reported previously (4) (Supplementary Table S5).

MiRME detected 17 putative C-to-U editing sites (Supplementary Figure S7A and Table S6). The −1 position of these C-to-U editing sites has a weak preference to C (Supplementary Figure S7D). Most C-to-U editing sites, including C-to-U editing sites in hsa-mir-125b-1/b-2 (Supplementary Figures S7B and S7F), hsa-mir-219-1/-2 (Supplementary Figures S7C and S7G) and hsa-mir-3653, happened in glioma (GLI in Supplementary Figure S7A) and Alzheimer's disease (EAD and LAD in Supplementary Fig-

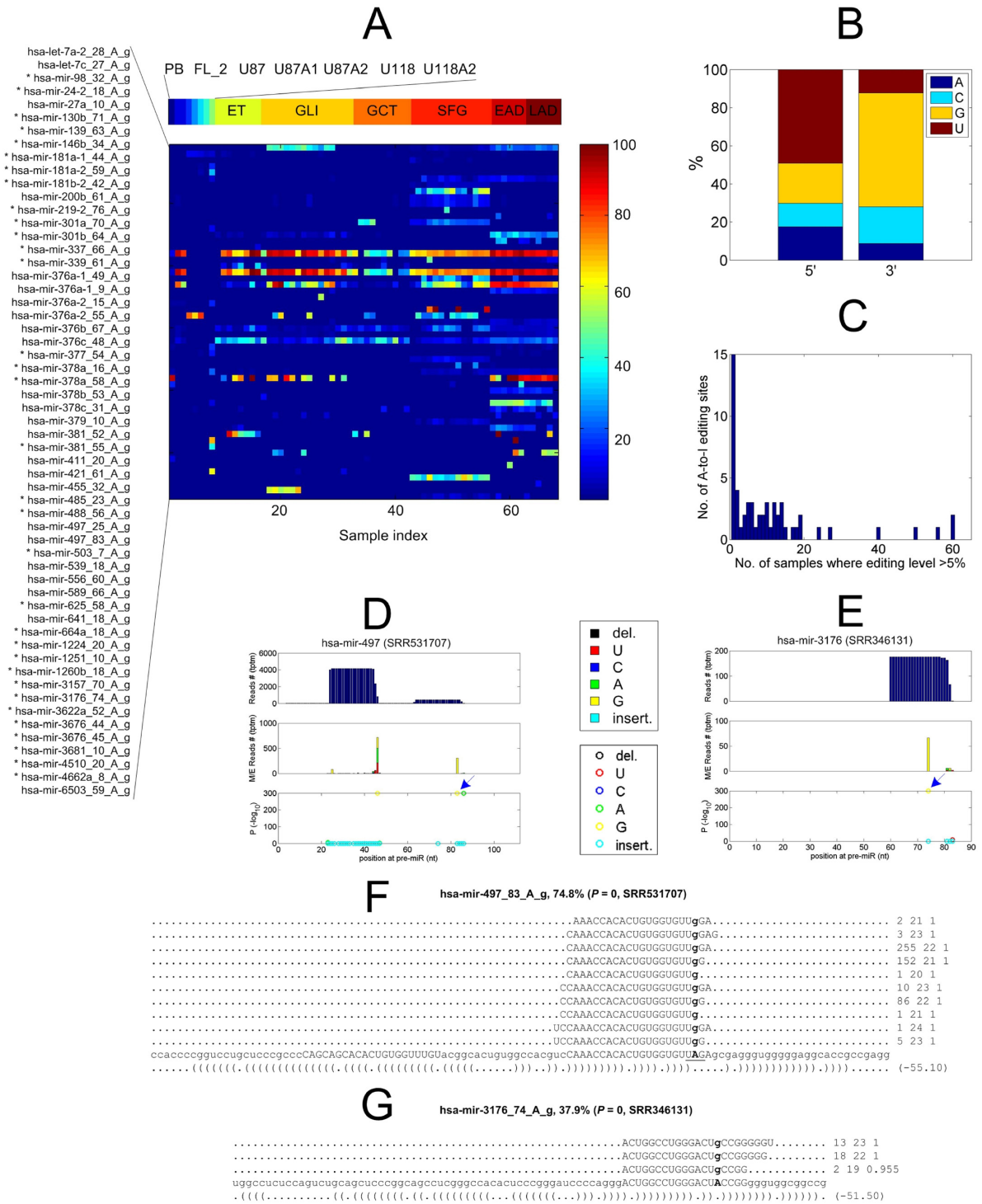

Figure 3. The details of our identified A-to-I editing sites in miRNAs. (**A**) The editing levels of the 57 A-to-I editing sites in the 68 selected brain data sets. The sites marked with stars are newly identified sites. The tissues of the 68 samples are the pooled human brain (PB), frontal lobe (FL_2), brain-originated U87 glioblastoma cell-line (U87), U87 stably transfected with ADAR1 (U87A1), U87 stably transfected with ADAR2 (U87A2), brain-originated U118 glioblastoma cell-line (U118), U118 stably transfected with ADAR2 (U118A2), embryonal tumor (ET), glioma (GLI), germ cell tumor (GCT), superior frontal gyrus of the brain (SFG), prefrontal cortex of early Alzheimer's disease (EAD), and prefrontal cortex of late Alzheimer's disease (LAD). (**B**) The percentages of nucleotides around the 57 A-to-I editing sites. (**C**) The distribution of the number of A-to-I editing sites versus the numbers of samples where editing levels are higher than 5%. (**D**) The MiRME map of hsa-mir-497 in one ET sample (SRR531707). (**E**) The MiRME map of hsa-mir-3176 in U118A2 (SRR346131). (**F**) The details of hsa-mir-497_83_A_g in SRR531707. (**G**) The details of hsa-mir-3176_74_A_g in SRR346131. In Part F and G, the edited nucleotides are shown in bold face.

ure S7A). The scores of the reads supporting hsa-mir-125b-1_25_C_u are shown in Supplementary Figure S7E, indicating this site is not caused by low scored reads. Most C-to-U editing events show modest editing levels (Supplementary Figure S7A) except hsa-mir-491_26_C_u and hsa-mir-128_51_C_u that have very high levels in one glioma sample and one early Alzheimer's disease (EAD) sample, respectively. Two C-to-U editing sites, hsa-mir-125b-1_25_C_u and hsa-mir-125b-2_27_C_u, were also detected in colon tissues (43). Another site, hsa-mir-100_25_C_u, was also reported in (27).

