SOFTWARE

FindAdapt: A python package for fast and accurate adapter detection in small RNA sequencing

Hua-Chang Chen[1,2], **Jing Wang**[1,2], **Yu Shyr**[1,2]*, **Qi Liu**[1,2]*

[1] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, [2] Center for Quantitative Sciences, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

* yu.shyr@vumc.org (YS); qi.liu@vumc.org (QL)

Abstract

Adapter trimming is an essential step for analyzing small RNA sequencing data, where reads are generally longer than target RNAs ranging from 18 to 30 bp. Most adapter trimming tools require adapter information as input. However, adapter information is hard to access, specified incorrectly, or not provided with publicly available datasets, hampering their reproducibility and reusability. Manual identification of adapter patterns from raw reads is labor-intensive and error-prone. Moreover, the use of randomized adapters to reduce ligation biases during library preparation makes adapter detection even more challenging. Here, we present FindAdapt, a Python package for fast and accurate detection of adapter patterns without relying on prior information. We demonstrated that FindAdapt was far superior to existing approaches. It identified adapters successfully in 180 simulation datasets with diverse read structures and 3,184 real datasets covering a variety of commercial and customized small RNA library preparation kits. FindAdapt is stand-alone software that can be easily integrated into small RNA sequencing analysis pipelines.

Introduction

Small RNA sequencing (small RNA-seq) is commonly used to profile the full spectrum of small RNAs, including microRNA (miRNA), Piwi-interacting RNA (piRNA), small interfering RNA (siRNA), tRNA, rRNA, and small nuclear RNA (snRNA) [1–7]. Small RNAs, typically 18 to 30 bp long, are generally shorter than sequencing reads, resulting in adapters included in raw reads. Therefore, the first essential step for small RNA-seq analysis is adapter trimming and recovery of true sequences [1,8,9].

There are many tools developed to trim adapters that generally require adapter information as input, such as Cutadapt [10], FASTX [11], AdapterRemoval [12], and Trimmomatic [13]. However, adapter information is not mandatory when data are deposited into public databases like the NCBI Gene Expression Omnibus (GEO). A recent study reported that around 53% of NCBI Sequence Read Archive (SRA) entries lack adapter-related information [14]. Even when adapter sequences are provided, it is often difficult to extract them from the record automatically and streamline the analysis. Manually retrieving adapter sequences from literature or raw

Citation: Chen H-C, Wang J, Shyr Y, Liu Q (2024) FindAdapt: A python package for fast and accurate adapter detection in small RNA sequencing. PLoS Comput Biol 20(1): e1011786. https://doi.org/10.1371/journal.pcbi.1011786

Editor: Sushmita Roy, University of Wisconsin, Madison, UNITED STATES

Received: June 16, 2023

Accepted: December 26, 2023

Published: January 22, 2024

Data Availability Statement: FindAdapt is freely available at https://github.com/chc-code/FindAdapt.

Funding: This work is supported by National Cancer Institute grants (U2C CA233291 to Y.S. and Q.L., P01CA229123 to Y.S. and Q.L., and U54 CA274367 to Q.L.), National Institutes of Health (P01 AI139449 to Q.L.), and Cancer Center Support Grant (P30CA068485 to Y.S.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

