analytical chemistry

pubs.acs.org/ac

Comment

Response to the Comment on "DpCoA tagSeq: Barcoding dpCoA-Capped RNA for Direct Nanopore Sequencing via Maleimide-Thiol Reaction"

Xiaojian Shao,[‡] Hailei Zhang,[‡] Zhou Zhu, Fenfen Ji, Zhao He, Zhu Yang,* Yiji Xia,* and Zongwei Cai*

Cite This: https://doi.org/10.1021/acs.analchem.3c05281

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

In our recent paper,[1] we reported the dpCoA tagSeq method for the identification of dpCoA-capped RNA (dpCoA-RNA). In this method, synthesized RNA containing thiol-reactive maleimide is used to label dpCoA-RNAs. After tagging, the RNA samples were sequenced by Oxford nanopore sequencing technology. We conducted tagSeq procedures on the RNA samples extracted from mouse liver combined with model dpCoA-RNA and identified 44 genes that may potentially transcribe dpCoA-RNA.[2−5] Jeppe Vinther recently analyzed our data and identified two issues in our data processing: the presence of chimeric reads blending the model RNA spike-in and mouse RNAs, and incorrect mapping of reads to multiple genes. We deeply appreciate these comments, which prompted us to carefully consider and reanalyze our data. Our new analysis indicated that some of the 44 putative dpCoA-RNAs we previously reported are like artifacts from chimeric reads likely generated during nanopore sequencing when the sequence of the spiked-in synthetic model dpCoA-RNA was joined with that of a cellular RNA.

In our corrective efforts, we implemented several measures (Figure 1). First, we introduced two more steps to our data processing protocol to remove potential chimeric reads: (i) eliminating reads containing the tagged model RNA sequence and (ii) filtering out reads with a 5′ end clip length of 20−100 nt (including ∼40 nt of tag RNA). This approach effectively removed not only the chimeric reads of the model RNA and cellular RNA molecules but also other possible chimeras.

Additionally, to avoid assigning a single read to multiple genes, we selected only the gene with the highest mapping score (MAPQ value). Further scrutiny involved using the integrative genome viewer (IGV) and mafft alignment tools (Figure 2).[6] The refined analysis dramatically reduced the number of tagged RNA reads, leaving us with seven potential dpCoA-RNA candidates.

RESULTS AND DISCUSSION

Our revised data analysis revealed that approximately 58% of the previously identified dpCoA-RNA reads contain the sequence that aligns with both spike-in model dpCoA-RNA and cellular RNA (Supporting Data 1). We agree with Jeppe Vinther that these chimeric reads may be false positives stemming from the potential issue of the nanopore sequencing method by erroneous merging of two successive RNAs traversing the nanopore within a short time interval (Figure 3).[7−9] The intrinsic errors associated with nanopore technology could explain, in part, the presence of 193 tagged model RNA instances in the PM−_1 group (Figure 2A). Furthermore, the inadvertent merging of two RNAs may contribute significantly to the approximately 58% of chimeric reads, particularly in the presence of a substantial quantity of model dpCoA-RNAs.

In addition to eliminating chimeric reads, we implemented the steps mentioned above to enhance the reliability of dpCoA-RNAs. Consequently, we refined the pool of identified dpCoA-RNAs to encompass only eight genes, characterized by tagged reads exceeding 5 in the PM+_1 group and a fold change (PM+_1/PM−_1) surpassing 10.

We further inspected the multiple mapping issue and eliminated multimapping of the RNA reads by mapping one read to one gene which has the highest MAPQ score. In this way, a substantial number of reads were identified as mapping to gene nos. 2, 4, and 7. Following this analytical step, no RNA reads were allocated to gene no. 4. Consequently, we refined the selection of dpCoA-RNAs, focusing on a subset of seven genes (see Table 1).

Upon closer examination of the reads using IGV and mafft, it was determined that they predominantly encompass both tag RNA sequences and gene sequences. Notably, the clip length falls within the 20−100 nt range, signifying the effectiveness of our data-filtering steps. This inspection serves to validate the potential transcription of dpCoA-RNAs by the seven identified genes. However, further verification of these genes may be warranted to ascertain the authenticity of dpCoA-RNAs.

We would like to clarify the length of the model dpCoA-RNAs mentioned in the Comment stating that "The dpCoA-tagSeq method is based on specific addition of a maleimide containing 25 nt oligonucleotide sequence tag to the 5′ end of dpCoA-capped RNAs." Please note that we employed a 40

Received: November 22, 2023
Revised: December 12, 2023
Accepted: December 12, 2023