Ten G-to-U editing sites were detected in our selected samples (Supplementary Table S7 and Figure S8). Four of these 10 G-to-U editing sites happen at the fourth position of the seed region in four mature let-7 members and appear in 10 of 16 glioma samples selected (GLI in Supplementary Figure S8A). For example, hsa-let-7a-1_9_G_u shows a level of 5.6% in one of the glioma samples (Supplementary Figure S8B and S8F). The raw reads carrying hsa-let-7a-1_9_G_u have a large variance at the fourth and sixth nucleotides (Supplementary Figure S8C). Another site, hsa-mir-4454_4_G_u appears in the same samples as hsa-let-7a-1_9_G_u, but with higher editing levels (Supplementary Figure S8D and S8G). The raw reads supporting this site do not show enhanced variances at specific sites (Supplementary Figure S8E).

The four G-to-U editing sites at the fourth positions in the four let-7 members severely change the potential targets of the mature miRNAs (Supplementary Table S8 and S9). For example, let-7a-5p and P4 G-to-U edited let-7a-5p share only 47 common targets, but each of them have more than 500 other targets (Supplementary Figure S9A). Consequently, the P4 G-to-U editing events could severely modify the GO terms of these miRNAs (Supplementary Tables S10 to S11). For example, the Molecular Function and Biological Process of let-7a-5p and P4 G-to-U edited let-7a-5p have changed remarkably (Supplementary Figure S9B and S9C). The P4 G-to-U editing sites also severely change the enriched KEGG pathways of let-7 miRNAs. There are several enriched pathways for original let-7 miRNAs (Supplementary Table S12), but the edited let-7 miRNAs have no significantly KEGG pathways.

Thirty seven other types of editing sites are shown in Supplementary Figure S10 and Table S13. For examples, hsa-mir-375_56_G_c and hsa-mir-378f_65_C_g are shown in Supplementary Figure S10B/D and S10C/E, respectively. The reads supporting these two sites have no increased variances at specific sites. hsa-mir-378i_15_A_u is also detected in the lymphoblastoid cell line and will be discussed in the following.

Putative small insertions and deletions in miRNAs

As mentioned early, some miRNAs may have undergone insertions and deletions during their biogenesis (see details in Supplementary Figure S11 and Table S14). Five of the 10 insertions are G insertions and there are more C/G deletions than A/U deletions (Supplementary Figure S11A). hsa-mir-378c_30_-_g (Supplementary Figure S11B and S11D) seems to be a widely existing event (see Supplementary Figure S11A), also detected in colon tissues (43). An example of deletion events was hsa-mir-26a-1_8_C_- (shown in Supplementary Figure S11C and S11E). Small insertions and deletions had been reported in mouse let-7 members (45). In comparison, hsa-let-7c has a significant G-insertion site in a few samples (Supplementary Figure S11A). Our results suggest that there may be small deletions and insertions in other miRNAs.

Detection of known and novel SNPs in miRNAs

We found 18 significant SNP sites from the 68 brain data sets (Supplementary Table S15). These SNP sites exhibit very different levels (Supplementary Figure S12A). Some of these SNP sites, such as hsa-mir-302b_34_G_a, hsa-mir-544b_27_U_g (Supplementary Figures S12B and S12D), hsa-mir-548al_72_A_g, hsa-mir-1304_65_C_a, hsa-mir-3152_57_G_a and hsa-mir-4804_15_C_g, had universal M/E levels of 100% or close to 100% in most samples. As the other SNP sites do not show 100% levels in some of the samples, it is suggested that they are heterozygotic or somatic mutations in the corresponding samples. For example, hsa-mir-627_17_U_g shows a level of only 29.6% (Supplementary Figure S12C and S12E) in one of the 68 data sets. Three SNPs, i.e. hsa-mir-1304_65_C_a, hsa-mir-146a_60_C_g and hsa-mir-627_17_U_g , are verified by using the sRNA and genome sequencing profiles of the lymphoblastoid cell line. The later two cases will be discussed in the following sections and Supplementary Figure S13.

The integrated analysis of sRNA and genome sequencing profiles of the neuroepithelial stem cell line leads to the discovery of 20 novel SNPs in miRNAs (Supplementary Table S16). For examples, hsa-mir-20b_52_A_g and hsa-mir-212_87_C_g are shown in Figure 4A and B, respectively. It is clear that the nucleotide on either sRNA or genome DNA-seq reads are different from the reference genome sequence, indicating these sites are SNPs and have not been reported after comparing them to the latest dbSNP (Figure 4C and D).

Analyzing non-canonical miRNA editing sites by integrating genome sequencing profiles

The availability of genome sequencing profiles for the two selected human cell lines makes it possible to exclude the possibility that the non-canonical editing sites are originated from mutations in genome. Two non-canonical editing sites, hsa-mir-378i_15_A_u and hsa-mir-1260a_22_U_g, have significant editing levels in the lymphoblastoid cell line and neuroepithelial stem cell line, respectively (Figure 5). From Figure 5A and B, it can be seen that the genome sequencing reads carry the same nucleotide as the reference genome sequence, but some of the sRNA sequencing reads have a different nucleotide from the reference genome sequence at the editing sites. Figure 5C and E show that these sites are not false positive predictions, because most obtained sequencing reads are produced from these two miRNAs based on the weights of these reads (the last columns in Figure 5C and E). Figure 5D and F demonstrate that these sites are not exclusively appearing in these two cell lines, but also have high editing levels in many of other selected samples. Finally, these two sites are not reported SNPs (Figure 5G and H).

