Published online 21 July 2021 *Nucleic Acids Research, 2021, Vol. 49, No. 18* ***e106***
https://doi.org/10.1093/nar/gkab610

Aberration-corrected ultrafine analysis of miRNA reads at single-base resolution: a *k*-mer lattice approach

Xuan Zhang [1], Pengyao Ping [1], Gyorgy Hutvagner [2], Michael Blumenstein [3] and Jinyan Li [1,*]

[1] Data Science Institute, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia, [2] School of Biomedical Engineering, Faculty of Engineering and IT, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia and [3] Faculty of Engineering and IT, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia

Received December 21, 2020; Revised July 01, 2021; Editorial Decision July 01, 2021; Accepted July 06, 2021

ABSTRACT

Raw sequencing reads of miRNAs contain machine-made substitution errors, or even insertions and deletions (indels). Although the error rate can be low at 0.1%, precise rectification of these errors is critically important because isoform variation analysis at single-base resolution such as novel isomiR discovery, editing events understanding, differential expression analysis, or tissue-specific isoform identification is very sensitive to base positions and copy counts of the reads. Existing error correction methods do not work for miRNA sequencing data attributed to miRNAs' length and per-read-coverage properties distinct from DNA or mRNA sequencing reads. We present a novel lattice structure combining kmers, (*k* – 1)mers and (*k* + 1)mers to address this problem. The method is particularly effective for the correction of indel errors. Extensive tests on datasets having known ground truth of errors demonstrate that the method is able to remove almost all of the errors, without introducing any new error, to improve the data quality from every-50-reads containing one error to every-1300-reads containing one error. Studies on experimental miRNA sequencing datasets show that the errors are often rectified at the 5′ ends and the seed regions of the reads, and that there are remarkable changes after the correction in miRNA isoform abundance, volume of singleton reads, overall entropy, isomiR families, tissue-specific miRNAs, and rare-miRNA quantities.

INTRODUCTION

With rapid developments of sequencing technology, high-throughput platforms have inexpensively produced huge amounts of genomic reads at unprecedented speed (1), for example by whole genome sequencing, total RNA sequencing, mRNA sequencing and small RNA sequencing. Recently, sequencing of miRNAs (a special type of small RNA molecules containing about 22 nucleotide bases) has been widely used to examine tissue-specific expression patterns, to identify isomiRs (mature miRNA variants) and to discover previously uncharacterized miRNAs (2–6). As key regulators in various biological processes, miRNA dysregulation is implicated in many diseases for example cancer and autoimmune disorders (7–11). Numerous studies also reaffirm that miRNA regulatory functions are involved in post-transcriptional gene silencing (PTGS), transcriptional gene silencing (TGS) and transcriptional gene activation (TGA) (12,13), in which miRNAs bind to nascent RNA transcripts, gene promoter regions or enhancer regions and exert further effects via epigenetic pathways (8,14).

Fine-granulated analysis of miRNA reads at single-base resolution for uncovering their isoforms (isomiRs) and alternative splicing is one of the most frontier research areas in this field (4,15–20). IsomiRs vary in size and base content, due to the alternative and imprecise cleavage of Drosha and Dicer, or the turnover of miRNAs (20,21). IsomiRs have been classified into four categories: 5′ trimmed isomiRs, 3′ trimmed isomiRs, 3′ addition isomiRs, and polymorphic isomiRs (22). 5′/3′ trimmed or addition isomiRs are defined as those miRNA sequences which have one or more bases trimmed or added respectively at the 5′ or 3′ end from the canonical miRNAs, while polymorphic isomiRs usually have substitution mutations, causing one or more bases different from the canonical miRNA. For such broad range of miRNAome analysis, super high-quality sequencing data is demanded because the definitions are very

[*] To whom correspondence should be addressed. Tel: +61 295149264; Fax: +61 295149264; Email: jinyan.li@uts.edu.au

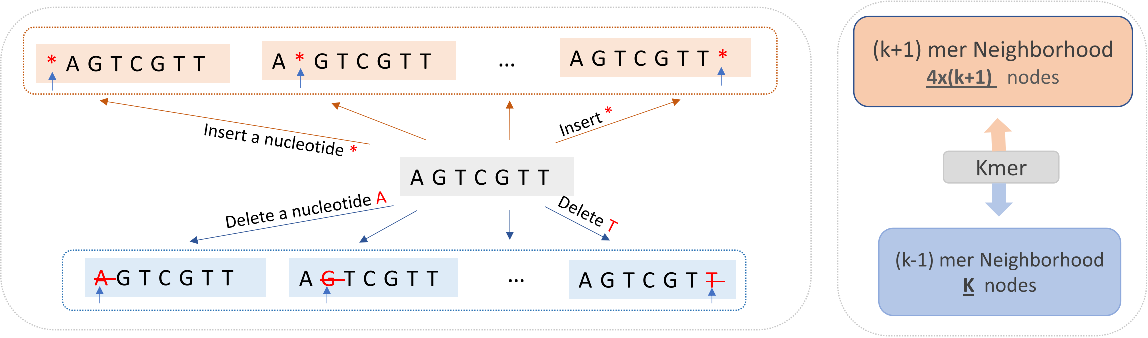

Figure 1. A 3-layer kmer lattice structure. The red * symbol represents any nucleotide (A, G, T or C).

sensitive to the base positions—one base difference can lead to entirely different read categorization. High-throughput sequencing technology produces short reads containing approximately 1% erroneous bases (1,23,24) such as aberrations of substitutions, base insertions, or deletions (indels). A previous study reported that the error percentage of most Illumina reads is ∼0.5% at best (25). These widely distributed errors or even erroneous bases fixed at only one position can cause lowered copy numbers for miRNA reads, and thus affect the calculation of miRNA expression levels and differential folds (26–30). Suppose a miRNA isoform has 100 copies in a diseased cell sample after library preparation, if there are substitution errors happened in five copies of them during the sequencing, three deletion errors in the other reads, and two insertion errors as well, then the reads count of this miRNA isoform would be tallied as 90 which is away from the ground truth 100. Further, the data may lead to wrong identification of isomiRs without correction of these errors. For example, those reads containing the deletion errors would be wrongly identified as a 5′ trimmed isoform of a canonical miRNA; If the errors occur at the seed region of a miRNA (conserved region of miRNAs), its target specificity analysis would be affected, potentially increasing the number of target transcripts (21,31,32). Although current research adopts 'abandon ambiguity reads or noise reads' to avoid misinterpreting erroneous sequence variants (ESVs) as isomiRs, the approach inevitably losses a part of the precious raw data (33,34). It is demanded to develop sophisticated algorithms to rectify these aberrations for truth-closer analysis of miRNA reads in a wide range of applications.

None of the existing error correction methods suits well for miRNA sequencing data, since they have not considered the unique characteristics of miRNA reads (short length and varying per read coverage). Besides, most of the methods, designed for DNA or mRNA sequencing reads, only focus on the correction of substitution errors and do not support indels error correction. So far, these methods have taken two streams of different correction ideas. The first one is a kmer-based error correction idea, represented by BCOOL (35), BFC (36), ACE (37) and BLESS (38). The key step is to examine the frequencies of kmers to distinguish between solid and weak k-mers according to a fixed global frequency threshold. Then the solid kmers (assumed as error-free) are referred as templates to rectify weak kmers (assumed as error-containing) to obtain correct reads. The second idea is a multi-alignment based error correction approach, represented by coral (39), ECHO (40) and Karect (41). These methods usually group those reads sharing the same kmer and then concatenate such a group of reads to form a long consensus contig. The contig is assumed error-free to correct erroneous bases. There are also a few methods designed for RNA sequencing reads error correction, for example Seecer (42) and Rcorrector (43). These approaches do not work for miRNA sequencing data error correction. For example, the consensus idea is not applicable to miRNA data because each read already encompasses one entire miRNA sequence. Our study did verify that the existing methods tend to significantly under-correct the errors and are prone to introducing tremendous number of new errors.

We present an error RECtification method for miRNA sequencing reads (named miREC), which is the first tool to address the problem of miRNA sequencing errors. Unlike the existing methods which have the primary goal of correcting substitution errors, our miREC concentrates more on insertion and deletion errors for excellent correction performance. The novel step of our method is the use of a 3-layer $(k-1)$mer–k-mer–$(k+1)$mer lattice structure to maintain the frequency differences of the kmers (Figure 1). These superset-subset frequency differences are very effective to detect the errors especially the indel errors. The lattice structure is also a moving structure where k is set continuously from a small number to a big number 23 or 25 for a full coverage of error correction. Extensive tests on both simulated and experimental raw miRNA sequencing datasets show that miREC can excel performance in all of the precision, recall and gain.

MATERIALS AND METHODS

A miRNA sequencing read r is a sequence $r_1r_2\cdots r_n$, $r_i \in \Sigma = \{A, C, G, N, T\}$, where A, C, G and T stand for the nucleotide bases Adenine, Cytosine, Guanine and Thymine respectively, and the character N stands for a uncertain nucleotide; n is the length of r. Usually, the length n of a miRNA read ranges from 15 to 28 in a dataset, but each read encompasses one entire miRNA. A kmer *substringk* is a contiguous subsequence in a read r.

Neighborhood of a *k*-mer and a 3-layer kmer lattice structure

Given a miRNA sequencing read multi-set *RS* and a setting k, the copy count (or frequency) of a distinct read *r* in *RS* is the total number of its copies in *RS*, and the copy count (or frequency) of a distinct kmer in *RS* is the total number of its copies in *RS*. KMC3 (44) is used by this work as a kmer counter for these calculations.

Consider a kmer *substringk*, this kmer's k-neighborhood is defined as the set of kmers *H*(*k*, *substringk*) containing all possible distinct kmers of *RS* that each have only one base difference from *substringk*. Similarly, *substringk*'s (*k* – 1)-neighborhood is defined as the set of (*k* – 1)mers *H*((*k* – 1), *substringk*) containing all possible distinct (*k* – 1)mers of *RS* each of which is an immediate subset of *substringk*, and *substringk*'s (*k* + 1)-neighborhood is defined as the set of (*k* + 1)mers *H*((*k* + 1), *substringk*) containing all possible distinct (*k* + 1)mers of *RS* each of which is an immediate superset of *substringk*.