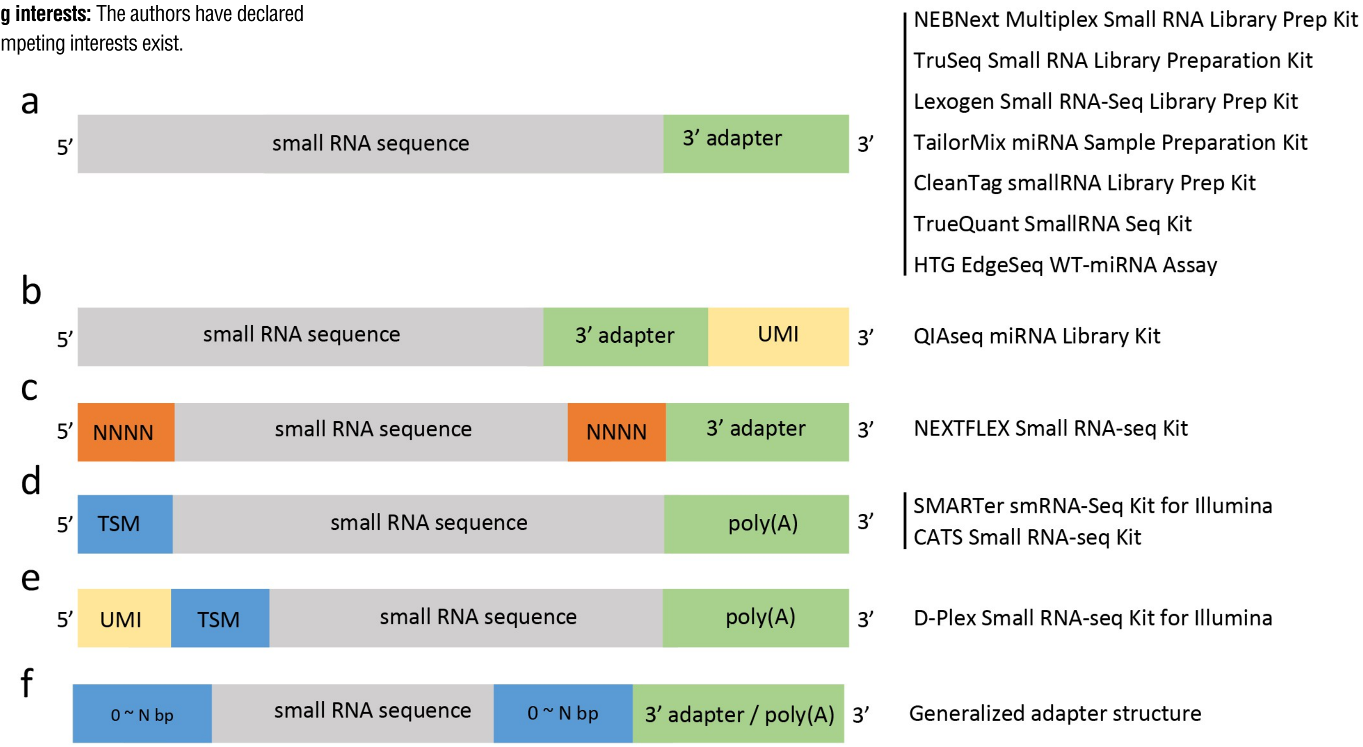

a
5′
small RNA sequence
3′ adapter
3′
NEBNext Multiplex Small RNA Library Prep Kit
TruSeq Small RNA Library Preparation Kit
Lexogen Small RNA-Seq Library Prep Kit
TailorMix miRNA Sample Preparation Kit
CleanTag smallRNA Library Prep Kit
TrueQuant SmallRNA Seq Kit
HTG EdgeSeq WT-miRNA Assay
b
5′
small RNA sequence
3′ adapter
UMI
3′
QIAseq miRNA Library Kit
c
5′
NNNN
small RNA sequence
NNNN
3′ adapter
3′
NEXTFLEX Small RNA-seq Kit
d
5′
TSM
small RNA sequence
poly(A)
3′
SMARTer smRNA-Seq Kit for Illumina
CATS Small RNA-seq Kit
e
5′
UMI
TSM
small RNA sequence
poly(A)
3′
D-Plex Small RNA-seq Kit for Illumina
f
0 ~ N bp
small RNA sequence
0 ~ N bp
3′ adapter / poly(A)
3′
Generalized adapter structure

Fig 1. Read structures from different small RNA library preparation kits. (a) The 3' adapter right after the true small RNA sequence by seven Kits. (b) The true small RNA sequence followed by the 3' adapter and the UMI of 12bp by the QIAseq Kit. (c) The true small RNAs sequence surrounded by random tetramer (NNNN) at both 5' and 3' ends, and the 3' adapter sequence after the 3' random tetramer by the NEXTFLEX Kit. (d) The true small RNA sequence surrounded by poly(A) tail at the 3' end and a TSM (template switching motif) at the 5' end for the SMARTer and CATS Kits. e) The UMI sequence of 12bp at the 5' before the TSM by the D-Plex Kit. f) The generalized read structure from all kits.

https://doi.org/10.1371/journal.pcbi.1011786.g001

reads is labor-intensive and error-prone, especially when there are many datasets to be analyzed. The absence of readily available adapter information greatly hampers the reproducibility and reusability of publicly available datasets.

Different small RNA-seq library preparation methods create different adapter patterns. Fig 1 illustrates five adapter patterns and read structures generated from 12 small RNA library preparation kits. Among them, two strategies are used to mitigate sequence-specific biases during the ligation step [1,15–20]. One is to introduce random sequence at both 5' and 3' end, such as NEXTFLEX Small RNA-seq Kit with random tetramer (Fig 1c). The other is template switching, which contain non-template oligo sequence at the 5' end and poly(A) tail at the 3' end, such as SMARTer, CATS and D-Plex Small RNA-seq Kits (Fig 1d and 1e). Diverse read structures, randomized adapters, and low complexity regions make adapter detection challenging.

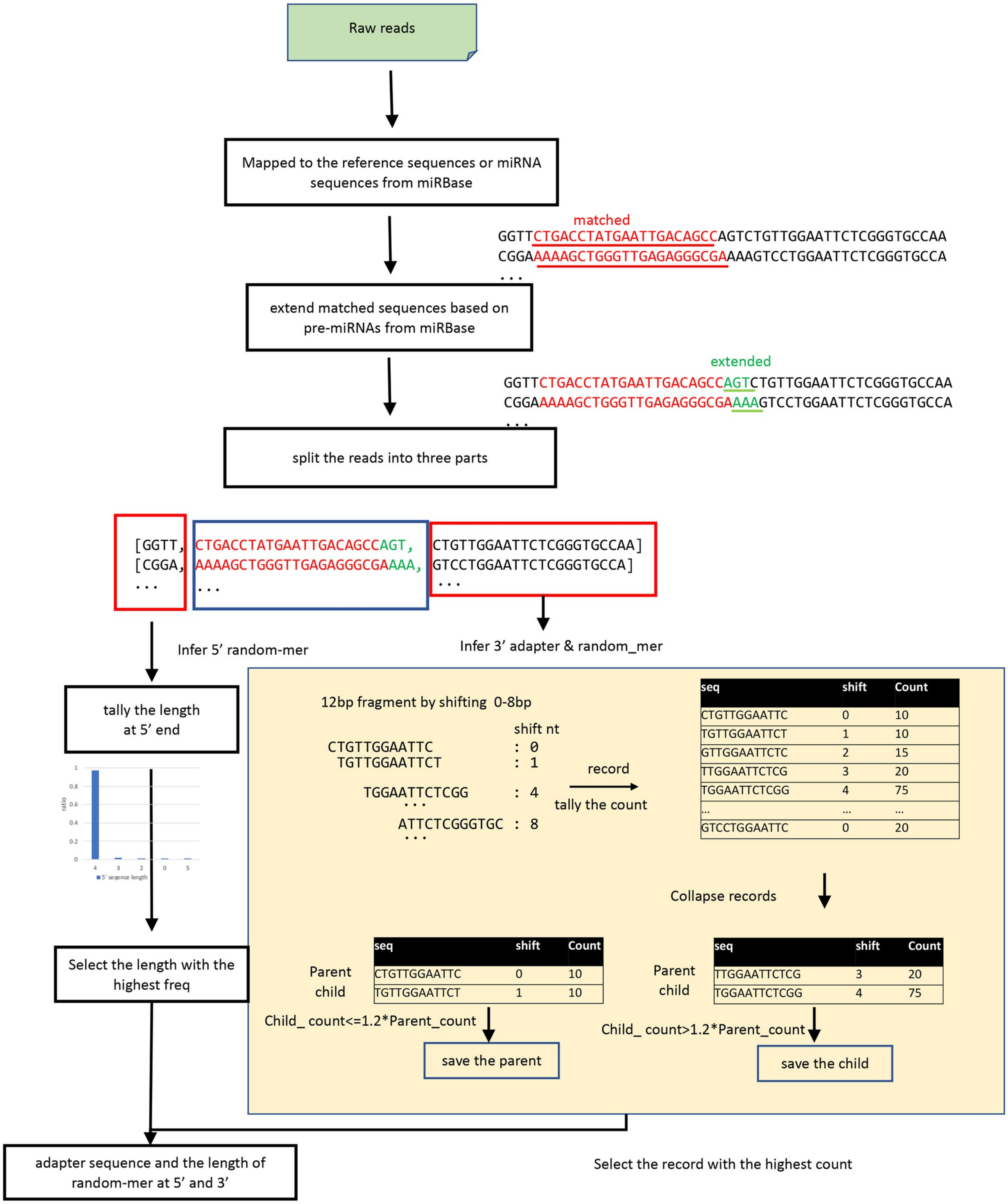
Several tools have been developed to identify adapter sequences from raw reads. For examples, Atropos [21], fastp [22], DNApi [23], and EARRINGS [24] count k-mers in reads to infer adapter sequences. However, k-mer counting methods do not consider random-mers at each side of the ligation site, such as the random tetramer at both the 5' and 3' ends produced by the NEXTFLEX Kit (Fig 1c). Additionally, most methods struggle to identify adapters with low complexity, such as the poly(A) tail introduced when libraries are prepped with SMARTer or CATS Kit (Fig 1d). The latest approach, adapt_find, uses a BLAST-based [25] strategy to identify adapter sequences [26]. It blasts reads against the reference genome to locate small RNA sequences and then infers the adapter pattern, which is computationally intensive and time-consuming. Although it is supposed to work theoretically, the package failed to return any results due to one logic error on the BLAST E-value in the script.

