non-coding RNA

MDPI

Commentary

Accurate Adapter Information Is Crucial for Reproducibility and Reusability in Small RNA Seq Studies

Xiangfu Zhong [1,*], Fatima Heinicke [1], Benedicte A. Lie [1] and Simon Rayner [1,2,*]

[1] Department of Medical Genetics, Oslo University Hospital and University of Oslo, 0450 Oslo, Norway; fatima.heinicke@medisin.uio.no (F.H.); b.a.lie@medisin.uio.no (B.A.L.)

[2] Hybrid Technology Hub—Centre of Excellence, Institute of Basic Medical Sciences, University of Oslo, 0372 Oslo, Norway

* Correspondence: zhong.xfz@gmail.com (X.Z.); simon.rayner@medisin.uio.no (S.R.); Tel.: +47-9380-6657 (S.R.)

Received: 17 September 2019; Accepted: 25 October 2019; Published: 28 October 2019

Abstract: A necessary pre-processing data analysis step is the removal of adapter sequences from the raw reads. While most adapter trimming tools require adapter sequence as an essential input, adapter information is often incomplete or missing. This can impact quantification of features, reproducibility of the study and might even lead to erroneous conclusions. Here, we provide examples to highlight the importance of specifying the adapter sequence by demonstrating the effect of using similar but different adapter sequences and identify additional potential sources of errors in the adapter trimming step. Finally, we propose solutions by which users can ensure their small RNA-seq data is fully annotated with adapter information.

Keywords: adapter trimming; adapter; small RNA; microRNA; NGS; reproducibility; reusability

Small non-coding RNAs comprise short RNAs less than 200 nt in length, including microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs) and small nucleolar RNAs (snoRNAs) and which have a variety of functions. Next generation sequencing (NGS) is the most common technique for studying small RNA expression. An NGS experiment requires the construction of a library, which includes ligation of an adapter that acts as a binding site for priming the sequencing reaction and capturing the endogenous small RNA inserts [1,2]. Due to the short insert size, the 3′ adapter sequence is commonly included in the raw data reads and removing the adapter from raw reads is a necessary data analysis step. Most studies consider adapter trimming a straightforward pre-processing step. For other NGS experiments such as RNA-seq and ChIP-seq, the insert size is generally longer and the adapter trimming step has less impact. However, there are many issues that can affect the trimming output, which might influence further downstream analysis. For example, under- or overtrimming a read can make a significant difference in the quantification of small RNAs of interest. The problem is further exacerbated when mismatches are allowed during mapping. However, an even more serious challenge is ensuring that the adapter sequence is available, or correctly specified. For example, the adapter sequence in the instruction manual for the NEBNext Small RNA Library Prep Kit for Illumina is specified as:

5′-rAppAGATCGGAAGAGCACACGTCT-NH2-3′

where the *rApp*, a 5′-adenylated termini, is removed when the 3′ adapter and inserted RNA fragment are ligated by the T4 RNA ligase [3].

However, in several studies, for example, [4–6] the *rApp* was replaced with an additional Adenine so the 3′ adapter sequence was specified as:

AAGATCGGAAGAGCACACGTCT

It is unclear whether this modified adapter sequence was used for adapter trimming, but it highlights how mistakes can occur and be propagated in subsequent publications.

Many tools are available for 3′ adapter trimming such as cutadapt [7], fastx_clipper (http://hannonlab.cshl.edu/fastx_toolkit/) and Trimmomatic [8] and they generally require a specified adapter sequence as mandatory input. However, while data portals such as The Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) require deposition of original raw data files, they do not require adapter information as mandatory data. Thus, several tools have been developed to identify adapter sequence from the raw sequencing data, e.g., minion (part of the kraken package [9]), DNApi [10] and AdapterRemoval v2 [11]. However, these predicted adapter sequences should be used with care.