For example, when the kmer is given as *GTC* and assume that all its proper supersets and subsets exist in *RS*, then its (*k* + 1)-neighborhood $H(4, GTC) = \{\underline{A}GTC, \underline{T}GTC, \underline{C}GTC, \underline{G}GTC, G\underline{A}TC, G\underline{T}TC, G\underline{C}TC, G\underline{G}TC, GT\underline{A}C, GT\underline{T}C, GT\underline{C}C, GT\underline{G}C, GTC\underline{A}, GTC\underline{T}, GTC\underline{C}, GTC\underline{G}\}$. Its (*k* – 1)-neighborhood $H(2, GTC) = \{TC, GC, GT\}$. These three neighborhoods of kmer *substringk* can be combined and it is called a 3-layer kmer lattice structure of *substringk*. A schematic example of this lattice structure is shown in Figure 1.

Error correction steps

The first step of the algorithm is to rectify substitution errors in *RS*. The algorithm traverses all of the distinct kmers. If a kmer *substringk* has a frequency lower than a threshold τ (a small integer like 1, 2 or 3) and there exist at least one kmer in *substringk*'s k-neighborhood *H*(*k*, *substringk*) whose frequency is larger than τ, we conjecture that *substringk* contains a substitution error. We choose the kmer with the highest frequency in *H*(*k*, *substringk*) as template to rectify the erroneous base in *substringk*. In the case where more than one kmer neighbors have the same high frequency, we choose the smallest kmer according to the alphabetical order as the template. After the change in *substringk*, those reads in *RS* containing the original *substringk* are changed accordingly; some of them may become identical with other reads in *RS*. We introduce a double-checking technique to decide whether we eventually accept the correction—we double-check the updated frequencies of the distinct reads in the updated *RS*. Only when the corrected reads become identical with a read having a frequency higher than τ, we confirm the correction; Otherwise, we abandon the modification. With this double-checking strategy, we can avoid the issue of over-correction.

The second step is to rectify indel errors in the updated *RS* after the correction of substitution errors. The procedure is similar to correcting the substitution errors. But the concept is fundamentally different. We traverse all of the distinct kmers in the updated *RS*. If a kmer *substringk* has a frequency lower than a threshold τ and there exist at least one kmer in *substringk*'s (*k* – 1)-neighborhood *H*((*k* – 1), *substringk*) whose frequency is larger than τ, we conjecture that *substringk* contains an insertion error. On the other hand, if there exists at least one kmer in *substringk*'s (*k* + 1)-neighborhood *H*((*k* + 1), *substringk*) whose frequency is larger than τ, we conjecture that *substringk* contains a deletion error. We choose the kmer with the highest frequency in *H*((*k* – 1), *substringk*) or in *H*((*k* + 1), *substringk*) as template to rectify the insertion error in *substringk* or to add the deleted base into *substringk*. After the change in *substringk*, those reads in *RS* containing the original *substringk* are changed accordingly; some of them may become identical with other reads in *RS*. Again we use the double-checking strategy to decide whether we eventually accept the correction. We iterate these two steps by setting *k* from k_1 (usually 8) to k_{end} (usually 20 or 25). Setting a start *k* as 8 is because of that we find low-frequency kmers (e.g. frequency equal to 1) at this *k* but we cannot find such low-frequency (<τ) kmers for *k* = 7. Starting from *k* = 8, we correct substitution errors first, then we perform the indel error correction, till *k* reaches k_{end}. Our method is named miREC built from a 3-layer kmer lattice structure for effective correction of miRNA sequencing errors especially those insertion and deletion errors. The pseudo code of our algorithm is shown in Algorithm 1.

Algorithm 1: miRNA Sequencing Reads Error Correction

```
Input: A read set $\mathcal{RS} = \{r^1, r^2, \cdots, r^n\}$, a frequency
       border $\tau$, a k value region $(k_1, k_{end})$
Output: A corrected read set $\mathcal{S}$
Function ERROR_CORRECTION $(\mathcal{RS}, (k_1, k_{end}), \tau)$
begin
    $H_r[1 \ldots m] \leftarrow$ hash table ;                ▷ read info
    $H_a[1 \ldots m], H_b[1 \ldots m], H_c[1 \ldots m] \leftarrow$ hash table ;
     ▷ kmer info; $H.[i]$ is an array; Each element of
     $H.[i]$ is a tuple composed of sequence and its
     frequency
    for $k = k_1$ to $k_{end}$ do
        $H_a[1 \ldots m]$ to $KMC3(\mathcal{RS}, k)$
        $H_b[1 \ldots m]$ to $KMC3(\mathcal{RS}, k-1)$
        $H_c[1 \ldots m]$ to $KMC3(\mathcal{RS}, k+1)$
        for $i = 1$ to $n$ do
            $(s, f) \leftarrow Count(r^i)$
            Append $(s, f)$ to $H_r[s].f$
        $C_{kmer} \leftarrow \emptyset$ ; ▷ the kmer with highest frequency
        for $i = 1$ to $n$ do
            foreach $kmer \in r^i$ do
                if $H_a[kmer].f < \tau$ then
                    $C_{kmer} \leftarrow FindNeighbor(H_a)$
                    $C_r \leftarrow Replacekmer(C_{kmer}, r^i)$
                    if $H_r[C_r].f > \tau$ then
                        $r^i \leftarrow C_r$
        for $i = 1$ to $n$ do
            foreach $kmer \in r^i$ do
                if $H_a[kmer].f < \tau$ then
                    $C_{kmer} \leftarrow FindNeighbor(H_c, H_c)$
                    $C_r \leftarrow Replacekmer(C_{kmer}, r^i)$
                    if $H_r[C_r].f > \tau$ then
                        $r^i \leftarrow C_r$
    $\mathcal{S} \leftarrow \mathcal{RS}$
    return $\mathcal{S}$
```

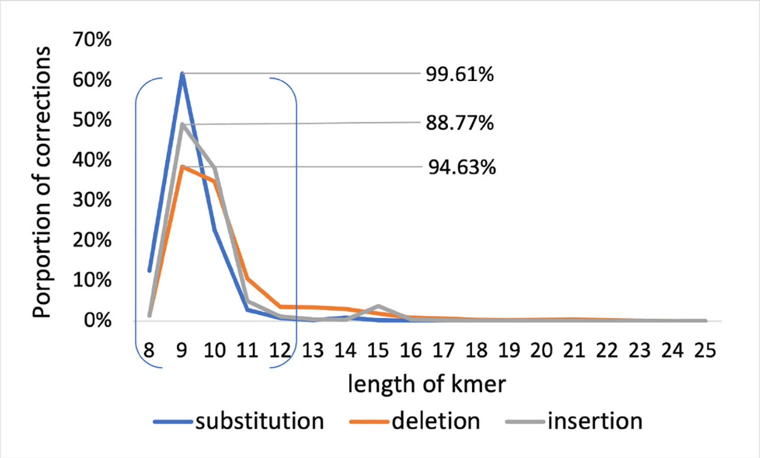

Figure 2. The proportion of corrections varies at different lengths of *k*-mer.

Our miREC has been implemented as a software prototype. It provides several parameters for users to specify their tasks. Three most useful settings are: the error types, the frequency threshold τ, and the kmer range [k_1, k_{end}]. miREC has two running modes: one is for the substitution error correction only, the other is for the correction of both indel and substitution errors. Based on our experience, the frequency threshold τ is best recommended as 5 by default, and the *k*-mer range parameter is set as [8,15]. The higher frequency τ is set, the bigger number of bases might be considered as errors. Thus, users should be cautious about using a too large frequency threshold to avoid over-correction.

Every iterative step of miREC with the increasing length of *k*-mer each time by 1 in the range [k_1, k_{end}] actually corrects different amounts of errors. As shown in Figure 2, after five consecutive lengths of *k* are iterated, about 99.61% of substitution errors, 88.77% of insertion errors and 94.63% of deletion errors can be corrected on average over 12 wet-lab salmon datasets (Table 2) if k_1 is set as 8. With more loops of correction, more erroneous bases are detected and corrected. As each iterative loop consumes the same order of time complexity, users are suggested to narrow the kmer range (by setting k_{end} smaller) to shorten the program running time while correcting almost all of the errors for those miRNA sequencing datasets of huge size.

The source codes of miREC are publicly available online at https://github.com/XuanrZhang/miREC.

Simulated datasets (with known ground truth) and public wet-lab miRNA sequencing reads both for performance evaluation

To evaluate the performance of error correction methods, simulated datasets are required and the ground truth of the errors should be known. We introduce a novel process to generate simulated datasets that would have a close nature to wet-lab miRNA sequencing reads. We have two considerations in the process. One is to computationally replicate lab-verified miRNA sequences as templates to form the basic sequences of the simulated datasets, then we duplicate these basic sequences such that the copy counts of them follow a real distribution from a wet-lab dataset of miRNA sequencing reads. In fact, we replicated the mature miRNA sequences in miRBase (45) as the templates, and made the copy count distribution of these template sequences to follow the distribution drawn from a typical miRNA dataset (accession number SRR866573). In other words, the sequences in our simulated datasets are not random sequences (they are real lab-verified miRNA sequences); their copy count distribution is not random either. Then we injected errors into the simulated datasets under an error rate of 0.1% per base (24). Specifically, we randomly selected two reads from every 100 reads in the dataset; then for each selected read, we injected an erroneous base (substitution, deletion or insertion) randomly at any position of the read. We recorded all of these randomly and purposely injected errors for performance evaluation.

Considering that some existing methods only support substitution error correction, we synthesized 8 simulated datasets: four datasets containing substitution errors only (denoted as D_sub1, D_sub2, D_sub3 and D_sub4), and four datasets containing a mixture of 80% substitution and 20% indel errors (denoted as D_mix1, D_mix2, D_mix3 and D_mix4). More details of the simulated datasets are shown in Table 1.

Table 1. Description of our simulated datasets

		ID	Total erroneous bases	Per read error rate
Simulated Datasets	subs only	D_sub1	3071	3.03%
		D_sub2	3022	2.98%
		D_sub3	2973	2.93%
		D_sub4	3124	3.08%
	mix errors	D_mix1	1602 213 211	2.00%
		D_mix2	1618 188 206	1.98%
		D_mix3	1598 184 177	1.93%
		D_mix4	1625 226 217	2.04%

Notes: '_sub' means datasets contain substitution errors only and '_mix' means datasets contain both substitution and indel errors. Total erroneous bases list substitution, insertion and deletion errors respectively.

Table 2. 12 wet-lab salmon miRNA sequencing datasets and four human miRNA sequencing datasets.