Here, we developed FindAdapt, a fast and light-weight Python package that identifies complex and various adapter patterns accurately without prior information. FindAdapt is far superior to existing tools, with 100% accuracy in simulated and real datasets. FindAdapt makes it easy to streamline small RNA-seq analysis and provides a hassle-free way to analyze a large number of datasets, even those generated by different commercial and customized preparation kits.

Design and implementation

FindAdapt is designed to identify any adapter patterns, including but not limited to the variety of read structures generated by different small RNA-seq preparation kits. Fig 1 illustrates read structures from 12 commonly used kits. The simplest structures are generated by NEBNext, TruSeq, Lexogen, TailorMix, CleanTag, TrueQuant, and HTG EdgeSeq WT-miRNA, wherein the 3' adapter immediately follows the true small RNA sequence (Fig 1a). The QIAseq miRNA library kit introduces a UMI sequence after the true small RNA sequence and the 3' adapter (Fig 1b). NEXTFLEX adds random tetramer at both 5' and 3' sides of the true small RNA sequence to greatly reduce sequence biases, followed by the 3' adapter sequence (Fig 1c). The template switching based kits, such as SMARTer, CATS, and D-plex, have the poly(A) at the 3' end, and the template switching motif (TSM) at the 5' end of the true small RNA sequence (Fig 1d and 1e). The template switching motif is generally a random sequence of 3 bp. The D-plex kit also adds the UMI sequence of 12 bp before the TMS (Fig 1e). The generalized structure of various kits is illustrated in Fig 1f, which includes 0-N bp random-mers at the 5' end, 0-N bp random-mers at the 3' end and followed by the 3' adapter sequence or poly(A).

The workflow of FindAdapt is shown in Fig 2. To identify adapters without prior information about library preparation, the essential step is to locate the true small RNA sequence. Rather than BLAST against the reference genome to find true sequences, FindAdapt locates true miRNAs in raw reads using the Aho-Corasick algorithm [27], which is more computationally efficient and faster. The Aho-Corasick algorithm builds a fast finite state machine of all the given reference sequences for executing searches in linear time, which is one of the most powerful algorithms for searching for patterns in a large set of sequences. The reference sequences are provided by users via the '-seq' option or derived from miRBase v22.1 automatically from the organism specified by users. After locating regions of small RNAs, FindAdapt extends mapped regions based on pre-miRNA sequences from miRBase and then uses the mapped and extended regions to further define the exact boundary between true reads and adapters. FindAdapt splits reads into three regions: the upstream, the mapped and extended region, and the downstream. The upstream sequence is used to infer the 5' random-mer length, while the downstream sequence is used to infer the adapter sequence and 3' random-mer length. Each downstream sequence is shifted by 0-m bp (default m = 8) from the 5' end to generate a n-bp fragment (default n = 12). The default value of n set to 12 is because the adapter sequence is either longer than 12bp or empty, and 12bp is long enough to exclude random matches. FindAdapt records each n-bp fragment along with its shifted position and tallies counts from all downstream sequences (Fig 2). FindAdapt then collapses records based on their relationships to find the most likely adapter pattern. The record with (k+1)-bp shift is the child of a record with k-bp shift if its sequence can be generated from the sequence of that record by shifting one bp. For example, the record TGTTGGAATTCT with 1-bp shift is the


Raw reads
Mapped to the reference sequences or miRNA sequences from miRBase
matched
GGTTCTGACCTATGAATTGACAGCCAGTCTGTTGGAATTCTCGGGTGCCAA
CGGAAAAAGCTGGGTTGAGAGGGCGAAAAGTCCTGGAATTCTCGGGTGCCA
...
extend matched sequences based on pre-miRNAs from miRBase
extended
GGTTCTGACCTATGAATTGACAGCCAGTCTGTTGGAATTCTCGGGTGCCAA
CGGAAAAAGCTGGGTTGAGAGGGCGAAAAGTCCTGGAATTCTCGGGTGCCA
...
split the reads into three parts
[GGTT,
[CGGA,
...
CTGACCTATGAATTGACAGCCAGT,
AAAAGCTGGGTTGAGAGGGCGAAAA,
...
CTGTTGGAATTCTCGGGTGCCAA]
GTCCTGGAATTCTCGGGTGCCA]
...
Infer 5' random-mer
Infer 3' adapter & random_mer
tally the length at 5' end
1
0.8
0.6
0.4
0.2
0
ratio
4
3
2
0
5
5' seqence length
Select the length with the highest freq
12bp fragment by shifting 0-8bp
shift nt
CTGTTGGAATTC : 0
TGTTGGAATTCT : 1
TGGAATTCTCGG : 4
...
ATTCTCGGGTGC : 8
...
record
tally the count
Collapse records
Parent
child
Child_ count<=1.2*Parent_count
save the parent
Parent
child
Child_ count>1.2*Parent_count
save the child
adapter sequence and the length of random-mer at 5' and 3'
Select the record with the highest count