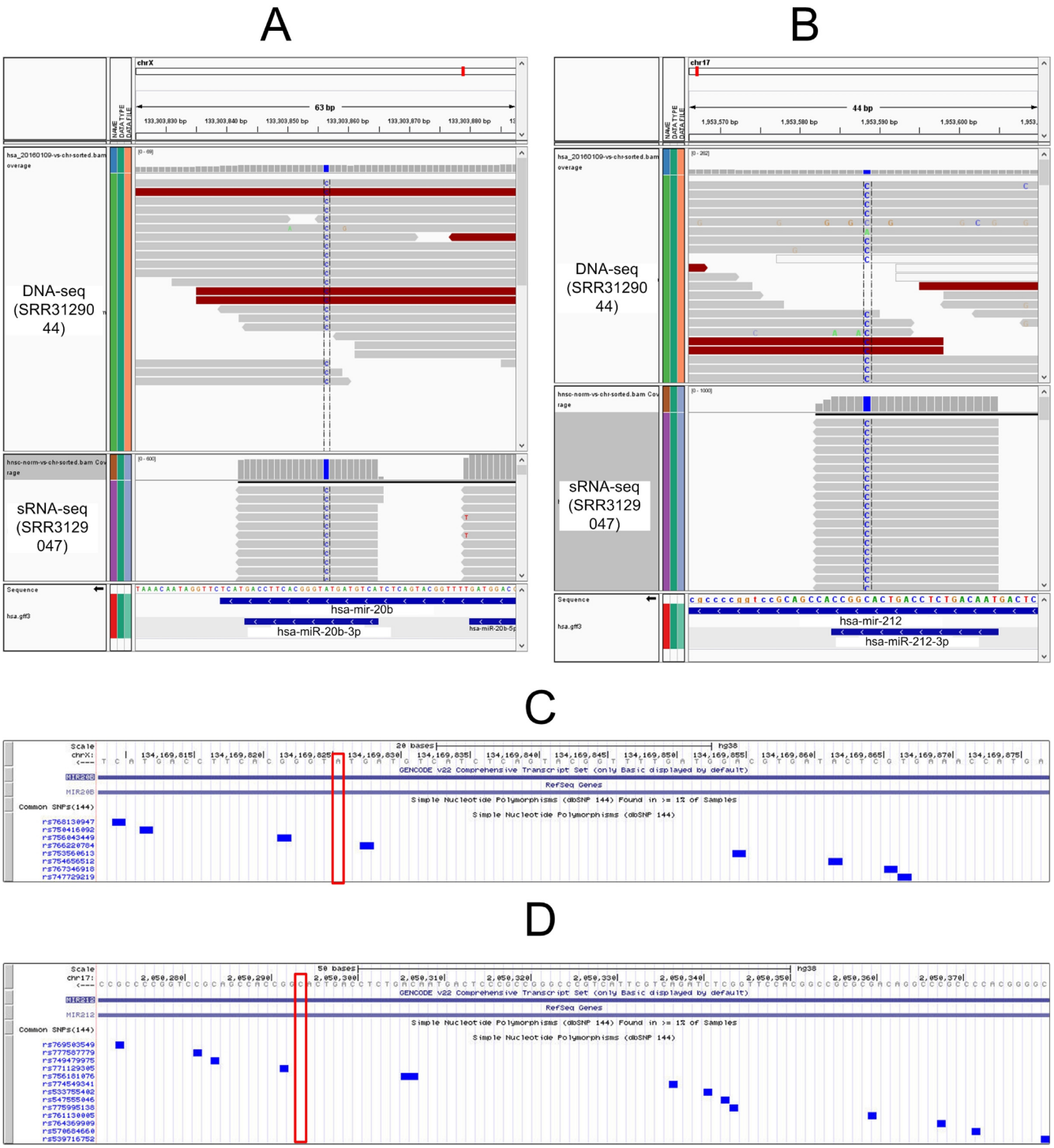

Figure 4. Novel SNPs in miRNAs found by integrating the genome sequencing profile of the neuroepithelial stem cell line. (**A**–**B**) Genome and sRNA sequencing profiles in the regions of hsa-mir-20b and hsa-mir-212, respectively. The nucleotides between the two dot dash lines are the M/E sites, hsa-mir-20b_52_A_g and hsa-mir-212_87_C_g in Part (A) and (B) respectively, reported by MiRME. (C–D) The reported SNPs (dbSNP, v144) in the regions of hsa-mir-20b and hsa-mir-212, respectively. The nucleotides enclosed in the red rectangles in Part (C) and (D) are the M/E sites in Part (A) and (B), respectively.

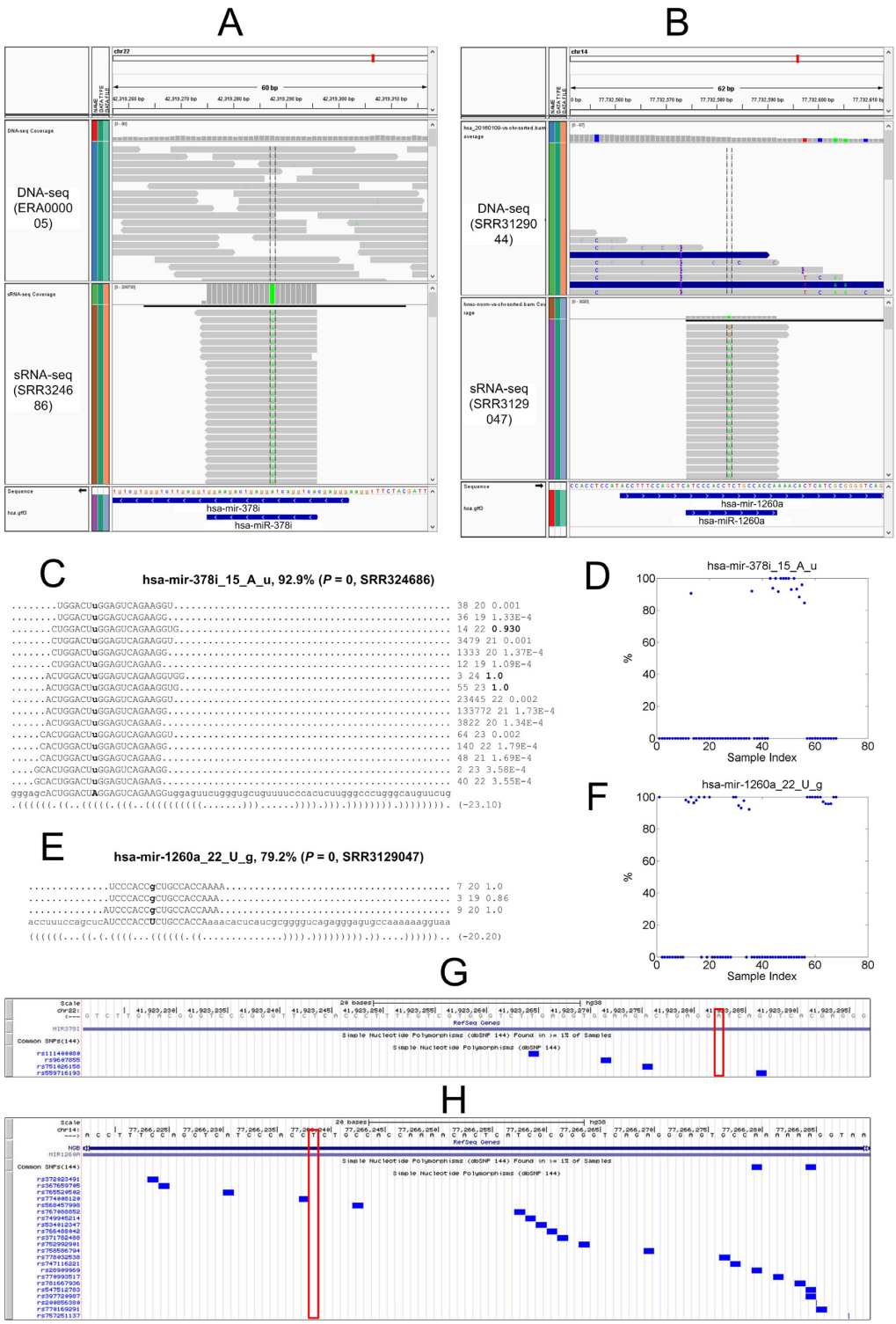

Figure 5. Two non-canonical editing sites verified by integrating the genome sequencing profiles. (**A**–**B**) Genome and sRNA sequencing profiles of the lymphoblastoid cell line and the neuroepithelial stem cell line in the regions of hsa-mir-378i and hsa-mir-1260a, respectively. (**C**) The details of hsa-mir-378i_15_A_u in SRR324686. (**D**) The editing levels of hsa-mir-378i_15_A_u in the 68 selected samples. (**E**) The details of hsa-mir-1260a_22_U_g in SRR3129047. (**F**) The editing levels of hsa-mir-1260a_22_U_g in the 68 selected samples. (**G**–**H**) The reported SNPs (dbSNP, v144) in the regions of hsa-mir-378i and hsa-mir-1260a, respectively. Legends are the same as those in Figure 4.