Tissue	Total reads	Unique reads	Accession ID
Liver	1 446 902	64 593	SRR866573
Liver	1 647 133	75 273	SRR866579
Spleen	8 597 057	295 940	SRR866583
Spleen	2 236 013	89 165	SRR866587
Kidney	10 065 660	243 430	SRR866589
Head kidney	7 375 957	246 444	SRR866590
Heart	2 812 993	118 366	SRR866605
Brain	6 331 448	132 558	SRR866611
Intestine	12 428 822	197 094	SRR866612
White muscle	5 972 384	142 444	SRR866613
Gills	6 240 735	132 038	SRR866614
One day old individual	18 041 561	172 048	SRR866615
human beta cells datasets			
In low glucose	63 008 516	5 803 166	SRR13208981
In high glucose	33 444 257	1 856 318	SRR13208980
human brain datasets			
Sample of aged 75	11 849 807	635 169	SRR12881030
Sample of aged 94	17 250 812	361 039	SRR12881018

Wet-lab miRNA sequencing datasets for our performance evaluation are all downloaded from the Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra/) under the accession numbers SRP022967, SRP296813 and SRP288246. These datasets have been originally studied for problems related to salmon fish miRNAs (46,47), human beta cells, or Alzheimer's disease. This work used 12 salmon miRNA sequencing datasets which were acquired from particular salmon tissue samples, including liver, spleen, kidney, heart, brain, etc; and used two human beta-cell miRNA sequencing datasets which are about miRNA expression comparison between those cells incubated with a solution of low glucose (2 mM) and those with a high glucose (20 mM) in extracellular vesicles. The two other human miRNA sequencing datasets analyzed here are about brain samples related to post-mortem Alzheimer's disease. One is from a male patient aged 75, the other is from a male patient aged 94. All the reads in the above datasets contain the sequences of adaptors; we used the cutadapt tool (48) to remove the adaptors before our error correction. More details of these cleaned datasets are shown in Table 2.

To rigorously evaluate the error correction performance, we also randomly and purposely inject a small number of errors into these wet-lab datasets, rather than into the simulated datasets, to see whether our algorithm can detect and correct these errors of ground truth, together with other errors without ground truth. Only when all of these artificial errors in the experimental miRNA sequencing reads can be detected and corrected, the corrections on the other bases can be highly trustable. This small number of artificial errors constitutes only 0.5% of total corrections in each dataset to avoid changing the original nature of the data. We have done these for three salmon datasets (liver, heart and spleen tissues), and one human brain miRNA dataset from the male patient aged 75. For each of these datasets, we injected small numbers of errors twice.

Evaluation metrics

As the ground truth of the errors in the simulated datasets are known, we can use recall, precision and gain to compare the correction performance between different methods. On the wet-lab miRNA sequencing datasets, we measure the copy count changes of the reads, the entropy changes of the whole set of reads, and locations of the rectifications to understand the importance of error correction. There is no recall or precision performance on the wet-lab miRNA sequencing datasets, because the ground truth of error distributions is unknown.

Performance evaluation metrics on the simulated miRNA sequencing datasets. Precision, recall and gain rate are given as follows to assess the correction performance on the simulated datasets:

- Precision: $TP/(TP{+}FP)$, shows the fraction of truly corrected bases among all the changed bases.
- Recall: $TP/(TP{+}FN)$, shows the fraction of truly corrected bases among all the bases which are supposed to be corrected.
- Gain: (TP – FP)/(TP + FN), shows the fraction of removed errors without inducing additional errors.

where true positives (TP) correspond to corrected errors; true negatives (TN) correspond to initially correct bases left untouched; false positives (FP) correspond to newly introduced errors; and false negatives (FN) correspond to unidentified errors.

Metrics used for performance evaluation on wet-lab miRNA sequencing datasets. We examine the changes of miRNA copy counts and dataset entropy changes before and after error correction for multiple salmon fish miRNA sequencing datasets. Besides, we also summarize the position information of the corrections in the reads to record the proportion of corrections in the seed region. We define the miRNA count, dataset entropy and errors in the seed region as follows.

- miRNA count: the copy count of miRNA appearing in the datasets, which is corresponding to miRNA expression level or miRNA abundance.
- Dataset entropy: $-\sum_{i=1}^{n} p_i \cdot \log p_i$, where p_i is the proportion of reads whose frequency is small than i. We calculate the entropy for low-frequency reads and sum up to interpret the degree of disorder in the read dataset. When the entropy turns to be small, it means the certainty of the miRNA expression becomes higher.
- Errors in seed region: erroneous bases in the seed region, which is a conserved sub-sequence of miRNA (mostly situated at positions 2–8). Precise bases in the seed region are vital since the seed sequence must be perfectly complementary with its target mRNA to make the miRNAs function.

RESULTS

Our analysis and results are presented in five main parts. The first part is about the correction performance on the 8

Table 3. Outstanding error correction performance by our miREC in comparison with the best available tools

	Gain(%)					Recall(%)					Precision (%)				
	miREC	BFC	Cor	Kar	Rcor	miREC	BFC	Cor	Kar	Rcor	miREC	BFC	Cor	Kar	Rcor
D_s1	**97.88**	69.85	5.44	94.2	5.99	**97.88**	70.63	9.96	94.43	5.99	99.64	82.07	67.4	**99.76**	92
D_s2	**97.98**	84.98	4.8	94.44	5.43	**97.98**	85.7	9.46	94.87	5.43	99.5	99.08	65.75	**99.55**	91.62
D_s3	**97.61**	83.99	5.01	94.01	5.85	**97.61**	84.76	9.72	94.55	5.85	99.42	99.06	66.9	**99.43**	91.1
D_s4	**97.86**	84.51	5.83	94.01	5.19	**97.86**	85.21	10.34	94.37	5.19	99.48	99.07	68.87	**99.63**	90
AVE	**97.83**	80.83	5.27	94.17	5.62	**97.83**	81.58	9.87	94.55	5.62	99.51	94.82	67.23	**99.59**	91.18
D_m1	**95.96**	65.93	2.22	65.98	0.05	**95.96**	67.11	9.22	75.64	0.05	**98.78**	95.78	55.16	88.67	100
D_m2	**95.73**	68.24	1.04	66.95	0.1	**95.73**	69.43	8.2	77.53	0.1	**98.67**	96.28	52.05	87.99	100
D_m3	**96.02**	68.56	1.43	71.47	0.05	**96.02**	69.78	8.88	78.61	0.05	**98.58**	96.68	53.05	91.67	100
D_m4	**96.76**	65.57	2.47	65.28	0	**96.76**	66.73	9.33	74.95	0	**99.12**	96.17	57.27	88.57	0
AVE	**96.12**	67.07	1.79	67.42	0.05	**96.12**	68.26	8.91	76.68	0.05	**98.79**	96.23	54.38	89.22	75

Notes: AVE indicates the average score over the four datasets. Bold font indicates the best result in the row. Cor, Kar and Rcor stand for the Coral method, the Karect method and the Rcorrect method respectively. D_s indicates datasets containing only substitution errors, while D_m indicates datasets containing mixed substitution, insertion and deletion errors. The underline 100 precision of Rcor on D_m1, D_m2 and D_m3 stands for only one, two and one base is corrected respectively.

simulated datasets; the second part is about correction performance on the wet-lab miRNA sequencing datasets after a small number of artificial errors are injected; the third part is about copy abundance recovery, entropy change and rectification site summary on the recently published salmon fish miRNA sequencing datasets after error correction; the fourth part provides detailed case studies on the change of isomiR families, tissue-specific isoforms, differentially expressed biomarkers and rare-miRNA quantity enhancement after the error correction on some of the wet-lab datasets, including the human miRNA sequencing datasets. The fifth part presents our verification results on sequencing reads datasets of 963 miRXplore Universal Reference miRNAs (three replicates) and their spike-in at eukaryotic cells.

Gain, recall and precision performance on the simulated miRNA sequencing datasets

The correction performance of our miREC is presented in Table 3 in comparison with algorithms Karect (41), Coral (39), BFC (36), Rcorrector (43) and Bcool (35). Coral and Karect are multi-alignment based error correction methods. BFC is a representative of the k-mer based error correction methods. BFC requires a prior-setting of the *k* parameter; the best *k* in this work is 21 (namely, under other k settings, BFC did not exceed the performance of when $k = 21$). Karect is one of a few correction tools which supports the correction of indel errors. Rcorrector, a RNA reads error correction method, has a performance higher than another RNA correction method Seecer (42). Rcorrector also needs to set the *k* parameter and the best k in this work is 17. Even using the optimal k settings, only a few bases can be corrected by Rcorrector. A very recent error correction algorithm Bcool (35), which uses a de Bruijn graph as the platform to correct errors, could not detect any errors in the simulated datasets. This surprising performance is not included in the table.

Our method miREC has excelled in the correction performance:

- It did not introduce any new error, namely, it achieved the same gain and recall rates on all of the 8 datasets;
- It detected and corrected almost all of the errors including the indel errors; the recall rate ranges between 96.0–97.9%; the precision ranges between 98.6–99.5%;
- It improved the overall data quality remarkably: (i) from every 50 reads containing one error to every 1300 reads containing one error for the four error-mixed datasets and (ii) it improved the data quality from every 30 reads containing one error to every 1650 reads containing one error for the four substitution-only datasets.

The average recall and gain rate of miREC are much superior to Karect (the second-best method) respectively by 3.28% and 3.66% on the four substitution-only datasets, and by 19.44% and 28.7% on the four error-mixed datasets. Specifically, the average recall rates of miREC are 97.83% and 96.12% on the four D_sub datasets and on the four D_mix datasets respectively, which are 16.25%, 87.96% and 3.28% (on the D_sub datasets) and 27.86%, 87.21% and 19.44% (on the D_mix datasets) better than BFC, Coral and Karect. This implies that there are lots of errors undetected by these baseline methods meanwhile they introduced a lot of new errors (gains and recall not equal). The multi-alignment method performed worst on these miRNA datasets. A possible reason is that the alignment strategy could not differentiate miRNA reads well due to the short length of miRNAs. Rcorrector had a very low recall and gain performance as well, that means most of the errors were not detected by the method.

The performance of miREC is robust across all the 8 datasets including the four mixed-error datasets, in contrast to the baseline methods which exhibited a poor performance on the detection and correction of the indel errors. The gain rate of BFC drops from 80.83% (on the four D_sub datasets) to only 67.07%(on the four D_mix datasets), and the gain of Coral drops from 5.27% to 1.79%. It suggests that the performance of these methods on the substitution error correction was interrupted and affected by the addition of the indel errors into the datasets. As real-life wet-lab sequencing reads more or less company with a small amount of indel errors, our miREC provides an unalterable advantage over the baseline methods for the correction of all types of aberrations.