seq	shift	Count
CTGTTGGAATTC	0	10
TGTTGGAATTCT	1	10
GTTGGAATTCTC	2	15
TTGGAATTCTCG	3	20
TGGAATTCTCGG	4	75
...	...	...
GTCCTGGAATTC	0	20

seq	shift	Count
CTGTTGGAATTC	0	10
TGTTGGAATTCT	1	10

seq	shift	Count
TTGGAATTCTCG	3	20
TGGAATTCTCGG	4	75

Fig 2. An overview of FindAdapt workflow.

https://doi.org/10.1371/journal.pcbi.1011786.g002

child of the record CTGTTGGAATTC with 0-bp shift. To increase the speed, FindAdapt first chooses the records with the top 5 highest counts in each shifted position. FindAdapt then collapses those top5 records by constructing their parent-child relationships. FindAdapt saves the child record to the candidate pool if the ratio between the count of the child and the parent is greater than a certain cutoff (default: 1.2); otherwise, the parent record is stored. The default cutoff of 1.2 is based on the fact that sequencing errors may be over 10% but less than 20%. FindAdapt finally reports the record with the highest count, where the 12bp fragment is the adapter sequence and its shifted position is the length of the 3' random-mer (Fig 2) (Algorithm 1 in S1 Text). To be noted, FindAdapt only scans the first 10K reads and stops if it successfully obtains 1,000 matches to the reference; otherwise, it keeps loading 10K reads until it finds enough number of matches.

FindAdapt includes miRBase v22.1 in the package, from which it automatically derives miRNA sequences to use as the reference. Therefore, it supports all 271 organisms that miRBase contains. FindAdapt is implemented in Python 3 (> = 3.6) with multi-threaded support. No installation or dependencies are required. FindAdapt can identify adapters for one dataset or multiple datasets simultaneously. Its input is either a single FASTQ file or a tab-delimited file list, in which the first column is the study name or ID and the second column is the path to the FASTQ files. The output is the length of the random-mer at the 5' and 3' ends and the adapter sequence. In addition, FindAdapt provides the option to run adapter detection and trimming in one step. FindAdapt uses Cutadapt as the default trimming software [10], which is executable if its installation directory is listed in the $PATH environment variable or provided by the -pw_cutadapt argument.

Results

Performance on simulation datasets

We simulated various read structures to evaluate the performance of FindAdapt. We generated nine scenarios: 1) a simple read structure with only the 3' adapter sequence; 2) 1~5bp random-mers at 5' end and the 3' adapter sequence; 3) 1~5bp random-mers at 3' end and the 3' adapter sequence; 4) 1~5bp random-mers at both 5' and 3' end and the 3' adapter sequence; 5) The poly(A) tail; 6) 1~5bp random-mers at 5' end and the poly(A) tail; 7) 1~5bp random-mers at 3' end and the poly(A) tail; 8) 1~5bp random-mers at both 5' and 3' end and the poly(A) tail; 9) no adapters. Scenarios 1–8 included, but were not limited to, adapter patterns generated by commonly used library preparation kits. Scenario 9 served as a negative control in that adapters were not included in the reads. The simulated datasets were created by randomly sampling 10,000 post-trimmed reads from a real small RNA-seq dataset (SRA ID: SRR6502962), which mimicked realistic levels of sequencing errors. To build the random-mer sequence at the 5' and/or 3' ends, a number ranging from 1–5 was randomly chosen to determine its length (n), and then [A, T, C, G] was randomly sampled n times to generate the random-mer sequence. For scenarios 1–4, the 3' adapter sequence was randomly chosen from an adapter pool compiled from real datasets with 0.25% sequencing error [28]. For scenarios 5–8, 12-nt poly(A) was used as the 3' adapter. Random nucleotides were added to the tail to reach 50 bp if simulated reads were shorter than 50 bp, which is the read length of the real small RNA-seq dataset.

We generated 20 datasets for each scenario. The approach is successful if it recognizes both the length of the random-mer and the 3' adapter sequence correctly when there are adapters (scenarios 1–8). For negative controls, as in scenario 9, the approach is successful if it reports random-mers of length zero and empty 3' adapter sequences, or if it doesn't return any adapters.

Notably, FindAdapt achieved 100% accuracy for all scenarios involving 180 datasets. We then compared the performance of FindAdapt with DANpi, EARRINGS_sensitive, EARRINGS_default, Atropos_khmer (default), Atropos_heuristic, and fastp only on scenarios 1, 5 and 9 since they cannot handle random-mers at the 5' and/or 3' ends (Fig 3a). EARRINGS_sensitive and Atropos_heuristic are more sensitive but computationally intensive than their default modes. In contrast, Atropos_heuristic and fastp failed in every scenario, even in the simplest scenario 1, where the regular 3' adapter immediately followed the true sequence. DNApi achieved 100% accuracy in scenario 1, but failed in scenario 5 with poly(A) tail as adapters and scenario 9 by reporting false adapters when none were included. EARRINGS_sensitive achieved 95% accuracy in scenario 1 and 55% accuracy in scenario 5 with poly(A) tail. It successfully reported no adapters in scenario 9. EARRINGS_default worked by not returning adapters in scenario 9, but failed in scenarios 1 and 5. Atropos_khmer only achieved 40% accuracy in scenarios 1 and 9 and failed in scenario 5 (Fig 3a).