Our results also indicate that the C-to-U editing do happen in some miRNAs. At least one C-to-U editing site, hsa-mir-93_10_C_u, is detected in the neuroepithelial stem cell line. Except a low scored reads, all other genome sequencing reads are the same as the reference genome at the site of hsa-mir-93_10_C_u. This site is not a reported SNP after checking the latest dbSNP.

Comparisons with related works

MiRME was applied to the same sRNA HTS profiles used by two exiting studies (11,26) to understand whether previously detected miRNA editing sites can be re-detected by MiRME and whether our method could detect more. The result is that 31 of the 35 A-to-I editing sites reported by (11) can be re-detected by our method (Supplementary Table S18 and Figure 6A). Another two sites on miR-376b can be found when using a sequencing score threshold of 20 in MiRME (Supplementary Table S17). Two sites are not produced by MiRME because the supporting reads of one site have two mismatches and the other site has no reads with the editing events (namely, no supporting reads). On the other hand, MiRME detects 12 significant A-to-I editing sites which are not detected in (11) (as shown in Figure 6A and listed in Supplementary Table S18). Furthermore, our results also include >800 3′-editing, >500 5′-editing, 2 C-to-U editing, 5 other editing and 6 SNP sites (Supplementary Table S18).

MiRME re-detects 35 of the 44 editing sites reported in the related work (26) (see Supplementary Table S19 and Figure 6B). MiRME does not report the other 9 sites because they have no supporting reads (4 sites), or the supporting reads are removed due to many low scored nucleotides (2 sites), or the supporting reads are perfectly matched to many other loci (3 sites) (see Supplementary Table S19). MiRME successfully excludes these false positive predictions.

We carefully examined the 35 identified sites and found that these sites actually belong to much diverse categories. Two editing sites (hsa-mir-146a_60_C_g and hsa-mir-627_17_U_g) reported by (26) are actually two SNPs (rs2910164 and rs2620381). The genome sequencing results of the same individual show that some of the genome sequencing reads do carry the expected mutated nucleotides at these two positions (Supplementary Figures S13A and S13B, respectively).

In fact, two editing sites of (26) are pseudo sites, i.e. not real editing sites. The weights of the reads supporting these two sites are very small and they are actually produced from other loci in the genome (Supplementary Figure S14). In Supplementary Figure S14A, the weights of the reads supporting hsa-mir-422a_19_A_g are smaller than 0.001, meaning that they are produced from other loci in the genome. For example, a unique sequence ACUGGACUUGGgGU CAGAAGGC (blue in Supplementary Figure S14A) has a very small weight, with 120 reads, and is actually produced from miR-378a (with a weight of 0.5, Supplementary Figure S14B) and another locus in the genome (chr3:32027799-32027820, minus strand). hsa-mir-378c_32_G_c is a pseudo editing site too because the weights of the edited reads are also very small (Supplementary Figure S14C), <0.01. For example, a unique sequence ACUGGACUUGGAGUCA GAAGAc (blue in Supplementary Figure S14C), with 109 reads, has a perfectly matched locus at chr14:55108399-55108420, minus strand. Therefore, this locus is supposed to produce most of this sRNA read, with a weight of 0.936.

Furthermore, 16 editing sites of (26) are predicted as 3′-editing sites because all or most editing events happened at the 3′ end of their supporting reads (Supplementary Table S19 and Figure S15).

In addition to these 35 editing sites previously discovered by (26), MiRME detects many other significant M/E sites, including more than two hundred 3′-editing, eight 5′-editing, one novel A-to-I (hsa-mir-3609_76_A_g, Supplementary Figure S4C and S4F), four other editing sites and three SNPs from the same data set (see details from Supplementary Table S20).

The editing levels of some editing sites are different from those reported previously (11,26), which might be resulted from different methods to handle the cross mapping problem in these two studies (11,26) (see Supplementary Tables S17 and S19).

We also compared our results with two other related works (4,27) which analyzed much more data sets than this study (see Supplementary Tables S21 and S22). MiRME detects 406 of the 426 editing sites reported by (4) (Figure 6C). The exclusion of 20 sites is attributed to two reasons. First, three sites are not reported because the supporting reads of these sites have low scored nucleotides. Second, the remaining sites occur in different tissues from those used in our study. MiRME finds all the 56 editing sites reported in (27) (Figure 6D). The results of (27) were predicted by method proposed in (11). Thus, these results again show that MiRME has better performance than the method in (11) since MiRME predicts these sites using much smaller number of samples than those used in (27).

MiRME was compared to two methods introduced recently (28,29). Chimira allowed up to 2 mismatches when aligning reads to pre-miRNAs and was designed to identify editing sites at 3′ end, 5′ end, A-to-I editing sites and SNPs in the mature miRNAs (28). miTRATA was designed to find 3′ editing sites (29). We submitted two sRNA profiles, with accession numbers SRR448330 and SRR324686, to Chimira and miTRATA respectively, then compared their results with the results of MiRME for the same data sets (see Figure 6E and F).

As shown in Figure 6E, MiRME and Chimira totally find 6120 and 3544 M/E sites on the selected data set (SRR448330), and 2370 are commonly predicted by both of the two methods. Chimira exclusively predicts 1174 sites that include 102 sites of the newly found miRNAs in the release 21 of miRBase and should not be compared with MiRME's results using miRBase release 19. We find that some of the remaining sites are supported by reads with low scored nucleotides (smaller than 30 used by MiRME). Thus, when we adjust the sequencing threshold of MiRME to 20, MiRME additionally finds 243 M/E sites from the 1174 sites exclusively reported by Chimira (see MiRME_sc20 in Figure 6E). Finally, 829 M/E sites are exclusively reported by Chimira. We examine some of these 829 sites and find they are supported by either reads with more than one mismatch to the corresponding pre-miRNAs, or reads with se-