Thus, there are two situations that may lead to incorrect specification of adapter sequence in read trimming. Firstly, different vendors use distinct or similar 3′ adapter sequences in their small RNA sequencing library preparation kits (Supplementary Table S1). Moreover, some kits have been discontinued or outdated. For example, the ScriptMiner Small RNA-Seq Library Preparation Kit from Epicentre was discontinued in December 2013 and version 1 of the TailorMix kit is outdated. Secondly, adapter information is not always available. As mentioned above, the GEO and the SRA do not require adapter related data as mandatory information. A report in 2016 found that only one third of NGS entries in the GEO provide adapter information [10]. Moreover, even when the library preparation kit is specified, it is not always straightforward to find the corresponding adapter sequence. For example, NEBNext, NEXTflex and TruSeq specify the sequence in their product instruction manuals; Lexogen and QIAseq provide the information on their websites; the CleanTag 3′ adapter sequence is specified in a publication [12]; and the TailorMix, the SMARTer and the TrueQuant (used by GenXPro) adapters are declared neither on their websites nor manuals. Additionally, incorrect kit information is sometimes given in publications. For example, some studies have reported preparing libraries using the TailorMix miRNA Sample Preparation Kit v7 [13,14] instead of the currently available kit v2 (as of August 2019, https://www.seqmatic.com/products/tailormix-mirna-sample-preparation-kit-v2/).

To review the situation in 2019, we retrieved small RNA-seq data entries from the SRA by searching for "miRNA-Seq[Strategy]" and obtained 41,607 entries (access date 27 May 2019). TruSeq is the most used kit (although 335 experiments specified this as *TrueSeq*) followed by NEBNext (see Supplementary Table S2). However, still less than half (47%) of entries provide kit information. For the 53% SRA entries are lacking this data, software tools are needed to determine the best guess for the adapter sequence to proceed with trimming. As a simple test, we searched for the adapter in the data from [15] using *minion* [9], *DNApi* [10] and *AdapterRemoval* [11]. Of these tools, only *DNApi* gave the correct, but partial, sequence. The results are shown in Supplementary Table S3.

To determine the impact of adapter sequence and trimming protocol on trimming results, we used data from a study by Dard-Dascot et al. [15]. For details, see "Materials and Methods" in the Supplementary Materials. The goal of this study was to identify bias in various library kits by resequencing the same samples using different kits. One sample included six synthetic RNAs, RNA1 to RNA6, of known sequence (see sequences in Supplementary Table S4, from [15]) and these were the focus of our study. Firstly, we investigated the effect of using three similar adapter sequences on the dataset prepared by the NEBNext kit. The three sequences were the correct adapter sequence from the kit manual (NEBNext_trim01), the sequence with an additional A at the 5′ end (NEBNext_trim02) and the partial adapter sequence from the CATS trimming instructions (NEBNext_trim03) with two nucleotide changes relative to the correct adapter sequence (Figure 1A). The results are shown in Figure 1B–D. Figure 1B shows the effect of using the three different adapters (NEBNext_trim01, NEBNext_trim02 and NEBNext_trim03) on the read count of RNA1 to RNA6 in terms of the percentage of perfectly matched reads identified in the trimmed data. RNA6 is not detected by any of the kits, suggesting an issue with sequencing rather than a trimming problem. When the correct adapter, NEBNext_trim01, is used all the RNAs are detected to varying degrees. RNA1, RNA2 and RNA5, are

the three most highly represented—see Supplementary Table S5 for counts—with more than 90% of these RNAs perfectly trimmed from the raw data. However, when the highly similar NEBNext_trim02 and NEBNext_trim03 adapters were used, fewer than 1% of the reads were correctly trimmed. In this case, an analysis would fail to detect the presence of these RNAs.