In summary, FindAdapt is far superior to existing approaches (Fig 3a). FindAdapt successfully identified adapters with random-mers at the 5' and/or 3' ends, which were not considered in existing approaches. For scenarios without random-mers at the 5' and 3' ends, only EARRINGS_sensitive identified poly(A) tails with low accuracy (55%) at the cost of much longer computational time. It took FindAdapt 0.2s to analyze one typical dataset in scenario 5, compared to 7.6s for EARRINGS_sensitive (>35 times). Additionally, Atropos_heuristic,

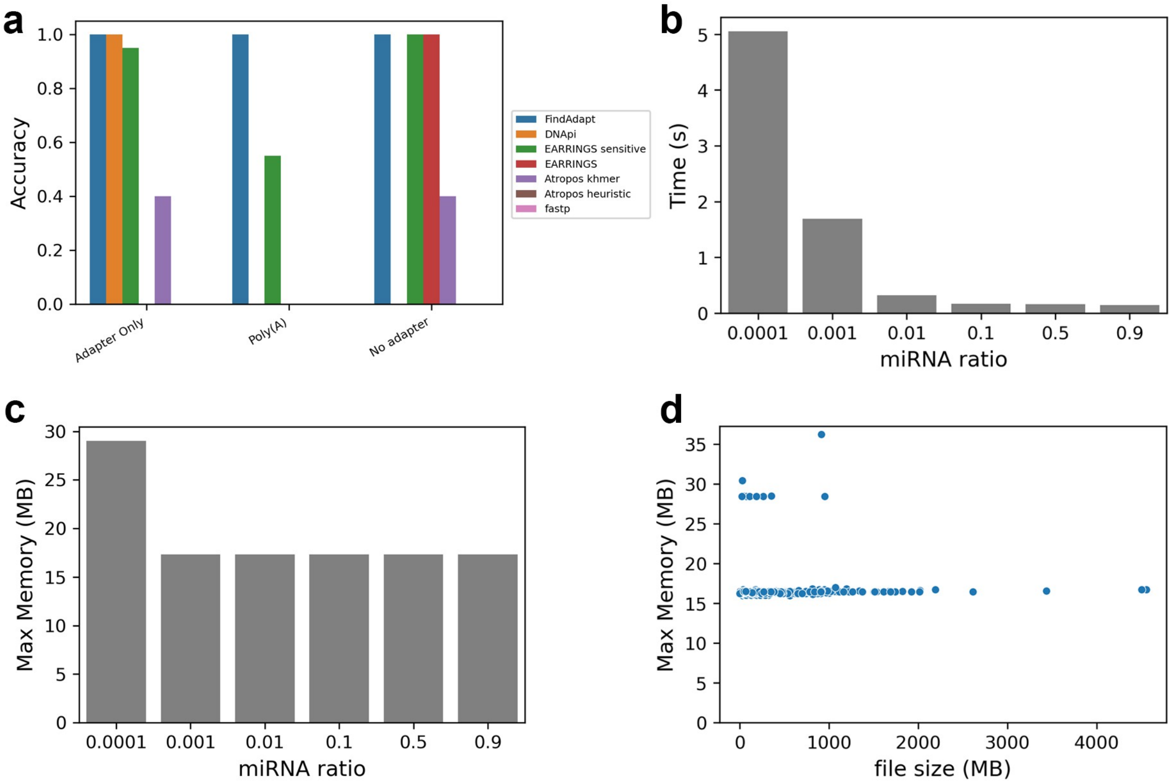


Fig 3. The performance of FindAdapt on simulation datasets. (a) The accuracy of FindAdapt, DNApi, EARRINGS_sensitive, EARRINGS_default, Atropos_khmer, Atropos_heuristic, and fastp on scenarios 1 (adapter only), 5 (poly (A) tail), and 9 (no adapter) where random-mers were not inserted. (b) The computational time of FindAdapt when miRNA percentages range from 0.0001 to 0.9. (c) The memory usage of FindAdapt when miRNA percentage range from 0.0001 to 0.9. (d) The memory usage of FindAdapt when file sizes range from 10Mb to 3,000Mb.

https://doi.org/10.1371/journal.pcbi.1011786.g003

Atropos_khmer, DNApi, and fastp reported false adapters when no adapters existed (scenario 9 in Fig 3a).

We further evaluated FindAdapt in terms of computational time and memory usage. We simulated small RNA-seq datasets with different miRNA percentages and file sizes. The simulated datasets were generated from a real dataset by randomly sampling miRNA reads and non-miRNA reads. FindAdapt was very fast, only taking less than 1s to process most datasets. Additionally, we observed that the computational time of FindAdapt was uncorrelated to the file size but negatively correlated to miRNA percentages in the data (Fig 3b). More sequences were required for screening to find 1,000 matches if a lower percentage of miRNAs was present. In the simulated datasets, for example, it took 0.17s for FindAdapt to analyze a dataset with 10% miRNA, whereas it took ~5s to analyze a dataset with 0.01% miRNA (Fig 3b). FindAdapt requires very little memory to run, and its memory usage is invariant with respect to the miRNA percentage and file size unless the miRNA percentage is extremely low (< = 0.01%) (Fig 3c and 3d). In summary, FindAdapt is fast and memory-efficient.

Performance on real datasets

We applied FindAdapt to 87 GEO [29] datasets involving 3,148 samples from extracellular vesicle-centered studies, which are challenging to analyze due to low percentages of miRNAs and high contamination. They covered a variety of adapter patterns and read structures (Fig 4a), which reflect a realistic situation when researchers analyze public datasets (S1 Table). Of these datasets, 70 studies provided information on adapters or small RNA library preparation kits, including widely used kits (NEBNext, TruSeq, QIAseq, and SMARTer smRNA-seq Kit) as well as rarely used ones (ion total RNA-seq Kit v2, TrueQuant SmallRNA Seq Kit, and HTG Edge-Seq WT-miRNA Assay) and customized adapters (GSE158948 and GSE72183) [30,31]. 17 studies did not specify the adapter or kit information (Fig 4a). For the 70 studies with kits or adapters available, FindAdapt identified patterns matching the provided information in 51 datasets (including two studies with customized adapters), mismatched in six datasets, and reported no adapters in 13 datasets (S1 Table). Manually checking raw reads, we found that incorrect adapters or kits were provided in the six studies and post-trimmed reads were deposited in the 13 studies. For the 17 studies lacking information, we generated miRNA count tables very close to those provided in the GEO, demonstrating that the identified adapters were correct (Fig 4a and S1 Table). In summary, FindAdapt identified adapters in all 87 studies accurately (Fig 4a).