Correction performance on wet-lab miRNA sequencing datasets injected with small numbers of artificial errors

We made 27 random base modifications (total 21 substitutions, 3 insertions and 3 deletions) at the salmon liver miRNA sequencing dataset (SRR866573). These mannual modifications introduced/injected 18 *genuine* errors into the dataset, where a random base modification is *not* considered as a genuine error if its correspondingly modified read becomes identical with another read having a high frequency (i.e. copy count > 5).

Our algorithm corrected all of these 18 genuine errors (100% recall). For example, read @SRR866573.64765 (TGCGGACCAGGGGAATCCGACT) had a random manual deletion at base position 5, becoming TGCGACCAGGGGAATCCGACT; our miREC detected this error and restored it to its original base. As another example, read @SRR866573.212344 (AAGCTGCCAGCTGAAGAACTG) had a random manual substitution from C to G at position 8, becoming AAGCTGCGAGCTGAAGAACTG; our miREC corrected it successfully. Read @SRR866573.1103128 (AAGCGGGCCCCCAAACTTCTGT) had a random manual insertion of G at position 16, becoming AAGCGGGCCCCCAAAGCTTCTGT; again, our miREC successfully detected this error and corrected it. For the remaining 9 randomly injected base modifications, they did not cause genuine errors because each of their reads was transformed into another read that has a high copy count in the same dataset. For example, read @SRR866573.360151 (ATGACCTATGAATTGACAGCCT) had a random manual substitution from T to C at position 21 (the last position). With this modification, the read becomes another read ATGACCTATGAATTGACAGCCC which has 156 copies. This modification was unable to be restored to its original base because every *k*-mer in ATGACCTATGAATTGACAGCCC was highly frequent (at least 156 copies), namely, containing no error. Note that this modification should not be restored to ensure no over-correction would happen in practice, otherwise the correction would be of guilty. For performance comparison, the second-best method Karect was applied to the same error-injected salmon liver dataset, but it corrected only 5 of the 18 genuine errors.

We repeated this test with another round of manual base modifications at SRR866573 (total 28 modifications including 20 substitutions, 6 insertions and 2 deletions). Our miREC detected and corrected all of the 20 genuine errors (100% recall again). In comparison, Karect corrected only 9 of them.

Similarly, our miREC corrected all of the genuine errors caused by small numbers of random base modifications at other wet-lab miRNA sequencing datasets (40 substitutions, 3 insertions and 6 deletions; or second round 45 substitutions, 7 insertions and 7 deletions at the salmon heart dataset. 38 substitutions, 4 insertions and 4 deletions; or second round 43 substitutions, 6 insertions and 6 deletions at the salmon spleen dataset). However, Karect corrected only 8 of the 27 genuine errors or only 8 of the 35 errors on these two error-injected salmon heart datasets, and had similar performance on the two error-injected salmon spleen datasets.

On the two human brain datasets, our miREC achieved the same perfect performance (100% recall) to correct all of the genuine errors caused by small numbers of random base modifications (about 300 base modifications which had resulted in 130 and 120 genuine errors). However, Karect could only fix 12 or 20 genuine errors in these two datasets. Our source codes for the random error injection into wet-lab miRNA sequencing datasets and more detailed correction results are available at github link https://github.com/XuanrZhang/miREC.

Changes in isoform abundance, whole set entropy and base positions after error correction at the salmon fish miRNA sequencing reads

The perfect recall performance on the small numbers of errors injected into wet-lab miRNA sequencing datasets and the excellent gain performance on the simulated datasets are strong combined evidence to support our correction results on wet-lab datasets where the ground truth of errors are not available.

The salmon liver miRNA sequencing dataset (SRR866573) has a total of 900 814 reads, containing 32,972 distinct reads before error correction. After error correction by our miREC, there are only 27 299 distinct reads some of which gained plenty of abundance. In other words, most of the error-contained reads were corrected and turned to be identical with some other reads, making the abundance merging meanwhile the disappearance of the originally error-contained reads.

See Figure 3 for an average percentages of the distinct miRNAs over the 12 datasets that have a high- or low-level abundance recovery. There are around 47.3% of the distinct miRNAs whose copy counts have increased by more than 10% after the corrections, in particular, about 5.5% of the distinct miRNAs have obtained above 50% abundance increase. These corrections are useful to draw more reliable conclusions about miRNA discovery or isomiR classification or tissue-specific biomarker discovery (case studies presented later).

The reads abundance recovery of the miRNA isoforms after error rectification in a dataset implies that the numbers of distinct reads are decreased as reported above. We present Figure 4 to illustrate the overall entropy change of every entire dataset before and after the error correction to quantify this point. On average the entropy of the 12 datasets is shrank by 15.11% when the parameter *k* of miREC ranges from 8 to 20, and the entropy score decreased by 14.51% when *k* ranges from 8 to 25. These entropy declines (with slight variance) in the 12 datasets theoretically mean that the certainty of the miRNA expression levels is greatly improved. In other words, our miREC can enhance the data quality in the perspective of achieving a lower entropy or a higher certainty.

We found that the aberrations could occur at every base position of the reads. But, one-third of the errors are detected and corrected at the seed region of the miRNAs (Figure 5). These corrections at the seed region provide great benefits for miRNAs' target prediction analysis. There are also a high percentage of the indel or substitution correc-

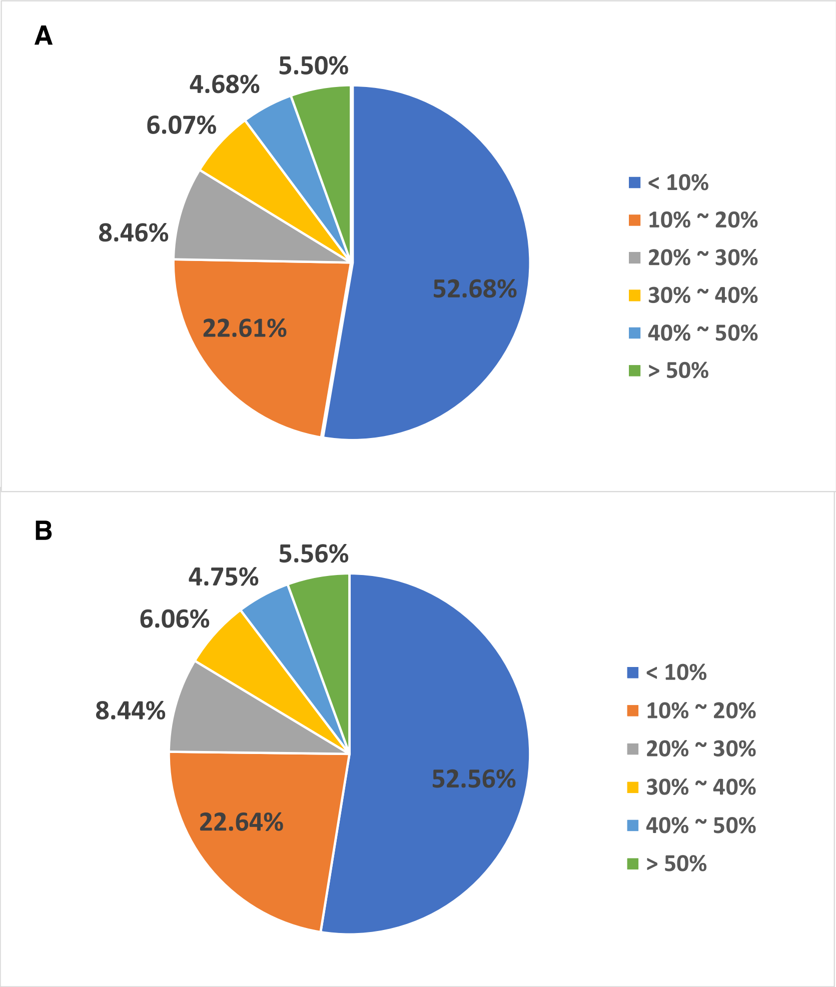

Figure 3. Proportions of distinct-read counts are changed compared with uncorrected data in average of 12 salmon datasets. (**A**) The miREC runs with continuous k value from 8 to 20. (**B**) The miREC runs with continuous *k* value from 8 to 25.

tions at position 1 which is a base position very sensitive to the definition of trimmed or addition isomiRs.

Case studies related to isomiR families, tissue-specific miRNAs and rare-miRNA quantity recovery

We report examples of miRNAs whose read counts have changed a lot after error correction. We also show examples of tissue-specific miRNAs after error correction, and describe the change in the ranking lists of differentially expressed miRNAs.

Case study 1: big abundance recovery. In the salmon heart dataset, read *TTGGTCCCCTTCAACCAGCTGT AAT* (mapped to miR-133a-1 in miRBase (45)) had 10 copies. Our miREC detected 13 erroneous reads related to this miRNA. Eight substitution errors happened at position 24 base A (sequenced to G or T), and five happened at position 25 base T (sequenced to A or G). After our correction, the abundance level of miR-133a-1 increased from 10 to 23, a 130% abundance recovery. Other two miRNAs (ssa-miR-133a-3p and ssa-miR-133a-5p) from the same miRNA family also recovered their abundance in the sequencing reads. See the read counts and change details in Table 4. We note that currently annotated functions of miR-133a-1 are related to conventional central osteosarcoma and heart conduction disease (47,49). With the refined abundance understanding, its functions can be re-examined more deeply.

Another example in Table 5, read *ATCCCGGACGAG CCCCCAA*, had 18 copies and its abundance increased to 31 after miREC correction. The aberrations include four deletion errors at position 1 (base A deleted), four substitution errors at position 19 base A (sequenced to C) and two insertion errors at position 1 and 2 (base A inserted). This error distribution implies that the sequencing mistakes can occur at multiple base positions with multiple times; and that our miREC is capable of correctly detecting these errors and performing accurate rectifications.