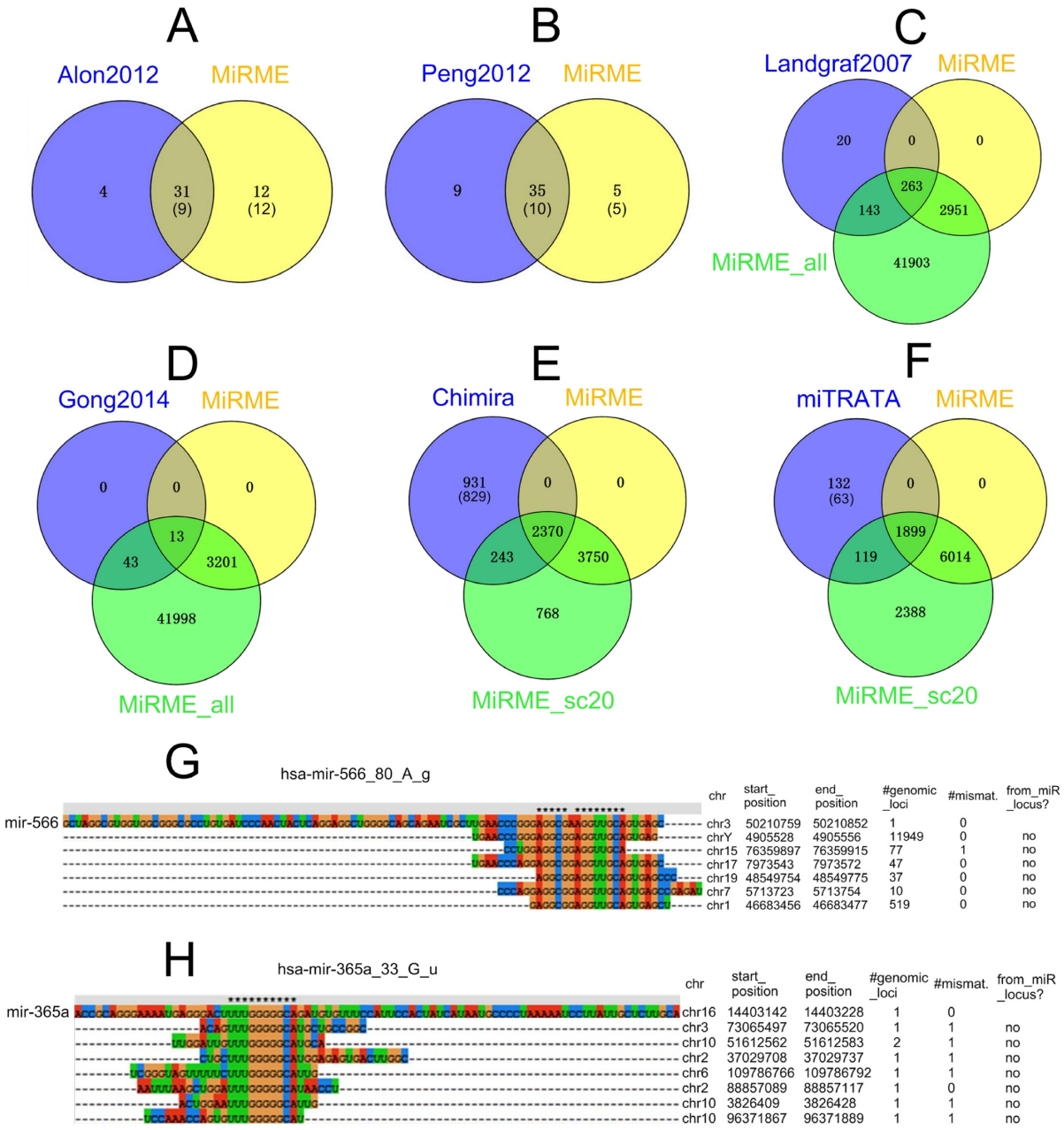

Figure 6. Comparisons between MiRME and six existing studies or methods. (**A**) The numbers of predicted non-3′ editing sites in one existing study (11), marked as Alon2012, and MiRME using the same data sets. (**B**) The numbers of predicted non-3′ editing sites in another existing study (26), marked as Peng2012, and MiRME using the same data sets. In Part A and B, the numbers in parenthesis are the numbers of significant M/E sites. (**C**) The number of predicted editing sites in (4), marked as Landgraf2007, and those predicted by MiRME using the 68 selected data sets. MiRME_all represents the 45253 M/E sites with at least 1 normalized supporting read and 7 M/E sites with >0 normalized supporting reads predicted using the 68 selected brain samples. (**D**) The number of predicted editing sites in (27), marked as Gong2014, and those predicted by MiRME using the 69 selected data sets. MiRME_all represents the 45253 M/E sites with at least 1 normalized supporting read using 68 selected brain samples and 2 M/E sites with >1 normalized supporting read predicted using SRR324686. (**E**) The number of M/E sites predicted by Chimira (28), marked as Chimira, and those predicted by MiRME using one of the selected sRNA profile (SRR448330). (**F**) The number of editing sites predicted by miTRATA (29), marked as miTRATA, and those predicted by MiRME using one of the selected sRNA profile (SRR324686). In Part E and F, the numbers in parenthesis are the number of predictions after removing editing sites on newly identified miRNAs in the release 21 of miRBase. (**G**) The details of hsa-let-7c_17_A_g that is exclusively predicted by Chimira. (**H**) The details of hsa-mir-365a_33_G_u that is exclusively predicted by miTRATA. Reads with scores larger than or equal to 30 were examined in Part G and H. In Part G and H, the columns, '#genomic_loci', '#mismat.', 'from_miR_locus?' mean the number of genomic loci of the read, the number of mismatches between the read and genome at the locus shown in the same line, and whether the read is generated from the miRNA locus, respectively. If a read has more than one genomic locus with the same number of mismatches, one of the genomic loci is shown, but all loci of the read have been checked with the locus of the hsa-let-7c and hsa-mir-365a in Part G and H, respectively.

quencing scores even under 20 because 419 of these 829 sites can be found in the 68 selected brain samples. For example, hsa-mir-566_80_A_g is exclusively reported by Chimira. We search reads without low-scored nucleotides (<30) that carry the editing site, and align them to precursor of hsa-mir-566 with ClustalX (46) and to genome with Bowtie by allowing at most 1 mismatch (shown in Figure 6G). The reads supporting this editing site have many genomic loci with less mismatches (0 or 1) and all these loci are different from the locus of hsa-mir-566. Thus these reads are not generated from hsa-mir-566, indicating this is a false positive prediction. Six other editing sites exclusively reported by Chimira are similar to hsa-mir-566_80_A_g and are false positive predictions as well (see Supplementary Figure S16). At last, Chimira has a size limit of inputs files, which makes Chimira infeasible to analyze the large libraries such as SRR324686 with over 70 million raw reads.