We compared the performance of FindAdapt with DANpi, EARRINGS_sensitive, EARRINGS_default, Atropos_khmer, Atropos_heuristic, and fastp on the 77 studies where random-mers were not included in the 5' and/or 3' ends. Consistent with simulation results, FindAdapt is far superior to other approaches (Fig 4b). Although DNApi obtained good performance on datasets prepared by TruSeq, NEBNext, QIAseq, and other kits, it failed when adapters were already trimmed or contained poly(N). EARRINGS and EARRINGS_sensitive performed poorly on most studies except those from the NEBNext kit. This was mainly due to the fact that the two approaches reported the adapter from the NEBNext kit as the default when no adapters were detected. Atropos_heuristic performed well on studies produced by TruSeq, NEBNext, and other kits. However, it failed to identify adapters prepared by QIAseq and poly(N) tail, and it reported adapters falsely when no adapters were included. Atropos_khmer failed in most studies, and fastp only worked well on specific adapters generated by the NEBNext kit (Fig 4b).

In terms of computational time and memory, FindAdapt is much faster and more memory-efficient than other approaches (Fig 4c and 4d). It took FindAdapt 0.1s–9s (median = 0.2s)

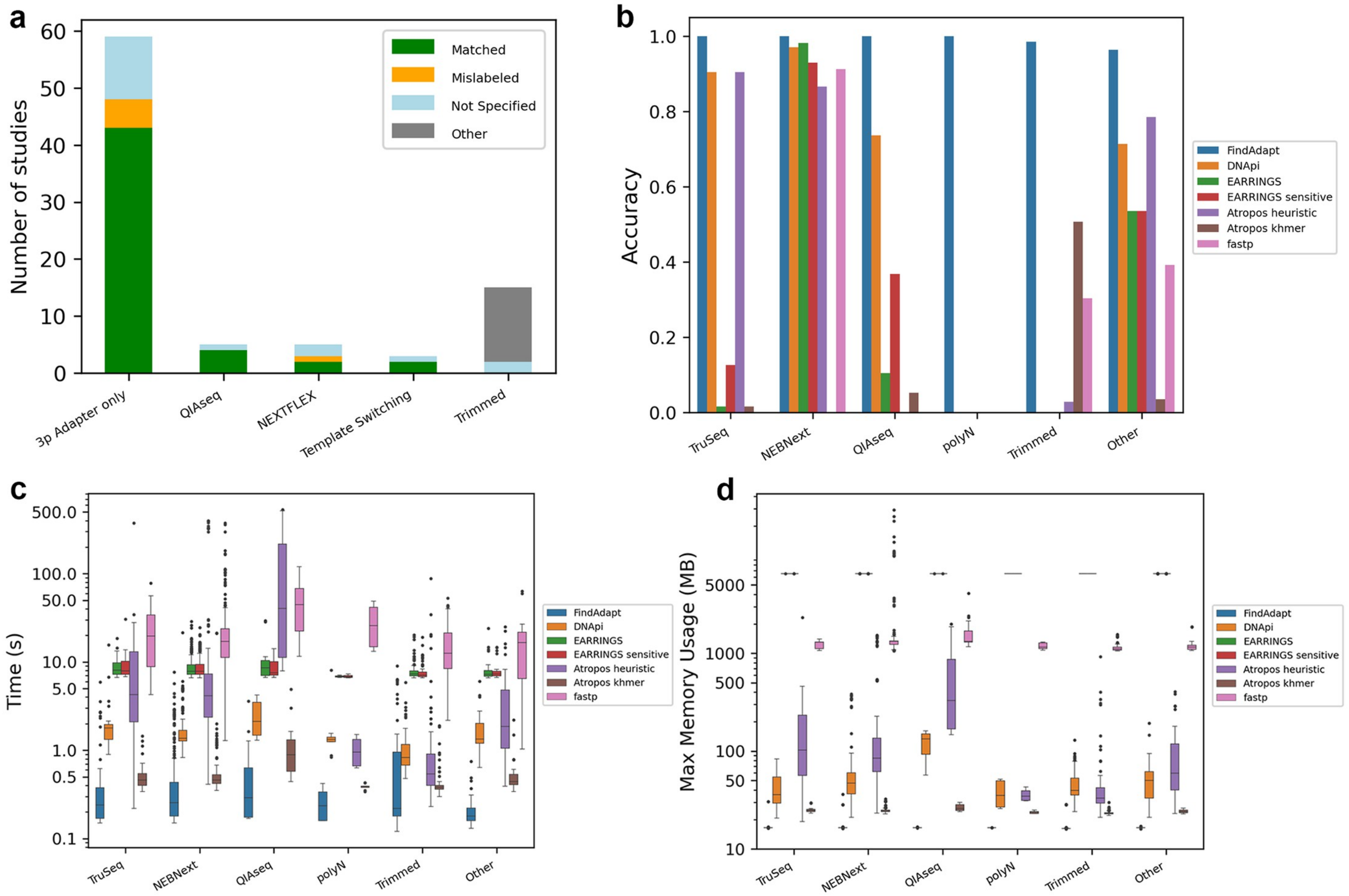


Fig 4. The performance of FindAdapt on the 87 human small RNA-seq studies. (a) The comparison of adapters identified by FindAdapt with the information provided in original studies in each type of adapter patterns. Matched (denoted in green) means the adapter identified by FindAdapt match the kit information provided in the original study. Mislabeled (denoted in orange) means the kit information provided by the original study was incorrect but FindAdapt identified adapters correctly. Not specified (denoted in blue) indicates the kit information was not provided in the original study. Other (denoted in gray) suggests that the kit information was provided but the reads were already post-trimmed. (b) The accuracy of FindAdapt and other methods on the 77 studies where read structures did not contain random-mers at the 5' and/or 3' ends. (c) The runtime of FindAdapt and other methods; d) The memory usage of FindAdapt and other methods.