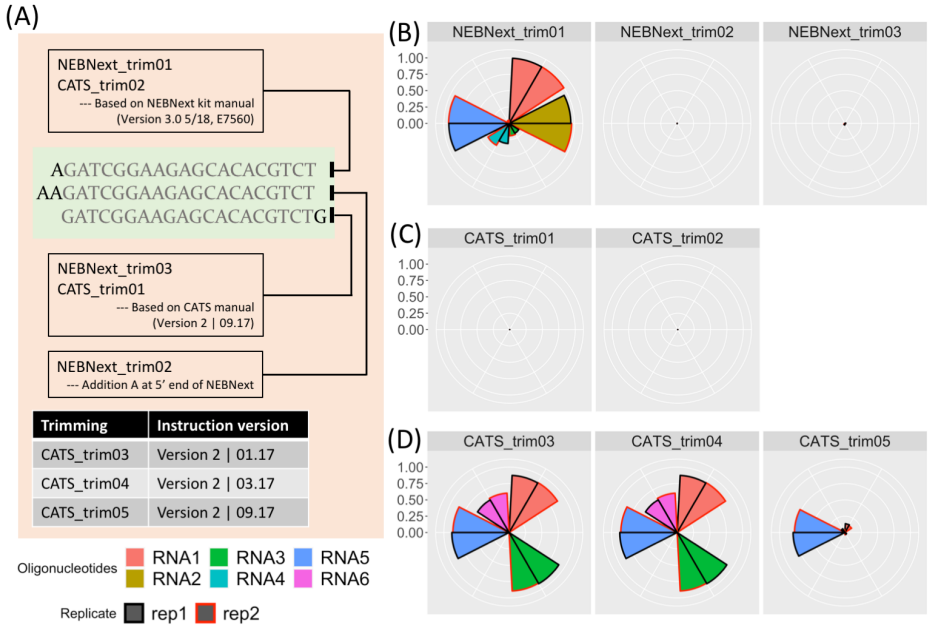

Figure 1. Use of incorrect adapter sequence or trimming protocol can lead to incorrectly trimmed reads and miscounting of reads mapping to features. The selected datasets are originally from [15], samples containing six synthetic small RNAs prepared by the NEBNext and CATS kits. After trimming, the Linux command *grep -c* was used for counting. (**A**) Top part of figure: Schematic of the different but highly similar adapter sequences used in (**B**–**D**). Bottom section of figure: Different versions of the CATS manual used for trimming protocol applied in (**D**). Legend under (**A**) Fill colour corresponds to the six synthetic RNAs in the dataset, line colour corresponds to the two replicates for each sample. Sequence for oligonucleotides are listed in Supplementary Table S5. (**B**) Choice of adapter sequence can have a major impact on downstream analysis. Left: Use of the correct adapter sequence (NEBNext_trim01) identifies the presence of 5 out of 6 synthetic small RNAs present in the NGS dataset. Middle and right (NEBNext_trim02 and NEBNext_trim03): Using a highly similar adapter sequence that differs by one or two nucleotides has a drastic effect on mapped reads with less than 1% of reads identified. (**C**) In some case detailed trimming instructions are required in addition to the adapter sequence. The trimming sets CATS_trim01 and CATS_trim02 were trimmed by specifying the correct adapter sequence, but few perfectly trimmed reads were detected. (**D**) The problem extends to incorrect application of manufacturer's protocol during read trimming. From left to right, trimming results after following trimming instructions specified in the January 2017, March 2017 and September 2017 releases of the manual. The instructions in the latest version are distinct from those provided in the previous two versions and this is reflected in the number of identified reads, with the latest protocol identifying notably fewer reads associated with the synthetic RNAs. CATS_trim01 was trimmed using the same adapter sequence as CATS_trim05, demonstrating that for some kits, specifying the adapter alone is not sufficient to achieve efficient read trimming.