MiRME could find 1899 of the 2150 editing sites reported by miTRATA. The 251 sites exclusively reported by miTRATA include 69 sites for newly identified miRNAs in release 21 of the miRBase. When relaxing the threshold of sequencing score to 20, 119 of these 251 sites can be identified further by MiRME (Figure 6F). We carefully examined six of the 63 remaining sites (Figure 6H and Supplementary Figure S17). As shown in Figure 6H, the supporting reads of hsa-mir-365a_33_G_u have more than one mismatch to hsa-mir-365a and there are genomic loci with less mismatches for these reads. Five more sites examined in Supplementary Figure S17 are similar to hsa-mir-365a_33_G_u. Thus, these sites are false positive predictions.

As shown in Figure 6A, B, E and F, MiRME finds most or all M/E sites reported by other methods and exclusively reports many additional M/E sites using the same data sets, suggesting that MiRME has better sensitivity than these compared methods.

MiRME finds all 3′-A and 3′-U editing sites reported in (7) using the 68 selected data sets. Furthermore, the results of MiRME also indicate that there are 3′-editing in hundreds of other miRNAs (Supplementary Table S2). We examine several 3′-editing sites not reported in (7) by integrating the analysis of genome sequencing profiles (Supplementary Figure S18). Two miRNAs, hsa-mir-132-3p and hsa-mir-127-3p, have both 3′-A nad 3′-U editing (Supplementary Figures S18A/E/F and S18C/H/I, respectively), which are not caused by mutations in genome (Supplementary Figure S18A and S18C). Two other miRNAs, let-7a-3p and hsa-mir-143-3p, show significant 3′-U editing events that are not resulted from mutations in genome as well (Supplementary Figure S18B/G and Supplementary Figure S18D/J, respectively).

Efficiencies of MiRME and two compared methods

We performed our analysis on an HP DL580 server running CentOS 7.0 64 bit operating system. Normally, it takes tens to hundreds of minutes to finish all steps of the MiRME pipeline using one 2.8 Ghz processor. For example, the MiRME pipeline analyzed one embryonal tumor of human brain (SRR531683) consisting of 24 299 190 raw 35 nt sequencing reads with 3′ adapters (see Supplementary Table S1) in 2 h and 5 min. Around half of the run time for this data set was used to remove 3′ adapters in raw reads. If there are no 3′ adapters, the MiRME pipeline is even more efficient. For example, it only took around 16 min to finish all steps of MiRME on one frontal lobe data set (SRR448330) with more than 30 million raw reads without 3′ adapters. In comparison, Chimira used around 4 min to analyze the same data set (SRR448330). miTRATA used several days to analyze the selected sRNA library (SRR324686), but MiRME only used less than 4 h to analyze the same data set.

DISCUSSION

Advantages of using MiRME to detect miRNA mutation and editing sites

MiRME has several advantages over existing methods. First of all, MiRME uses a unique three-round sequence alignment strategy which is critical and necessary to correctly identify false positive predictions. For example in Figure 6G, the unique read on the second line has 1 mismatch to hsa-mir-566, but it has 11 949 genomic loci with 0 mismatches. If this read, as well as other reads in Figure 6G, is not aligned to genome with a second round of alignment, hsa-mir-566_80_A_g will be predicted as a true editing site based on these reads, as done by other method (28). Thus, the second round of alignment is necessary to correctly eliminate this kind of false positive predictions. The third round of alignment is also needed to integrate the results from the first two rounds and to calculate statistics for evaluating the significance of identified M/E sites, as well as to visualize the results. When compared with the several related works (11,26,28,29) using the same data sets, MiRME finds most M/E sites of these studies as well as many other M/E sites. MiRME has comparable efficiency to Chimira (28) and uses much less time than miTRATA (29), for analyzing the same data set. Second, MiRME can remove the false positive predictions raised by the reads mapped to multiple genomic loci by using the cross-mapping correction method (6) (see examples in Supplementary Figure S14). Although the cross-mapping problem was noticed by (6), but the computational pipeline in (6) aligned millions of reads in the sRNA HTS profiles to genome which makes their computational process inefficient. In comparison, our three-round alignment strategy avoids the complex alignments of all reads to genome by only aligning reads that are mapped to pre-miRNAs to the genome. Third, MiRME can identify and visualize all kinds of editing sites and SNPs.

MiRME is easy to use. One command line can finish all the analysis starting from sRNA HTS profiles in SRA or FASTQ format to the final report. Detailed manual and script are available in the Supplementary Information.

Non-canonical editing events

Before this work, some non-canonical editing sites had been reported (4,26,27,43). Here, our results show that three types of editing events, U-to-G, C-to-U and G-to-U, might be biologically relevant because their frequencies are much higher than other types of editing events in our samples selected (Figure 2B). By integrating the genome sequenc-

ing profiles, we verify that the variations for two non-canonical editing sites, hsa-mir-378i_15_A_u and hsa-mir-1260_22_U_g shown in Figure 5, as well as one C-to-U editing site (hsa-mir-93_10_C_u) in the neuroepithelial stem cell line, are not originated from mutations in genome. These editing events may represent unrecognized approaches of miRNAs to diversify their functions by targeting another set of genes, such as shown in Supplementary Table S8 and Figure S9.

There could be alternative mechanisms for some M/E sites

The categories of predicted editing sites are determined based on the most reasonable way of biogenesis. However, there could be other mechanisms to generate the M/E sites.

For example, hsa-mir-26a-1_8_C_- in Supplementary Figure S11E could be originated from 5′ addition events instead of being a deletion event. For another example, hsa-mir-183_49_G_a in Supplementary Table S2 is classified as a 3′-A event with a maximal level of 14.9% in all 68 samples examined (see Supplementary Table S2). Meanwhile, this site also has an SNP, rs41281222. However, this sites is reported as a 3′-A event based on two considerations. First, the editing level is much less than about 100% for homozygotic or 50% for heterozygotic genotypes. Second, the position is at the +1 position of hsa-mir-183-5p. In comparison, hsa-mir-302b_34_G_a is another site at 2 nt downstream of hsa-mir-302b-5p, it overlaps with an SNP, rs190807868 and is reported as an SNP because its high level of 94.7% in one of samples examined (SRR531691) (see Supplementary Figure S12A and Table S16).

The reliability and the repeatability of the predicted M/E sites

Some predicted M/E sites only appear in one or a few samples examined. For example, hsa-mir-3176_74_A_g and six other A-to-I sites only appear in U118A2 and/or U82A2. In practice, the M/E sites that are significant in more biological samples or verified with other approaches are more reliable and suggested for further studies.

Although MiRME uses a very strict sequencing score threshold of 30, the variance of scores should be considered for some special cases. For example, as shown in Supplementary Figure S8C, the score variances of reads that support hsa-let-7a-1_9_G_u in SRR531702 are severely increased at position 4 and 6, suggesting that further studies or experiments are necessary to verify the site.