https://doi.org/10.1371/journal.pcbi.1011786.g004

to process every dataset, followed by Atrophos_khmer 0.3s–21.4s (median = 0.4s), DNApi 0.5s–22s (median = 1.4s), Atrophos_heuristic 0.2s–530.5s (median = 3.2s), EARRINGS (default and sensitive) 6.6s–30.7s (median = 7.6s), and fastp 1s–276s (median = 17.1s) (Fig 4c). FindAdapt required memory of 16–36 Mb (median = 16 Mb), followed by Atrophos_khmer 22–32 Mb (median = 24 Mb), DNApi 20–385 Mb (median = 47 Mb), Atrophos_heuristic 19–2,332 Mb (median = 79 Mb), fastp 1,057–2,9240 Mb (median = 1,283 Mb), and EARRINGS (default and sensitive) 6,539 Mb (Fig 4d). The evaluation was performed on a Linux server with 96 cores Intel(R) XeonGold 6246 CPU (3.2GHz) and 1T RAM.

Availability and future directions

FindAdapt is fast, memory-efficient, and powerful for identifying a variety of adapter patterns, which facilitates small RNA-seq analysis and the reproducibility of publicly available datasets. FindAdapt achieved much better performance than existing approaches since it distinguishes adapters from true sequences by matching reads to a reference instead of identifying overrepresented k-mers. FindAdapt is also flexible on adapter patterns by considering random-mers at the 5' and 3' ends.

FindAdapt uses miRNA sequences automatically derived from miRBase v22.1 as the reference. Therefore, FindAdapt supports all 271 organisms contained in the miRBase, including human, mouse, rat, fly, and Arabidopsis. We evaluated the performance of FindAdapt on 50 small RNA-seq datasets, which were randomly chosen from the GEO across multiple organisms. As a result, FindAdapt achieved 100% accuracy as well (S2 Table). Besides, FindAdapt provides the option to use a list of sequences as a reference in cases where the miRNA sequences of the organism are not included in the miRBase or when other types of small RNAs rather than miRNAs are the major focus.

FindAdapt achieved 100% accuracy on 3,184 small RNA-seq datasets from human and 50 from non-human organisms. In extreme scenarios when non-template isomiRs become the dominant form, however, FindAdapt might not be able to identify adapters correctly. FindAdapt assumes a general structure of adapters consisting of 0-Nbp random-mer at the 5', and 0-N bp random-mer at the 3' ends, followed by adapter sequences (Fig 1f), which covers various adapter patterns from existing small RNA library preparation kits. Further development or adjustment of FindAdapt might be required if a new kit generates an adapter pattern unmatching the assumed general structure.

FindAdapt is freely available at https://github.com/chc-code/FindAdapt. The installation instruction, manual, output format, and examples of FindAdapt are also provided at the GitHub.

Supporting information

S1 Table. Adapter patterns identified by FindAdapt on 87 human small RNA-seq studies.
(XLSX)

S2 Table. Adapter patterns identified by FindAdapt on 13 studies involving 50 small RNA-seq datasets from non-human species, including mouse, rat, Arabidopsis, rice, and grape.
(XLSX)

S1 Text. Pseudocode for identifying the adapter sequence and determining 3' random-mer length.
(PDF)

Author Contributions

Data curation: Hua-Chang Chen.

Formal analysis: Hua-Chang Chen.

Funding acquisition: Qi Liu.

Investigation: Qi Liu.

Methodology: Hua-Chang Chen.

Project administration: Jing Wang, Yu Shyr, Qi Liu.

Software: Hua-Chang Chen.

Supervision: Yu Shyr, Qi Liu.

Visualization: Hua-Chang Chen.

Writing – original draft: Hua-Chang Chen.

Writing – review & editing: Jing Wang, Qi Liu.