For comparison, we tested the second-best method Karect on this salmon heart dataset to see whether the same mistakes could be corrected. Take the cases in Table 4 as example, only three erroneous reads of the first read *TTGGTC CCCTTCAACCAGCTGTAAT* were detected by Karect (we detected 13); none of the erroneous reads of the other

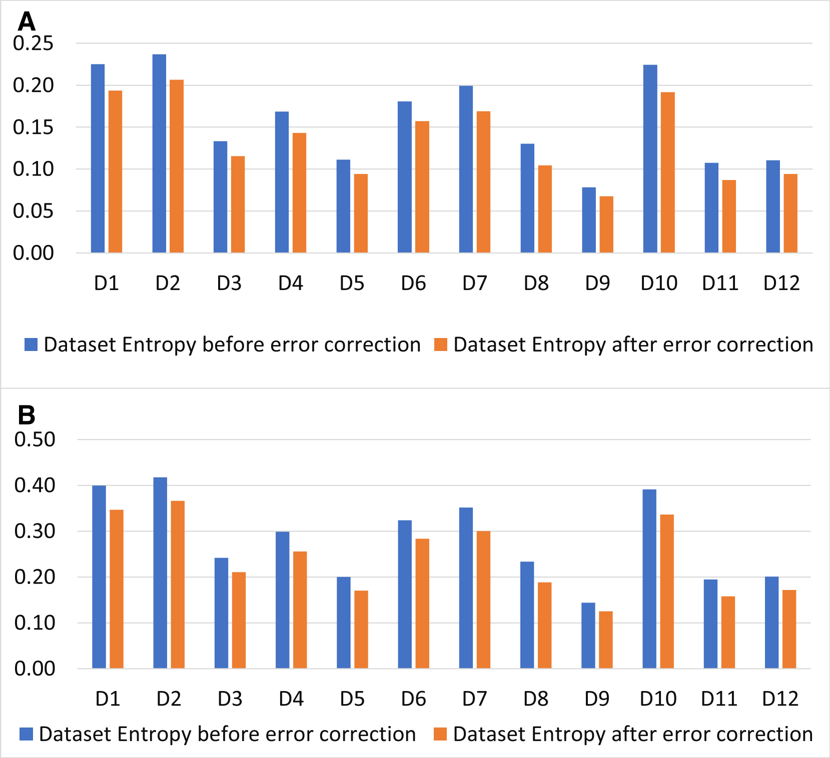

Figure 4. Dataset-entropy changes before and after the error correction by miREC on the 12 salmon miRNA datasets. (**A**) when the continuous k settings from 8 to 20; (**B**) when the continuous *k* settings from 8 to 25.

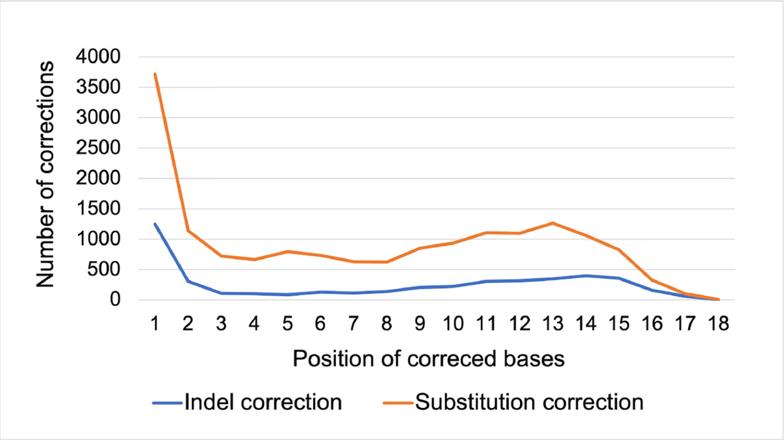

Figure 5. The distributions of corrections at different positions.

Table 4. Changes in the read counts of some miRNAs

Sequence	Read count before correction	Read count after correction	Abundance change percentage	Isoforms/editing events
TTGGTCCCCTTCAACCAGCTGTAAT	10	23	130.00%	
TTGGTCCCCTTCAACCAGCTGTA-T	44	45	2.27%	5′ deletion
TTGGTCCCCTTCAACCAGCTGTAA-	107	111	3.74%	3′ deletion
Related Erroneous Reads List				
TTGGTCCCCTTCAACCAGCTGTAGT	4	0	Error removal	Substitution error
TTGGTCCCCTTCAACCAGCTGTATT	4	0	Error removal	Substitution error
TTGGTCCCCTTCAACCAGCTGTAAA	3	0	Error removal	Substitution error
TTGGTCCCCTTCAACCAGCTGTAAG	2	0	Error removal	Substitution error

Table 5. isomiRNA detection

Sequence	Read count before correction	Read count after correction	Abundance change percentage	Isoforms/editing events
ATCCCGGACGAGCCCCCAA	18	31	72.22%	
ATCCCGGACGAGCCCCCA-	1489	1513	1.61%	Deletion
ATCCCGGACGAGCCCCCAT	16	20	25.00%	Substitution
ATCCCGGACGAGCCCCCAAA	9	17	88.89%	Insertion
Related Erroneous Reads List				
-TCCCGGACGAGCCCCCAA	4	0	Error removal	Deletion error
ATCCCGGACGAGCCCCCAC	4	0	Error removal	Substitution error
AATCCCGGACGAGCCCCCAA	1	0	Error removal	Insertion error
ACCCCGGACGAGCCCCCAA	1	0	Error removal	Substitution error
AATCCCGGACGAGCCCCCAA	1	0	Error removal	Insertion error

Table 6. Rank changes of the top-10 common miRNAs in salmon heart and brain tissues after error correction

miRNA sequence	After_rank	Before_rank
TCTTTGGTTATCTAGCTGTATG	1	2
TCTTTGGTTATCTAGCTGTAT	2	3
TTTGTTCGTTCGGCTCGCGTT	3	5
TCTTTGGTTATCTAGCTGTA	4	8
TTGCATAGTCACAAAAGTGATC	5	6
TCTTTGGTTATCTAGCTGTATGA	6	7
TGGAAGACTAGTGATTTTGTTG	7	10
TAAAGCTAGAGAACCGAATGTA	8	11
TAAGGCACGCGGTGAATGCC	9	12
ATGGCACTGGTAGAATTCACT	10	13

Notes: After_rank indicates the rank after error correction, while Before_rank indicates the rank before error correction.

two reads in the table were detected. Only one of the four related erroneous bases was corrected by Karect, while all of the related erroneous bases were corrected by our method.

Case study 2: miRNA isoforms and editing events. Editing events and isoform variations at the cleavage sites can cause slight but important difference in many miRNA sequences (50). In the miRNA sequencing dataset (SRR866605) of salmon fish heart, canonical miRNA read *ATCCCGGA CGAGCCCCCAA* co-exists with five isoforms having read counts 9, 1489, 16, 4 or 4; There are also three singleton reads having an editing distance with this canonical miRNA (Table 5). Our miREC grouped all of these reads and detected some of them as erroneous reads. After error correction, the abundance of the canonical miRNA increased from 18 copies to 31; the first three isoforms' abundance increased from 9 to 17, from 16 to 20 and from 1489 to 1513. The abundance recovery of the canonical miRNA is owned to the erroneous base correction of the 11 reads listed in the last five rows of Table 5.

The performance by the Karect method shows that only one of the eleven erroneous reads was corrected. Only the first and second miRNA sequence (Table 5) have different read counts after Karect's correction. The copy count of the first read was increased by 1 and the copy count of the second read was increased by 20, missing lots of corrections.

A more interesting point of the error correction is that the 11 erroneous reads of the canonical miRNA contain not only substitutions, but deletion and insertion errors distributed at multiple base positions. In particular, more than one third of erroneous bases happened at the seed region, important for gene target binding analysis.

Case study 3: upside down change in differential expression analysis. Analysis on tissue-specific uniquely expressed or top-ranked differentially expressed miRNAs in a specific tissue or at a disease stage is very sensitive to the sequencing data quality (9). Some uniquely expressed miRNAs can be identified only after error correction.

In our differential expression analysis between the salmon heart and brain tissues (SRR866605 vs SRR866611), we found that 5,675 miRNA did not co-exist in the two datasets, and the number of common miRNAs was reduced from 16 443 to 10 768 after error correction. For example, read *TGAGGTAGTTGGTTGTATGGTG* (mapped to ssa-let-7d-5p in miRBase), had four copies in the heart dataset and 26 copies in the brain dataset before correction, while its read count was changed to zero in the heart dataset and changed to 30 in the brain dataset after error correction. Two more examples: read *CTTTCAGTCGGATGTTTGCACCA* (mapped to ssa-miR-30d-3p in miRBase) had 152 copies in the heart dataset and two copies in the brain data before correction, while its quantity was changed to 155 in the heart dataset and to zero in the brain dataset. Another read *TTGCATAGTCACAAAAATGATC* (mapped to ssa-miR-153a-3p in miRBase) had three copies in the heart dataset and 14 434 copies in the brain dataset before correction, while the quantity dropped to zero in the heart dataset but increased to 14,498 in the brain dataset after error correction.

Top-rank differentially expressed miRNAs can become low-ranked ones, and vice versa after error correction. The reason is that the expression folds of miRNAs between two tissue types or between two disease stages are sensitive to the copy counts after erroneous reads are corrected in the two classes. We compared the expression folds of common miRNAs between the salmon heart tissue and brain tissue before and after our error correction. Table 6 presents the list of 10 miRNAs whose expression folds between the two tissues are top-ranked after the error correction, in comparison with their ranking positions before the error correction. The two ranking lists are quite different. For example, the rank of ssa-miR-9a-5p (*TCTTTGGTTATCTAGCTGTA*) is reverted from rank 8 to 4. Furthermore, the originally

Table 7. Ranking position change of tissue-specific miRNAs in the heart tissue (vs the liver tissue) before and after error correction

miRNA sequence	Rank after correction	Rank before correction	Read count before correction	Read count after correction	Read count increase
TTAAGACTTGTAGTGATGTTT	1	out of scope	47 546	47 583	37
TGGAATGTAAAGAAGTATGTAT	2	1	12 650	12 728	78
TTTGGTCCCCTTCAACCAGCTG	3	2	4954	4985	31
TTGGTCCCCTTCAACCAGCTG	4	3	2522	2541	19
TTAAGACTTGCAGTGATGTT	5	4	1665	1677	12
ACAGCTCATCCATTGGTC	6	5	1174	1188	14
TGGAATGTAAAGAAGTATGTA	7	6	879	892	13
AACATCACTTTAAGTCTCTGCT	8	7	876	892	16
TTGGTCCCCTTCAACCAGCTGTA	9	8	835	856	21
TGAGGTAGTTGGTTGTATTGTTT	10	Out of scope	780	791	11
TGGACGGAGAACTGATAAGGG	11	9	693	702	9
TTAAGACTTGTAGTGATGTTTAA	12	10	685	698	13
TGAGGTAGTTGGTTGTATTGT	13	Out of scope	659	666	7
TAAAGGGAATTTGCGACTGTTA	14	11	622	635	13
TGGAATGTAAAGAAGTATGTATT	15	12	616	629	13

top-ranked number-1, number-4 and number-9 miRNAs are all dropped below rank-10 after error correction.