The 3′-editing may happen to single-strand small RNAs

As shown in Supplementary Figure S1, we find several 3′-editing sites at the ends of reads mapped to the central loops of pre-miRNAs. This raises a question of how this type of editing is realized. Existing studies suggested that 3′-editing can happen at the end of mature miRNAs when a pre-miRNA or miRNA:miRNA* duplex is formed (7). The reads originated from the central loop are byproducts when Dicer cuts the loop end of pre-miRNA to form a miRNA:miRNA* duplex. Because the lengths of loop regions are too small to form hairpins, we thus speculate that the cut-out loop regions could become single-stranded small RNAs, as illustrated by hsa-mir-218-2 and hsa-mir-219-2 in Supplementary Figure S1. The existence of these 3′-editing sites suggests that the 3′-A and 3′-U might happen to some single-stranded sRNAs.

5′-editing is a potentially new type of editing

As shown in Supplementary Figure S2, Figure S3 and Table S3, many miRNAs can potentially have additional nucleotides at the 5′ end which might be another type of editing that has not been carefully studied. To the best of our knowledge, Chimira (28) seems to be the only published method for detecting 5′-editing of miRNAs. One reported A-to-I editing site, hsa-mir-27a_10_A_g, happened at the position 1 of the mature miR-27a and at the 5′ end of the supporting reads. This A-to-I editing is processed before the miRNA:miRNA* duplex is cut out from the hairpin of pre-miRNA. The 647 5′-editing sites might be processed in the same way as hsa-mir-27a_10_A_g is. The other possibility is that these editing events could be performed in the similar way as 3′-editing after the miRNA:miRNA* duplex is cut out from the hairpin of pre-miRNA. Or even after the single-stranded mature miRNA has been separated from the miRNA:miRNA* duplex, because of the possibility of 3′-editing to single-stranded sRNAs as discussed above.

CONCLUSION

MiRME is an effective and efficient computational pipeline for detecting and visualizing editing sites and SNPs in miRNAs. The unique idea is the three-round alignment strategy with a strict control of false positive predictions. Applying MiRME to 70 sRNA HTS profiles of human, we have found some novel canonical A-to-I editing sites, as well as some putative editing sites of other categories resulted from unknown mechanisms. By integrating the genome sequencing profiles, we verified that two non-canonical editing sites, hsa-mir-378i_15_A_u and hsa-mir-1260a_22_U_g, and one C-to-U editing site are not resulted from genomic mutations, and found 20 novel SNPs in miRNAs. MiRME, along with the results in the work, provides new insights into miRNA processing and makes it feasible to analyze miRNA M/E sites from a large number of sRNA HTS profiles.

AVAILABILITY OF DATA AND MATERIALS

The data sets used in the study, as listed in Supplementary Table S1, are publicly available from the NCBI SRA database. The MiRME package, including the main program and other supporting programs, is freely available for non-commercial purposes upon request.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Natural Science Foundation of China [31460295, in part]; Kunming University of Science and Technology

[14078285, 2015HC031, and 10978166 to Y.Z.]; Australian Research Council DP project [ARC DP130102124, to R.S. and J.L.]. Funding for open access charge: National Natural Science Foundation of China [31460295, in part]; Kunming University of Science and Technology [14078285, 2015HC031, and 10978166 to Y.Z.]; Australian Research Council DP project [ARC DP130102124, to R.S. and J.L.].
Conflict of interest statement. None declared.