To further confirm that this is not a kit-specific effect, we also tested the TruSeq dataset with eight different adapter sequences including the correct TruSeq 3′ adapter sequence. We found the same results as the NEBNext dataset. Any additional or removed nucleotide at the 5′ end of the adapter gives a different trimming result (Figure S1A). Thus, the correct adapter sequence is required for correct read trimming, even a single base difference in the adapter sequence can introduce large

changes in the digital gene expression (Figure S1B). Similar results are also found within the data from [6], see Supplementary Figure S2 and Supplementary Table S6.

However, providing the correct kit information does not always guarantee correct adapter trimming because of specific procedures in the library preparation protocol. For example, in the CATS Small RNA-seq protocol, single stranded RNAs are first polyadenylated at the 3′ end, and then cDNA synthesis is performed in the presence of a poly(T) anchored adapter. Therefore, in addition to the adapter sequence, the poly(A) between the RNA insert and adapter needs to be properly removed. Figure 1C shows the result of trimming the raw data prepared with the CATS Small RNA-seq Kit by simply specifying the provided adapter sequence. In this case, virtually no reads were retrieved after trimming. Once again, failure to following the manufacturer's protocol for adapter trimming will impact the trimmed output. Similar problems will arise in the dataset prepared by NEXTflex, which requires removal of four additional random nucleotides after adapter removal. Few adapter trimming tools provides built-in functions to handle these additional nucleotide removal steps.

Figure 1D shows the results from following the detailed trimming instructions provided in the CATS manual using three different versions of the user manual dated January 2017, March 2017 and September 2017. The instructions in the first two versions are identical, and detect notably more reads than following the instructions in the latest version of the manual, only read RNA5 is consistently reported after all three procedures. Thus, reporting the kit version is necessary to allow users to reproduce the trimming, however this information is not always provided, for example, in [15].

Additionally, the provided trimming instructions are not effective for the two synthetic RNAs that have an adenine at the 3′ end (RNA2 and RNA4 in this case). The perfectly trimmed reads from these two RNAs were relatively few with less than 10 untransformed raw counts, despite thousands of untrimmed reads present in the raw data (see Supplementary Table S5).

Thus, analysis of sequencing data is only reproducible with access to accurate and unambiguous specification of the adapter sequence that was used in library preparation. Missing adapter sequence represents a major obstruction to taking advantage of the vast amount of publicly available data. Providing comprehensive and consistent information regarding the adapter sequence benefits the research community in general. We therefore propose the three following solutions to increase the reproducibility and accuracy in small RNA-seq studies.

Firstly, most journals publishing articles on sequencing analysis require that users deposit accession numbers to public databases such as the SRA, GEO or ArrayExpress. At the same time, editors could require the adapter sequence to be specified in submitted small RNA sequencing studies. Researchers can also take responsibility by ensuring they include all relevant detail about the adapter, as well as kit information.

Secondly, the fastest and most straightforward approach is to include adapter information as mandatory metadata which needs to be filled out as part of the process of submitting sequencing data to public data repositories such as the SRA. Implementing the FAIR (findable, accessible, interoperable, reusable) principles in data repositories could maximize the sharing and reproducibility of sequencing data [16]. The detailed information regarding library kit version and adapter sequence should be included at this point.

Thirdly, consolidated information from kit manufacturers, offering practical adapter trimming advice, as well as archived versions of adapter sequences, would also be helpful to the research community. Biases, including barcoding, structural and ligation bias are well known and have been discussed in the context of small RNA sequencing [17–21]. Manufacturers are continuously developing or revising kits to achieve more efficient ligation, reduce bias and improve performance, e.g., the NEBNext and NEXTflex kits are currently at version 3, and the current TailorMix kit is version 2. Any information about adapter updates is essential for users and information should be readily available. As poly(A) based approaches are becoming more popular in library preparation, protocols for poly(A) removal should also be defined for accurate adapter trimming. Many small RNAs (such as miRNAs) contain adenines at their 3′ end and these may be mis-trimmed as part of poly(A) and lead

to mis-quantitation. The CATS kit from Diagenode provides ready-to-use trimming instructions in their manuals, and these were updated when they renewed their kit. Nevertheless, our simple analysis indicates that trimming these sequences ending in one or more adenines remains a challenge.