In detail, the original top-one miRNA (*TCTTTGGTTATCTAGCTGTATGT*) had 16 776 copies in the brain tissue. However, the corrected top-one miRNA is *TCTTTGGTTATCTAGCTGTATG*, whose copy count is 48 092 in the brain tissue after error correction. It is interesting to note that:

- The two miRNAs only have one base difference at the 3′ end. The corrected top-one miRNA after error correction has one base trimmed at the 3′ end, compared to the original top-one ranked miRNA. The two miRNAs can be considered as 3′ end trimmed/addition isoforms each other.
- The original top-ranked miRNA and the corrected top-one miRNA have a huge abundance difference (31 316 copies = 48 092 − 16 776) in the brain tissue. One is extremely high-level expressed; the other is median-level expressed. This suggests that we would concentrate on wrong top-ranked miRNA biomarkers if the sequencing reads had not been cleaned by good error correction algorithms.

New top-ranked tissue-specific miRNAs (or called no-presence miRNAs or tissue- and disease-subtype dependent miRNAs by (9)) were found in the heart tissue (SRR866605) after error correction when the liver tissue (SRR866579) was compared. Table 7 presents two rankings of top-15 miRNAs specifically expressed in the heart tissue before and after error correction. Without our correction, the top-1, top-10 and top-13 tissue-specific miRNAs in salmon heart would be not detected because erroneous reads which are identical with these reads also exist in the liver tissue. Moreover, after our error correction, the quantity of the top-ranked miRNAs increases. These recovered read counts and accurate abundance measurement would help make more convincing conclusions in the down stream analysis.

Case study 4: class-specific miRNAs and rare-miRNA analysis for human miRNA sequencing datasets. Ranking positions of class-specific miRNAs and rare miRNA quantity recovery analysis are also conducted on human miRNA sequencing datasets (acquired from beta cells and brain samples).

The human beta cells were incubated with solution of low glucose or high glucose. It's expected to reveal novel differentially expressed miRNAs between these two classes. We found that the number of distinct reads decreased by 8.85% from 5 803 166 to 5 289 466 in the low glucose solution cell, and reduced by 12.44% from 1 856 318 to 1 625 453 in the high glucose solution cell after error correction. For the top-ranked differentially expressed miRNAs between the two datasets, only slight rank changes were observed (Table 8). Some of the top-ranked miRNAs were just swapped ranking positions within top 10 after error correction. The copy counts of these top-ranked miRNAs all had small increases after the error correction. Note that these changes on glucose-level specific miRNAs in human beta cells after error correction is not as big as those changes made in the salmon heart-head tissue pair comparison by our error correction.

However, such big changes on age-specific miRNAs in brain samples can be observed again when we compared between miRNA sequencing reads of an Alzheimer's disease patient aged 75 and a patient aged 94. The number of distinct miRNA reads decreased by 33.6% from 361 039 to 239 667 in the patient aged 75, and decreased by 16.1% from 635 169 to 532 708 in the patient aged 94, after error correction. Table 9 provides two rankings of top-10 age-specific miRNAs expressed only in the patient aged 94 before and after correction. New top-ranked age-specific miRNAs were identified in the patient aged 94. Without our error correction, top-1, 2, 3, 4, 6, 8 and 10 age-specific miRNAs would not be detected because erroneous reads which are identical with these reads also exist in the patient aged 75, with copy counts 3, 2, 2, 3, 4, 2 and 4 respectively.

Discovery of rare miRNAs is of strong interests. We examined the read counts of low-expression miRNAs (or rare miRNAs) before and after error correction in the Alzheimer's disease patient aged 94. Note that all these rare miRNAs here are defined to have no expression in the patient aged 75. Table 10 shows top-10 read-count greatly-changed rare miRNAs before and after error correction. It suggests that the read counts of these rare miRNAs were all enhanced by about 2 or 3-fold after error correction.

Table 8. Read count changes and ranking changes of top-10 differentially expressed miRNAs in the high glucose incubated human beta cell after error correction, and those in the low glucose incubated human beta cell after error correction

miRNA sequence **Among high glucose**	Rank After	Rank Before	Read count Before	Read count After	Read count Increase
GTGGGGCCACGAGCTGAGTGCGT	1	1	86	92	6
AGCAGGGTCGGGCCTGGTTAGTA	2	2	68	69	1
GAGTTCGCGCTTTCCCCT	**4**	**3**	61	69	8
GCCGCAGGTGCAGATCTTGGTGG	**3**	**4**	62	65	3
TCGGGCCTGGTTAGTACTTGGAT	5	5	60	62	2
GTGGAGCCTGCGGCTTAAT	6	6	60	61	1
TCGGAAGCTAAGCAGGGTCGGGCC	7	7	57	57	0
GGCTCAGCGTGTGCCTACC	8	9	55	56	1
GTCTACGGCCCTACCACCCTGAACG	9	8	54	55	1
GTCGGGCCTGGTTAGTACTTGGA	10	10	51	54	3

miRNA Sequence **among low glucose**	Rank After	Rank Before	Read Count Before	Read Count After	Read Count Increase
CGCCCGTCCCCGCCCCTT	1	1	640	651	11
TAGGGGTATGATTCTCGCTTCGG	2	2	582	586	4
GCCCGTCCCCGCCCCTT	3	3	565	572	7
CGCCCGTCCCCGCCCCTTGCC	4	4	484	489	5
CCAGTGGTTGTCGACTTGCG	5	5	428	436	8
CTCAGGTGCCCGAGGCCGAA	6	6	371	376	5
AAGACGGAGAGGGAAAGAG	7	7	317	324	7
ACGGGGAGGGCGGCGCCGCCGCC	**8**	**9**	292	300	8
TTCGGCTGAGTTCGTGATGGATTTG	**9**	**8**	297	298	1
CCACCGCCCGTCCCCGCCCCTTG	10	10	272	278	6

Table 9. Ranking position changes of age-specific miRNAs in brain tissue from an Alzheimer male patient aged 94 (vs a patient aged 75) before and after error correction

miRNA sequence	Rank after correction	Rank before correction	Read count before correction	Read count after correction	Read count increase	Original read count (aged 75)
TTTCTCACTACTGCACTTGACTAGTC	1	Out of scope	1416	1477	61	3
TTTCTCACTACTGCACTTGACC	2	Out of scope	839	877	38	2
TTTCTCACTACTGCACTTGAC	3	Out of scope	817	857	40	2
TTTCTCACTACTGCACTTGACTAG	4	Out of scope	726	761	35	3
TTTCTCACTACTGCACTTGACA	5	1	719	746	27	0
TTTCTCACTACTGCACTTGACTAGT	6	Out of scope	593	628	35	4
TGAGGTAGTACGTTGTATAGT	7	2	521	533	12	0
TTTCTCACTACTGCACTTGACTA	8	Out of scope	381	410	29	2
TGAGGAAGTAGGTTGTAGAGTT	9	3	365	375	10	0
TGAGGTAGTACATTGTATAGT	10	Out of scope	339	352	13	4

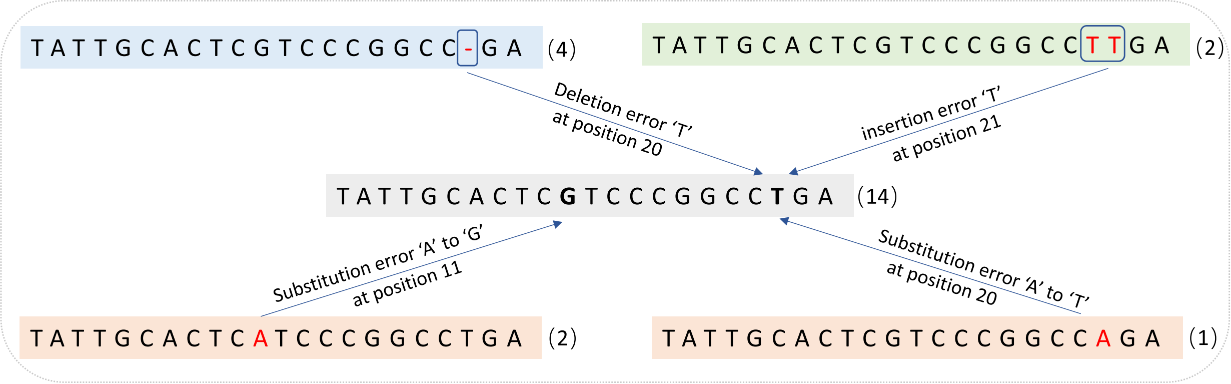
TATTGCACTCGTCCCGGCC-GA (4)
TATTGCACTCGTCCCGGCCTTGA (2)
Deletion error 'T' at position 20
insertion error 'T' at position 21
TATTGCACTCGTCCCGGCCTGA (14)
Substitution error 'A' to 'G' at position 11
Substitution error 'A' to 'T' at position 20
TATTGCACTCATCCCGGCCTGA (2)
TATTGCACTCGTCCCGGCCAGA (1)

Figure 6. A rare miRNA in the Alzheimer's disease patient aged 94 showing significant copy count change after error correction.

Table 10. Copy count enhancement of 10 rare miRNAs after error correction in a human brain dataset acquired from an Alzheimer's disease patient aged 94

miRNA sequence	Read count_before	Read count_after
TCATTGGTTATCTAGCTGTATGC	6	18
TAGAACTTCGTCGAGTACGCTC	9	26
AAAAGCTGGGTTGAGAGGGCGTGA	6	17
AGCAGGACGGTGGCCATGGA	8	22
TGAGGCAGTAGGTTGTGTGGTTAT	6	16
TCCAGCATCAGTGATTTTGTTGT	6	16
TCACAGACAGCCGGTCTCTTTT	6	16
GTTGGTCCGAGTGTTGTGGGC	6	16
TCCCCGGCATCTCCACCAT	9	23
AGGAGATGGAATAGGAGCTTGA	8	20

Notes: _after indicates after error correction, while _before indicates before error correction

Figure 6 depicts how the read quantity of a rare miRNA is enhanced from 14 copies to 23 in the correction process. The corrections were involved with four types of erroneous reads: four reads with a deletion error (labeled in blue), two reads with an insertion error (labeled in green), one read with a substitution error from A to G at position 11 and two reads with a substitution error from A to T at position 20 (labeled in orange). Our miREC can detect all of these erroneous reads and corrected them to recover this rare miRNA's quantity.