REFERENCES

1. Bass,B., Nishikura,K., Keller,W., Seeburg,P., Emeson,R., O'connell,M., Samuel,C. and Herbert,A. (1997) A standardized nomenclature for adenosine deaminases that act on RNA. *RNA*, **3**, 947–949.
2. Luciano,D.J., Mirsky,H., Vendetti,N.J. and Maas,S. (2004) RNA editing of a miRNA precursor. *RNA*, **10**, 1174–1177.
3. Blow,M., Grocock,R., Van Dongen,S., Enright,A., Dicks,E., Futreal,P., Wooster,R., Stratton,M. *et al.* (2006) RNA editing of human microRNAs. *Genome Biol.*, **7**, R27.
4. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
5. Kawahara,Y., Megraw,M., Kreider,E., Iizasa,H., Valente,L., Hatzigeorgiou,A. and Nishikura,K. (2008) Frequency and fate of microRNA editing in human brain. *Nucleic Acids Res.*, **36**, 5270–5280.
6. de Hoon,M.J.L., Taft,R.J., Hashimoto,T., Kanamori-Katayama,M., Kawaji,H., Kawano,M., Kishima,M., Lassmann,T., Faulkner,G.J., Mattick,J.S. *et al.* (2010) Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome Res.*, **20**, 257–264.
7. Burroughs,A.M., Ando,Y., deHoon,M.J.L., Tomaru,Y., Nishibu,T., Ukekawa,R., Funakoshi,T., Kurokawa,T., Suzuki,H., Hayashizaki,Y. *et al.* (2010) A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness. *Genome Res.*, **20**, 1398–1410.
8. Guo,L., Yang,Q., Lu,J., Li,H., Ge,Q., Gu,W., Bai,Y. and Lu,Z. (2011) A comprehensive survey of miRNA repertoire and 3′ addition events in the placentas of patients with pre-eclampsia from high-throughput sequencing. *PloS One*, **6**, e21072.
9. Wyman,S., Knouf,E., Parkin,R., Fritz,B., Lin,D., Dennis,L., Krouse,M., Webster,P. and Tewari,M. (2011) Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res.*, **21**, 1450–1461.
10. Mizuguchi,Y., Mishima,T., Yokomuro,S., Arima,Y., Kawahigashi,Y., Shigehara,K., Kanda,T., Yoshida,H., Uchida,E., Tajiri,T. *et al.* (2011) Sequencing and bioinformatics-based analyses of the microRNA transcriptome in Hepatitis B–related hepatocellular carcinoma. *PloS One*, **6**, e15304.
11. Alon,S., Mor,E., Vigneault,F., Church,G.M., Locatelli,F., Galeano,F., Gallo,A., Shomron,N. and Eisenberg,E. (2012) Systematic identification of edited microRNAs in the human brain. *Genome Res.*, **22**, 1533–1540.
12. Ekdahl,Y., Farahani,H., Behm,M., Lagergren,J. and Öhman,M. (2012) A-to-I editing of microRNAs in the mammalian brain increases during development. *Genome Res.*, **22**, 1477–1487.
13. Heo,I., Ha,M., Lim,J., Yoon,M.-J.J., Park,J.-E.E., Kwon,S.C., Chang,H. and Kim,V.N. (2012) Mono-Uridylation of Pre-MicroRNA as a key step in the biogenesis of group II let-7 MicroRNAs. *Cell*, **151**, 521–532.
14. García-López,J., Hourcade,J.d.D. and delMazo,J. (2013) Reprogramming of microRNAs by adenosine-to-inosine editing and the selective elimination of edited microRNA precursors in mouse oocytes and preimplantation embryos. *Nucleic Acids Res.*, **41**, 5483–5493.
15. Park,E., Williams,B., Wold,B.J. and Mortazavi,A. (2012) RNA editing in the human ENCODE RNA-seq data. *Genome Res.*, **22**, 1626–1633.
16. Yang,W., Chendrimada,T.P., Wang,Q., Higuchi,M., Seeburg,P.H., Shiekhattar,R. and Nishikura,K. (2005) Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat. Struct. Mol. Biol.*, **13**, 13–21.
17. Kawahara,Y., Zinshteyn,B., Chendrimada,T.P., Shiekhattar,R. and Nishikura,K. (2007) RNA editing of the microRNA-151 precursor blocks cleavage by the Dicer-TRBP complex. *EMBO Rep.*, **8**, 763–769.
18. Vesely,C., Tauber,S., Sedlazeck,F.J., vonHaeseler,A. and Jantsch,M.F. (2012) Adenosine deaminases that act on RNA induce reproducible changes in abundance and sequence of embryonic miRNAs. *Genome Res.*, **22**, 1468–1476.
19. Kawahara,Y., Zinshteyn,B., Sethupathy,P., Iizasa,H., Hatzigeorgiou,A.G. and Nishikura,K. (2007) Redirection of silencing targets by Adenosine-to-Inosine editing of miRNAs. *Science*, **315**, 1137–1140.
20. Morin,R.D., O'Connor,M.D., Griffith,M., Kuchenbauer,F., Delaney,A., Prabhu,A.L.L., Zhao,Y., McDonald,H., Zeng,T., Hirst,M. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells.. *Genome Res.*, **18**, 610–621.
21. Kim,Y.-K., Heo,I. and Kim,V.N. (2010) Modifications of small RNAs and their associated proteins. *Cell*, **143**, 703–709.
22. Duan,R., Pak,C. and Jin,P. (2007) Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. *Hum. Mol. Genet.*, **16**, 1124–1131.
23. Ryan,B., Robles,A. and Harris,C. (2010) Genetic variation in microRNA networks: the implications for cancer research. *Nat. Rev. Cancer*, **10**, 389–402.
24. Calin,G., Ferracin,M., Cimmino,A., Di Leva,G., Shimizu,M., Wojcik,S., Iorio,M., Visone,R., Sever,N., Fabbri,M. *et al.* (2005) A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *New Engl. J. Med.*, **353**, 1793–1801.
25. Shoshan,E., Mobley,A.K., Braeuer,R.R., Kamiya,T., Huang,L., Vasquez,M.E., Salameh,A., Lee,H.J., Kim,S.J., Ivan,C. *et al.* (2015) Reduced adenosine-to-inosine miR-455-5p editing promotes melanoma growth and metastasis. *Nat. Cell Biol.*, **17**, 311–321.
26. Peng,Z., Cheng,Y., Tan,B.C., Kang,L., Tian,Z., Zhu,Y., Zhang,W., Liang,Y., Hu,X., Tan,X. *et al.* (2012) Comprehensive analysis of RNA-seq data reveals extensive RNA editing in a human transcriptome. *Nat. Biotechnol.*, **30**, 253–260.
27. Gong,J., Wu,Y., Zhang,X., Liao,Y., Sibanda,V.L., Liu,W. and Guo,A.-Y. (2014) Comprehensive analysis of human small RNA sequencing data provides insights into expression profiles and miRNA editing. *RNA Biol.*, **11**, 1375–1385.
28. Vitsios,D.M. and Enright,A.J. (2015) Chimira: analysis of small RNA sequencing data and microRNA modifications. *Bioinformatics*, **31**, 3365–3367.
29. Patel,P., Ramachandruni,S.D., Kakrana,A., Nakano,M. and Meyers,B.C. (2015) mitrata: a web-based tool for microrna truncation and tailing analysis. *Bioinformatics*, doi:10.1093/bioinformatics/btv583.
30. Ai,Z., Xiang,Z., Li,Y., Liu,G., Wang,H., Zheng,Y., Qiu,X., Zhao,S., Zhu,X., Li,Y. *et al.* (2016) Conversion of monkey fibroblasts to transplantable telencephalic neuroepithelial stem cells. *Biomaterials*, **77**, 53–65.
31. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
32. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**(Suppl 1), D152–D157.
33. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, **57**, 289–300.
34. Han,M. and Zheng,Y. (2013) Comprehensive analysis of single nucleotide polymorphisms in human MicroRNAs. *PLoS One*, **8**, e78028.
35. Kiran,A. and Baranov,P.V. (2010) DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics*, **26**, 1772–1776.
36. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R10–R25.

37. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. *et al.* (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.
38. Robinson,J.T., Thorvaldsdottir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
39. Zheng,Y. and Zhang,W. (2010) Animal microRNA target prediction using diverse sequence-specific determinants. *J. Bioinform. Comput. Biol.*, **8**, 763–788.
40. Wang,C., Ren,R., Hu,H., Tan,C., Han,M., Wang,X. and Zheng,Y. (2014) Mir-182 is up-regulated and targeting cebpa in hepatocellular carcinoma. *Chinese J. Cancer Res.*, **26**, 17–29.
41. Li,R., Yu,C., Li,Y., Lam,T.-W., Yiu,S.-M., Kristiansen,K. and Wang,J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.
42. Cloonan,N., Wani,S., Xu,Q., Gu,J., Lea,K., Heater,S., Barbacioru,C., Steptoe,A., Martin,H., Nourbakhsh,E. *et al.* (2011) MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol.*, **12**, R126.
43. Zheng,Y., Li,T., Ren,R., Shi,D. and Wang,S. (2014) Revealing editing and SNPs of microRNAs in colon tissues by analyzing high-throughput sequencing profiles of small RNAs. *BMC Genomics*, **15**(Suppl 9), S11.
44. Bahn,J. H.H., Lee,J.-H.H., Li,G., Greer,C., Peng,G. and Xiao,X. (2012) Accurate identification of A-to-I RNA editing in human by transcriptome sequencing.. *Genome Res.*, **22**, 142–150.
45. Reid,J.G., Nagaraja,A.K., Lynn,F.C., Drabek,R.B., Muzny,D.M., Shaw,C.A., Weiss,M.K., Naghavi,A.O., Khan,M., Zhu,H. *et al.* (2008) Mouse let-7 miRNA populations exhibit RNA editing that is constrained in the 5′-seed/ cleavage/anchor regions and stabilize predicted mmu-let-7a:mRNA duplexes. *Genome Res.*, **18**, 1571–1581.
46. Larkin,M.A., Blackshields,G., Brown,N., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.