Based on their crucial role in small RNA-seq data analysis, adapter information should follow findable, accessible, interoperable and re-usable (FAIR) principles [22]. The reality is that accurate adapter trimming is not a straightforward process. Consequently, more attention needs to be paid to this step to ensure the correct sequence and protocol is used. Access to correct and detailed adapter information will help to minimize some of the problems associated with this issue.

Supplementary Materials: The following are available online at http://www.mdpi.com/2311-553X/5/4/49/s1. Figure S1: Use of incorrect adapter sequence or trimming protocol can lead to miscounting of reads mapping to features; Figure S2: Trimming analysis results for dataset from [6]. Table S1: 3′ adapters for small RNA-seq kits for the Illumina platform; Table S2: The number of SRA entries that mention library preparation kit in experimental description as of 27 May 2019; Table S3: Predicted adapter sequence from raw sequencing data using three tools, including Minion, DNApi and AdapterRemoval; Table S4: Sequences for the six synthetic oligonucleotides from [15] shown in Figure 1 and Table S5; Table S5: Un-transformed count table for samples in dataset[15] for raw data and using the various adapter sequences as specified in Figure 1; Table S6: Trimming results for top 15 most abundant miRNAs in [6]; Table S7: Datasets used in this analysis. Including their accession number, library preparation kit, sample information and reference; Table S8: Trimming sets generated in the analyses with different adapter sequence and corresponding trimming commands.

Funding: This research was funded by Helse Sør-Øst RHF grant number [2016122 to S.R., 2015034 to B.A.L.]. The APC was funded by Helse Sør-Øst RHF.

Acknowledgments: We would like to express our appreciation to the those authors sharing their sequencing data [6,15].

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

NGS	Next Generation Sequencing
SRA	Sequence Read Archive
GEO	Gene Expression Omnibus
FAIR	Findable, Accessible, Interoperable and Re-usable

References

1. Hafner, M.; Landgraf, P.; Ludwig, J.; Rice, A.; Ojo, T.; Lin, C.; Holoch, D.; Lim, C.; Tuschl, T. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* **2008**, *44*, 3–12. [CrossRef] [PubMed]
2. Lu, C.; Meyers, B.C.; Green, P.J. Construction of small RNA cDNA libraries for deep sequencing. *Methods* **2007**, *43*, 110–117. [CrossRef] [PubMed]
3. Raabe, C.A.; Tang, T.H.; Brosius, J.; Rozhdestvensky, T.S. Biases in small RNA deep sequencing data. *Nucleic Acids Res.* **2014**, *42*, 1414–1426. [CrossRef] [PubMed]
4. Van Goethem, A.; Yigit, N.; Everaert, C.; Moreno-Smith, M.; Mus, L.M.; Barbieri, E.; Speleman, F.; Mestdagh, P.; Shohet, J.; Van Maerken, T.; et al. Depletion of tRNA-halves enables effective small RNA sequencing of low-input murine serum samples. *Sci. Rep.* **2016**, *6*, 37876. [CrossRef] [PubMed]
5. Zovoilis, A.; Cifuentes-Rojas, C.; Chu, H.P.; Hernandez, A.J.; Lee, J.T. Destabilization of B2 RNA by EZH2 Activates the Stress Response. *Cell* **2016**, *167*, 1788–1802.e13. [CrossRef] [PubMed]
6. Gümürdü, A.; Yildiz, R.; Eren, E.; Karakülah, G.; Ünver, T.; Genç, Ş.; Park, Y. MicroRNA exocytosis by large dense-core vesicle fusion. *Sci. Rep.* **2017**, *7*, 45661. [CrossRef]
7. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **2011**, *17*, 10–12. [CrossRef]
8. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]