Verification results on the sequencing reads of the 963 miRXplore universal reference miRNAs (pure control and spike-in)

Our algorithm was tested on the sequencing reads of an equimolar mixture of synthetic miRNAs from the miRXplore Universal Reference that consists of 963 miRNAs from human, mouse, rat and viral sources (three replicate samples miRXploreUR rep1-3 corresponding to GSE139936.GSM4149813, GSE139936.GSM4149814 and GSE139936.GSM4149815 (51)). The test was to verify

- whether our detected erroneous reads can be each corrected into one of the 963 miRNA sequences, and
- whether any new sequences are introduced into the read dataset after the correction.

An ideal performance should be: every error-corrected read is turned to be an exact copy of one of the 963 miRNA sequences, and previously non-existing reads are never created by the correction step.

The correction performance by miREC in comparison with Karect (the best literature method (41)) are shown in Table 11. On the sequencing dataset named D18-6962_1 of GSE139936.GSM4149813, our algorithm detected a total of 43362 errors. After correction, the correspondingly rectified reads were each exactly matched with one of the 963 miRNA sequences. The total read count of the 963 miRNAs was therefore increased by about 19.59% (see Supplementary file S1 for details). During this correction step, previously non-existing reads were never generated/created. In fact, the number of distinct reads was decreased from 259867 to 212093. On the other hand, almost all (99.22%) of the remaining unchanged 231792 distinct reads were not considered as the erroneous reads of the 963 miRNAs by our algorithm. This is reasonable because each of them has a minimum editing distance of 2 or bigger with any of the 963 miRNA sequences. These remaining reads also have an extremely low counts such as 1, 2 or 3. They can be considered as noisy reads which may be caused by the library preparation noise or contaminates.

Karect detected total 127 642 errors, but only 18 225 of them were corrected into the sequencing reads of the 963 miRNAs, increasing their read counts by 8.22% in total. Meanwhile, the other base modifications have introduced a pool of 37 678 new sequences which did not exist in the dataset before Karect's correction.

From these comparisons, we note that our algorithm miREC has corrected almost all of those reads which should be rectified and that miREC has never introduced previously non-existing reads. This is true for all other datasets listed in Table 11. However, Karect introduced large pools of new reads which have never existed in the original reads set; also Karect corrected less than half of those reads which should be rectified.

On a spike-in sample of the 963 miRNAs at human cells (GSE159434.D19-10246.assembled.fastq (52)), our algorithm detected 89 301 erroneous reads of the 963 miRNAs. After correction, their read counts increased by 15.66% in total. The algorithm did not generate any previously non-existing reads, but decreased the number of distinct reads by 45 189, greatly diminishing the uncerternty/entropy of the data set. On the other hand, Karect detected and corrected 15885 erroneous reads of the 963 miRNAs, making their read counts increased by 7.62% in total. However, Karect created 14 462 new reads which were non-existing previously.

These comparative results on both the control and spike-in sample demonstrate that our modified reads are genuine correction and that our algorithms do not generate any previous non-existing reads after the correction process.

DISCUSSION

There are several aspects of complexities in miRNA sequencing datasets which can limit the performance of error correction algorithms. For example, some miRNAs (A versus B) can be very similar (just one base different) and have similar abundance in the sample. In this case, erroneous reads of A can be sometimes exactly the same sequence as B (or, erroneous reads of B can be sometimes exactly the same sequence as A). Such an error is unable to be detected by any error correction algorithm because A and B have similar abundance level (see examples in Section - Correction performance on wet-lab miRNA sequencing datasets injected with small numbers of artificial errors). However, we note that such sequencing errors (without correction) would not affect much about the true read counts of A and B. The reason is that the count of A's such erroneous reads should be at the same level of the count for B's such erroneous reads.

Some miRNAs have very low-level abundance and are prone of errors in sequencing. In this case, the read count of such a miRNA can be zero, meaning all of its reads are wrongly sequenced (see examples in the 'Editing distance = 1' column of Table 11). Our algorithm is unable to detect

Table 11. Correction performance by miREC[a] in comparison with Karect[b] (41) on the sequencing reads of the 963 miRXplore Universal Reference miRNAs (pure control and spike-in)

							Distinct reads with the same frequencies before and after correction								
			Total read count of the 963 miRNAs					Minimum editing distance ##							
Dataset[c](Total		Number of detected	before	after	Pct(%)	Introduced		## =0		## =1		## =2		## >=3	
read count)	Method	bases for correction	correction	correction	increased	new sequences	Total count	Count	Pct(%)	Count	Pct(%)	Count	Pct(%)	Count	Pct(%)
D18-6962_1	miREC	43 362	221 657	265 076	19.59	0	231 792	24	0.0104	1 809	0.7804	47 949	20.6862	182 010	78.5230
(544 056)	Karect	127 642		239 882	8.22	37 678	162 730	10	0.0061	11 031	6.7787	21 538	13.2354	130 151	79.9797
D18-6962_2	miREC	43 122	223 921	267 094	19.28	0	233 418	23	0.0099	1 813	0.7767	47 832	20.492	183 750	78.7214
(547087)	Karect	129 410		241 979	8.06	38 501	163 525	8	0.0049	11 095	6.7849	21 609	13.2145	130 813	79.9957
D18-6963_1	miREC	35 211	150 886	186 126	23.36	0	181 380	35	0.0193	1 310	0.7222	34 202	18.8565	145 833	80.4019
(402349)	Karect	108 583		167 678	11.13	30 885	124 456	11	0.0088	8 058	6.4746	13 726	11.0288	102 661	82.4878
D18-6963_2	miREC	35 006	152 201	187 237	23.02	0	180 491	35	0.0194	1 301	0.7219	34 154	18.9228	144 999	80.3359
(401407)	Karect	109 344		168 894	10.97	31 159	123 387	11	0.0089	8 040	6.5161	13 704	11.1065	101 632	82.3685
D18-6964_1	miREC	39 369	192 095	231 512	20.52	0	215 161	25	0.0116	1 372	0.6377	39 914	18.5508	173 850	80.8000
(490577)	Karect	113 706		209 543	9.08	32 412	154 167	12	0.0078	9 488	6.1544	17 004	11.0296	127 663	82.8083
D18-6964_2	miREC	39 210	194 177	233 441	20.22	0	213 499	26	0.0122	1 401	0.6562	39 802	18.6427	172 270	80.6889
(488317)	Karect	115 845		211 474	8.91	33 398	151 458	11	0.0073	9 513	6.2809	17 069	11.2698	124 865	88.4420
D19-10246	miREC	89 301	208 219	240 828	15.66	0	77 335	16	0.0207	1 782	2.3043	11 129	14.3906	64 408	83.2844
(767426)	Karect	85 491		224 089	7.62	14 462	78 927	7	0.0089	9 750	12.3532	14 004	17.7430	55 166	69.8950

[a] The parameter of kmer range of miREC: [8 25].

[b] The parameters of Karect: -matchtype=hamming -celltype=haploid.

[c] D18-6962_1 (180719Ded_D18-6962_1_sequence.fastq) and D18-6962_2 (180719Ded_D18-6962_2_sequence.fastq) from GSE139936.GSM4149813;
D18-6963_1 (180719Ded_D18-6963_1_sequence.fastq) and D18-6963_2 (180719Ded_D18-6963_2_sequence.fastq) from GSE139936.GSM4149814;
D18-6964_1 (180719Ded_D18-6964_1_sequence.fastq) and D18-6964_2 (180719Ded_D18-6964_2_sequence.fastq) from GSE139936.GSM4149815;
D19-10246 (D19-10246.assembled.2NN.fastq) from GEO accession GSE159434.

these errors for correction. If the read count of a miRNA is not zero but only 1, 2 or 3, again our algorithm is unable to detect the erroneous reads of this miRNA for correction (see examples in the 'Editing distance = 0' column of Table 11).

Although having these challenges, in this work, we have proposed an effective miRNA sequencing error correction method named miREC, which is the first tool to address the error correction problem in the area. The novelty of the method is a 3-layer k-mer–(k + 1)mer–(k – 1)mer lattice structure to hold the kmer's supersets and subsets' frequency differences which underline the locations of the errors and the correcting templates. Our miREC has showed excellent performance to rectify not only substitution errors but also indel errors at both simulated and real miRNA sequencing datasets. The experiments conducted with different running parameters showed that the miREC is insensitive to datasets and it has good robustness to guarantee high-quality correction performance. Our error correction performance have been also verified on the control and spike-in sequencing datasets of the 963 synthetic miRNAs from the miRXplore Universal Reference. With the precise aberration correction and free of new error introduction, we are able to conduct ultrafine analysis on miRNA sequencing data at the single base resolution. The method is immediately applicable to miRNA sequencing datasets from the fields of plant biology and cancer biology which are worth future investigation in detail.

DATA AVAILABILITY

The computational programs and datasets used in this work are available at https://github.com/XuanrZhang/miREC.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Professor Peter C. Dedon of MIT and his team members for sending us the latest versions of GSE139936 and GSE159434. We would also like to thank Dr Chaowang Lan and Dr Yuansheng Liu for discussions and expertise on preprocessing datasets of genomic reads.

FUNDING

Australian Research Council [DP 180100120]. Funding for open access charge: Australian Research Council [DP 180100120].

Conflict of interest statement. None declared.