References

1. Benesova S, Kubista M, Valihrach L. Small RNA-Sequencing: Approaches and Considerations for miRNA Analysis. Diagnostics (Basel). 2021; 11(6). Epub 2021/06/03. https://doi.org/10.3390/diagnostics11060964 PMID: 34071824.
2. Shi J, Zhou T, Chen Q. Exploring the expanding universe of small RNAs. Nat Cell Biol. 2022; 24 (4):415–23. Epub 2022/04/14. https://doi.org/10.1038/s41556-022-00880-5 PMID: 35414016.
3. Sadick JS, O'Dea MR, Hasel P, Dykstra T, Faustin A, Liddelow SA. Astrocytes and oligodendrocytes undergo subtype-specific transcriptional changes in Alzheimer's disease. Neuron. 2022; 110 (11):1788–805 e10. Epub 2022/04/06. https://doi.org/10.1016/j.neuron.2022.03.008 PMID: 35381189.
4. Li C, Zhang R, Zhang Z, Ren C, Wang X, He X, et al. Expression characteristics of piRNAs in ovine luteal phase and follicular phase ovaries. Front Vet Sci. 2022; 9:921868. Epub 2022/09/27. https://doi.org/10.3389/fvets.2022.921868 PMID: 36157184.
5. Ma D, Zhou X, Wang Y, Dai L, Yuan J, Peng J, et al. Changes in the Small Noncoding RNAome During M1 and M2 Macrophage Polarization. Front Immunol. 2022; 13:799733. Epub 2022/05/28. https://doi.org/10.3389/fimmu.2022.799733 PMID: 35619693.
6. Xiao L, Wang J, Ju S, Cui M, Jing R. Disorders and roles of tsRNA, snoRNA, snRNA and piRNA in cancer. J Med Genet. 2022; 59(7):623–31. Epub 2022/02/12. https://doi.org/10.1136/jmedgenet-2021-108327 PMID: 35145038.
7. Zhang C, Wang D, Wang L, Guan L, Yang H, Zhang Z, et al. Cause-aware failure detection using an interpretable XGBoost for optical networks. Opt Express. 2021; 29(20):31974–92. Epub 2021/10/08. https://doi.org/10.1364/OE.436293 PMID: 34615278.
8. Olivares D, Perez-Hernandez J, Perez-Gil D, Chaves FJ, Redon J, Cortes R. Optimization of small RNA library preparation protocol from human urinary exosomes. J Transl Med. 2020; 18(1):132. Epub 2020/03/20. https://doi.org/10.1186/s12967-020-02298-9 PMID: 32188466.
9. Chavez Montes RA, Jaimes-Miranda F, de Folter S. Bioinformatic Analysis of Small RNA Sequencing Libraries. Methods Mol Biol. 2019; 1932:51–63. Epub 2019/02/01. https://doi.org/10.1007/978-1-4939-9042-9_4 PMID: 30701491.
10. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011; 17(1):3. https://doi.org/10.14806/ej.17.1.200
11. Hannon. FASTX-Toolkit: FASTQ/a short-reads pre-processing tools. 2010.
12. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016; 9:88. Epub 2016/02/13. https://doi.org/10.1186/s13104-016-1900-2 PMID: 26868221.
13. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. Epub 2014/04/04. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404.
14. Zhong X, Heinicke F, Lie BA, Rayner S. Accurate Adapter Information Is Crucial for Reproducibility and Reusability in Small RNA Seq Studies. Noncoding RNA. 2019; 5(4). Epub 2019/10/31. https://doi.org/10.3390/ncrna5040049 PMID: 31661777.
15. Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res. 2011; 39(21):e141. Epub 2011/09/06. https://doi.org/10.1093/nar/gkr693 PMID: 21890899.
16. Raabe CA, Tang TH, Brosius J, Rozhdestvensky TS. Biases in small RNA deep sequencing data. Nucleic Acids Res. 2014; 42(3):1414–26. Epub 2013/11/08. https://doi.org/10.1093/nar/gkt1021 PMID: 24198247.
17. Fuchs RT, Sun Z, Zhuang F, Robb GB. Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. PLoS One. 2015; 10(5):e0126049. Epub 2015/05/06. https://doi.org/10.1371/journal.pone.0126049 PMID: 25942392 following competing interests: The authors were employees of New England Biolabs, Inc. when the experiments described were performed. The study was conducted in the Research Department. There is currently one patent application pending, and one product under development based on the findings presented here. This does not alter the authors' adherence to the PLOS ONE policies on data sharing and materials.
18. Maguire S, Lohman GJS, Guan S. A low-bias and sensitive small RNA library preparation method using randomized splint ligation. Nucleic Acids Res. 2020; 48(14):e80. Epub 2020/06/05. https://doi.org/10.1093/nar/gkaa480 PMID: 32496547.

19. Lama L, Cobo J, Buenaventura D, Ryan K. Small RNA-seq: The RNA 5'-end adapter ligation problem and how to circumvent it. J Biol Methods. 2019; 6(1). Epub 2019/05/14. https://doi.org/10.14440/jbm.2019.269 PMID: 31080843.

20. Hafner M, Renwick N, Brown M, Mihailovic A, Holoch D, Lin C, et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA. 2011; 17(9):1697–712. Epub 2011/07/22. https://doi.org/10.1261/rna.2799511 PMID: 21775473.

21. Didion JP, Martin M, Collins FS. Atropos: specific, sensitive, and speedy trimming of sequencing reads. PeerJ. 2017; 5:e3720. Epub 2017/09/07. https://doi.org/10.7717/peerj.3720 PMID: 28875074.

22. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018; 34(17):i884–i90. Epub 2018/11/14. https://doi.org/10.1093/bioinformatics/bty560 PMID: 30423086.

23. Tsuji J, Weng Z. DNApi: A De Novo Adapter Prediction Algorithm for Small RNA Sequencing Data. PLoS One. 2016; 11(10):e0164228. Epub 2016/10/14. https://doi.org/10.1371/journal.pone.0164228 PMID: 27736901.

24. Wang TH, Huang CC, Hung JH. EARRINGS: An Efficient and Accurate Adapter Trimmer Entails No a Priori Adapter Sequences. Bioinformatics. 2021. Epub 2021/01/19. https://doi.org/10.1093/bioinformatics/btab025 PMID: 33459339.

25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–10. Epub 1990/10/05. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712.

26. Mohideen A, Johansen SD, Babiak I. High-Throughput Identification of Adapters in Single-Read Sequencing Data. Biomolecules. 2020; 10(6). Epub 2020/06/12. https://doi.org/10.3390/biom10060878 PMID: 32521604.

27. AV A, MJ C. Efficient String Matching: An Aid to Bibliographic Search. Commun ACM. 1975; 18(6):40.

28. Pfeiffer F, Grober C, Blank M, Handler K, Beyer M, Schultze JL, et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Sci Rep. 2018; 8(1):10950. Epub 2018/07/22. https://doi.org/10.1038/s41598-018-29325-6 PMID: 30026539.

29. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30(1):207–10. Epub 2001/12/26. https://doi.org/10.1093/nar/30.1.207 PMID: 11752295.

30. Mitchell MI, Ben-Dov IZ, Liu C, Ye K, Chow K, Kramer Y, et al. Extracellular Vesicle Capture by AnTibody of CHoice and Enzymatic Release (EV-CATCHER): A customizable purification assay designed for small-RNA biomarker identification and evaluation of circulating small-EVs. J Extracell Vesicles. 2021; 10(8):e12110. Epub 2021/06/15. https://doi.org/10.1002/jev2.12110 PMID: 34122779.

31. Ben-Dov IZ, Whalen VM, Goilav B, Max KE, Tuschl T. Cell and Microvesicle Urine microRNA Deep Sequencing Profiles from Healthy Individuals: Observations with Potential Impact on Biomarker Studies. PLoS One. 2016; 11(1):e0147249. Epub 2016/01/20. https://doi.org/10.1371/journal.pone.0147249 PMID: 26785265.