9. Davis, M.P.A.; van Dongen, S.; Abreu-Goodger, C.; Bartonicek, N.; Enright, A.J. Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods* **2013**, *63*, 41–49. [CrossRef]
10. Tsuji, J.; Weng, Z. DNApi: A De Novo Adapter Prediction Algorithm for Small RNA Sequencing Data. *PLoS ONE* **2016**, *11*, e0164228. [CrossRef]
11. Schubert, M.; Lindgreen, S.; Orlando, L. AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **2016**, *9*, 88. [CrossRef] [PubMed]
12. Shore, S.; Henderson, J.M.; Lebedev, A.; Salcedo, M.P.; Zon, G.; McCaffrey, A.P.; Paul, N.; Hogrefe, R.I. Small RNA Library Preparation Method for Next-Generation Sequencing Using Chemical Modifications to Prevent Adapter Dimer Formation. *PLoS ONE* **2016**, *11*, e0167009. [CrossRef] [PubMed]
13. Niu, J.; Smagghe, G.; De Coninck, D.I.M.; Van Nieuwerburgh, F.; Deforce, D.; Meeus, I. In vivo study of Dicer-2-mediated immune response of the small interfering RNA pathway upon systemic infections of virulent and avirulent viruses in Bombus terrestris. *Insect Biochem. Mol. Biol.* **2016**, *70*, 127–137. [CrossRef] [PubMed]
14. Niu, J.; Meeus, I.; De Coninck, D.I.; Deforce, D.; Etebari, K.; Asgari, S.; Smagghe, G. Infections of virulent and avirulent viruses differentially influenced the expression of dicer-1, ago-1, and microRNAs in Bombus terrestris. *Sci. Rep.* **2017**, *7*, 45620. [CrossRef]
15. Dard-Dascot, C.; Naquin, D.; d'Aubenton Carafa, Y.; Alix, K.; Thermes, C.; van Dijk, E. Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *BMC Genom.* **2018**, *19*, 118. [CrossRef]
16. Corpas, M.; Kovalevskaya, N.V.; McMurray, A.; Nielsen, F.G.G. A FAIR guide for data providers to maximise sharing of human genomic data. *PLoS Comput. Biol.* **2018**, *14*, 1–10. [CrossRef]
17. Seguin-Orlando, A.; Schubert, M.; Clary, J.; Stagegaard, J.; Alberdi, M.T.; Prado, J.L.; Prieto, A.; Willerslev, E.; Orlando, L. Ligation bias in illumina next-generation DNA libraries: Implications for sequencing ancient genomes. *PLoS ONE* **2013**, *8*, e78575. [CrossRef]
18. Tian, G.; Yin, X.; Luo, H.; Xu, X.; Bolund, L.; Zhang, X.; Gan, S.Q.; Li, N. Sequencing bias: Comparison of different protocols of microRNA library construction. *BMC Biotechnol.* **2010**, *10*, 64. [CrossRef]
19. Zhuang, F.; Fuchs, R.T.; Sun, Z.; Zheng, Y.; Robb, G.B. Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Res.* **2012**, *40*, e54. [CrossRef]
20. Sorefan, K.; Pais, H.; Hall, A.E.; Kozomara, A.; Griffiths-Jones, S.; Moulton, V.; Dalmay, T. Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* **2012**, *3*, 4. [CrossRef]
21. Alon, S.; Vigneault, F.; Eminaga, S.; Christodoulou, D.C.; Seidman, J.G.; Church, G.M.; Eisenberg, E. Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Res.* **2011**, *21*, 1506–1511. [CrossRef] [PubMed]
22. Sansone, S.A.; McQuilton, P.; Rocca-Serra, P.; Gonzalez-Beltran, A.; Izzo, M.; Lister, A.L.; Thurston, M.; FAIRsharing Community. FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.* **2019**, *37*, 358–367. [CrossRef] [PubMed]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).