REFERENCES

1. Goodwin,S., McPherson,J.D. and McCombie,W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333.
2. Yeung,C.L.A., Co,N.-N., Tsuruga,T., Yeung,T.-L., Kwan,S.-Y., Leung,C.S., Li,Y., Lu,E.S., Kwan,K., Wong,K.-K. *et al.* (2016) Exosomal transfer of stroma-derived miR21 confers paclitaxel resistance in ovarian cancer cells through targeting APAF1. *Nat. Commun.*, **7**, 11150.
3. Xiao,Y. and MacRae,I.J. (2019) Toward a comprehensive view of microRNA biology. *Mol. Cell*, **75**, 666–668.
4. Tan,G.C., Chan,E., Molnar,A., Sarkar,R., Alexieva,D., Isa,I.M., Robinson,S., Zhang,S., Ellis,P., Langford,C.F. *et al.* (2014) 5 isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res.*, **42**, 9424–9435.
5. Trontti,K., Väänänen,J., Sipilä,T., Greco,D. and Hovatta,I. (2018) Strong conservation of inbred mouse strain microRNA loci but broad variation in brain microRNAs due to RNA editing and isomiR expression. *RNA*, **24**, 643–655.
6. Fernandez-Valverde,S.L., Taft,R.J. and Mattick,J.S. (2010) Dynamic isomiR regulation in Drosophila development. *RNA*, **16**, 1881–1888.
7. Meng,L., Liu,C., Lü,J., Zhao,Q., Deng,S., Wang,G., Qiao,J., Zhang,C., Zhen,L., Lu,Y. *et al.* (2017) Small RNA zippers lock miRNA molecules and block miRNA function in mammalian cells. *Nat. Commun.*, **8**, 13964.
8. Liu,H., Lei,C., He,Q., Pan,Z., Xiao,D. and Tao,Y. (2018) Nuclear functions of mammalian MicroRNAs in gene regulation, immunity and cancer. *Mol. Cancer*, **17**, 64.
9. Telonis,A.G., Magee,R., Loher,P., Chervoneva,I., Londin,E. and Rigoutsos,I. (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.
10. Dutta,R.K., Chinnapaiyan,S. and Unwalla,H. (2019) Aberrant micrornaomics in pulmonary complications: Implications in lung health and diseases. *Mol. Ther.-Nucl. Acids*, **18**, 413–431.
11. Dai,F.-Q., Li,C.-R., Fan,X.-Q., Tan,L., Wang,R.-T. and Jin,H. (2019) miR-150-5p inhibits non-small-cell lung cancer metastasis and recurrence by targeting HMGA2 and β-catenin signaling. *Mol. Ther.-Nucl. Acids*, **16**, 675–685.
12. Pisignano,G., Napoli,S., Magistri,M., Mapelli,S.N., Pastori,C., Di Marco,S., Civenni,G., Albino,D., Enriquez,C., Allegrini,S. *et al.* (2017) A promoter-proximal transcript targeted by genetic polymorphism controls E-cadherin silencing in human cancers. *Nat. Commun.*, **8**, 15622.
13. Yang,A., Shao,T.-J., Bofill-De Ros,X., Lian,C., Villanueva,P., Dai,L. and Gu,S. (2020) AGO-bound mature miRNAs are oligouridylated by TUTs and subsequently degraded by DIS3L2. *Nat. Commun.*, **11**, 2765.
14. Liu,C.-H., Wang,Z., Huang,S., Sun,Y. and Chen,J. (2019) MicroRNA-145 regulates pathological retinal angiogenesis by suppression of TMOD3. *Mol. Ther.-Nucl. Acids*, **16**, 335–347.
15. Liao,Z., Li,D., Wang,X., Li,L. and Zou,Q. (2018) Cancer diagnosis through IsomiR expression with machine learning method. *Curr. Bioinform.*, **13**, 57–63.
16. Liu,H.-P., Lai,H.-M. and Guo,Z. (2021) Prostate cancer early diagnosis: circulating microRNA pairs potentially beyond single microRNAs upon 1231 serum samples. *Brief. Bioinform.*, **22**, bbaa111.
17. Bilanges,B., Posor,Y. and Vanhaesebroeck,B. (2019) PI3K isoforms in cell signalling and vesicle trafficking. *Nat. Rev. Mol. Cell. Biol.*, **20**, 515–534.
18. Sänger,L., Bender,J., Rostowski,K., Golbik,R., Lilie,H., Schmidt,C., Behrens,S.-E. and Friedrich,S. (2021) Alternatively spliced isoforms of AUF1 regulate a miRNA-mRNA interaction differentially through their YGG motif. *RNA Biol.*, **18**, 843–853.
19. Pillman,K.A., Goodall,G.J., Bracken,C.P. and Gantier,M.P. (2019) miRNA length variation during macrophage stimulation confounds the interpretation of results: implications for miRNA quantification by RT-qPCR. *RNA*, **25**, 232–238.
20. Hoefer,I.E. (2020) Isolating functional (iso) miRNA targets during ischemia. *Mol. Ther.*, **28**, 7–8.
21. Neilsen,C.T., Goodall,G.J. and Bracken,C.P. (2012) IsomiRs—the overlooked repertoire in the dynamic microRNAome. *Trends Genet.*, **28**, 544–549.
22. Lan,C., Peng,H., McGowan,E.M., Hutvagner,G. and Li,J. (2018) An isomiR expression panel based novel breast cancer classification approach using improved mutual information. *BMC Med. Genomics*, **11**, 118.
23. Salk,J.J., Schmitt,M.W. and Loeb,L.A. (2018) Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.*, **19**, 269.

24. Laehnemann,D., Borkhardt,A. and McHardy,A.C. (2016) Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Brief. Bioinform.*, **17**, 154–179.
25. Mardis,E.R. (2013) Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.*, **6**, 287–303.
26. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *cell*, **116**, 281–297.
27. Chekulaeva,M. and Filipowicz,W. (2009) Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells. *Curr. Opin. Cell Biol.*, **21**, 452–460.
28. Yu,F., Pillman,K.A., Neilsen,C.T., Toubia,J., Lawrence,D.M., Tsykin,A., Gantier,M.P., Callen,D.F., Goodall,G.J. and Bracken,C.P. (2017) Naturally existing isoforms of miR-222 have distinct functions. *Nucleic Acids Res.*, **45**, 11371–11385.
29. Telonis,A.G. and Rigoutsos,I. (2018) Race disparities in the contribution of miRNA isoforms and tRNA-derived fragments to triple-negative breast cancer. *Cancer Res.*, **78**, 1140–1154.
30. van der Kwast,R.V., Woudenberg,T., Quax,P.H. and Nossent,A.Y. (2020) MicroRNA-411 and Its 5-IsomiR have distinct targets and functions and are differentially regulated in the vasculature under ischemia. *Mol. Ther.*, **28**, 157–170.
31. Cloonan,N., Wani,S., Xu,Q., Gu,J., Lea,K., Heater,S., Barbacioru,C., Steptoe,A.L., Martin,H.C., Nourbakhsh,E. *et al.* (2011) MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome biol.*, **12**, R126.
32. Mullany,L.E., Herrick,J.S., Wolff,R.K. and Slattery,M.L. (2016) MicroRNA seed region length impact on target messenger RNA expression and survival in colorectal cancer. *PloS one*, **11**, e0154177.
33. Guo,L. and Chen,F. (2014) A challenge for miRNA: multiple isomiRs in miRNAomics. *Gene*, **544**, 1–7.
34. Ebhardt,H.A., Tsang,H.H., Dai,D.C., Liu,Y., Bostan,B. and Fahlman,R.P. (2009) Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res.*, **37**, 2461–2470.
35. Limasset,A., Flot,J.-F. and Peterlongo,P. (2020) Toward perfect reads: self-correction of short reads via mapping on de Bruijn graphs. *Bioinformatics*, **36**, 1374–1381.
36. Li,H. (2015) BFC: correcting Illumina sequencing errors. *Bioinformatics*, **31**, 2885–2887.
37. Sheikhizadeh,S. and de Ridder,D. (2015) ACE: accurate correction of errors using K-mer tries. *Bioinformatics*, **31**, 3216–3218.
38. Heo,Y., Wu,X.-L., Chen,D., Ma,J. and Hwu,W.-M. (2014) BLESS: bloom filter-based error correction solution for high-throughput sequencing reads. *Bioinformatics*, **30**, 1354–1362.
39. Salmela,L. and Schröder,J. (2011) Correcting errors in short reads by multiple alignments. *Bioinformatics*, **27**, 1455–1461.
40. Kao,W.-C., Chan,A.H. and Song,Y.S. (2011) ECHO: a reference-free short-read error correction algorithm. *Genome Res.*, **21**, 1181–1192.
41. Allam,A., Kalnis,P. and Solovyev,V. (2015) Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data. *Bioinformatics*, **31**, 3421–3428.
42. Le,H.-S., Schulz,M.H., McCauley,B.M., Hinman,V.F. and Bar-Joseph,Z. (2013) Probabilistic error correction for RNA sequencing. *Nucleic Acids Res.*, **41**, e109.
43. Song,L. and Florea,L. (2015) Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience*, **4**, 48.
44. Kokot,M., Długosz,M. and Deorowicz,S. (2017) KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, **33**, 2759–2761.
45. Seppey,M., Manni,M. and Zdobnov,E.M. (2020) LEMMI: a continuous benchmarking platform for metagenomics classifiers. *Genome Res.*, **30**, 1208–1216.
46. Woldemariam,N.T., Agafonov,O., Høyheim,B., Houston,R.D., Taggart,J.B. and Andreassen,R. (2019) Expanding the miRNA repertoire in Atlantic salmon; discovery of isomiRs and miRNAs highly expressed in different tissues and developmental stages. *Cells*, **8**, 42.
47. Andreassen,R., Worren,M.M. and Høyheim,B. (2013) Discovery and characterization of miRNA genes in Atlantic salmon (Salmo salar) by use of a deep sequencing approach. *BMC Genomics*, **14**, 482.
48. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, **17**, 10–12.
49. Stelzer,G., Rosen,N., Plaschkes,I., Zimmerman,S., Twik,M., Fishilevich,S., Stein,T.I., Nudel,R., Lieder,I., Mazor,Y. *et al.* (2016) The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics*, **54**, 1.30.1–1.30.33.
50. Martí,E., Pantano,L., Bañez-Coronel,M., Llorens,F., Miñones-Moyano,E., Porta,S., Sumoy,L., Ferrer,I. and Estivill,X. (2010) A myriad of miRNA variants in control and Huntington—s disease brain regions detected by massively parallel sequencing. *Nucleic Acids Res.*, **38**, 7219–7235.
51. Hu,J.F., Yim,D., Ma,D., Huber,S.M., Davis,N., Bacusmo,J.M., Vermeulen,S., Zhou,J., Begley,T.J., DeMott,M.S. *et al.* (2021) Quantitative mapping of the cellular small RNA landscape with AQRNA-seq. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-021-00874-y.
52. Hu,J.F., Yim,D., Huber,S.M., Bacusmo,J.M., Ma,D., DeMott,M.S., Levine,S.S., de Crécy-Lagard,V., Dedon,P.C. and Cao,B. (2019) Sequencing-based quantitative mapping of the cellular small RNA landscape. bioRxiv doi: https://doi.org/10.1101/841130, 26 November 2019, preprint: not peer reviewed.