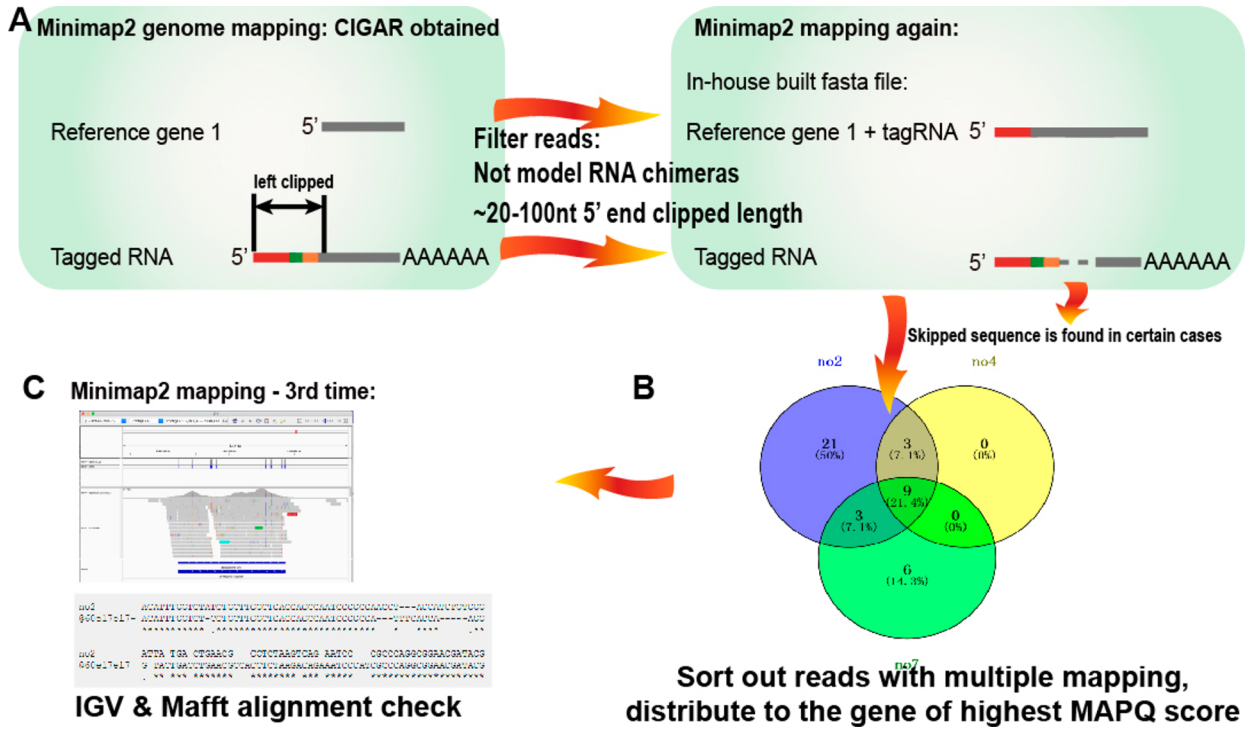

Figure 1. Data reprocessing procedure containing 3 gene mapping steps. Specific details regarding the steps taken for data reprocessing. (A) Reads that potentially amalgamated tagged model RNA with untagged mouse RNA reads were excluded, and then, CIGAR values were used to sort out the reads with clip lengths between 20 and 100 nt (possibly containing ~tag RNA). The filtered RNA reads were mapped against their tagged sequences. (B) Reads with multiple mappings were inspected, and only the mapping with the highest MAPQ value was retained. If the MAPQ value is the same, then the reads with the lowest clip length were maintained. If a read was mapped to multiple identical sequences (i.e., the exact repeats in the genome), all accession numbers and locus information were retained. (C) RNA reads were mapped against tagged gene sequences, and the mapping was examined using IGV and mafft to assess the alignment results.

Table 1. List of Gene Nos. 2−9 and the Filtered RNA Reads in the Reanalysis Procedure[a]

							PM+_1				PM-_1				PM+/PM-
	Geneid		Chr	Start	End	Strand	model-matched	not-model-matched	not-model #left 20:100	multiple mapping process	model-matched	not-model-matched	not-model #left 20:150	multiple mapping process	
no2	NONMMUG015781.2	LSU rRNA repeat	chr16	11143906	11144315	+	952	719	36	33	4	19	3	3	10.65
no3	NONMMUG044321.2	LSU rRNA repeat	chr9	123461799	123462187	+	317	202	9	9	1	7	0	0	90.00
no4	Unknown1(Gm25911)	LSU rRNA repeat	chr1	167340220	167340545	-	200	150	12	0	3	6	1	0	0.00
no5	RN45s	SSU rRNA repeat	chr17	39846353	39848201	+	267	217	26	26	1	4	1	1	23.64
			chr3;	5860338;	5860904;	+;									
			chr3;	5860338;	5860637;	+;									
no6	NONMMUG026007.2	LSU rRNA repeat	chr3;	5860339;	5860764;	+;	234	146	7	7	0	4	0	0	70.00
			chr3;	5860770;	5860826;	+;									
			chr3	6049675	6049710	+									
no7	Unknown2(GM24187)	LINE repeat	chr13	9834250	9834560	-	135	100	18	9	1	3	2	0	90.00
no8	NONMMUG009157.2	7SL RNA	chr12	69159295	69159591	+	112	87	10	10	0	2	1	0	100.00
no9	mt-Rnr2		chrM	1094	2675	+	146	104	7	7	0	2	0	0	70.00

[a]When calculating "Fold change (PM+/PM−)", "Reads in PM−" was +0.1 before getting divided.

nt tagRNA-PM for dpCoA tagSeq; the 25 nt tagRNA-PM was utilized only in gel analysis.

CONCLUSION

First and foremost, we express our sincere gratitude to Prof. Jeppe Vinther for his critical insights into the pitfalls of our bioinformatic analysis, which significantly enhanced the quality of our research. We hope the method development published in *Analytical Chemistry* may serve as an effective tool for determination of dpCoA-RNA after the discussions and modifications by experts in the field. Given the observation that chimeric reads primarily originate from the tagged model dpCoA-RNA, it is prudent to optimize the protocol for future applications. This optimization may encompass reducing the concentration of spiked-in model RNA and incorporating additional bioinformatic data processing steps to minimize chimeric reads from model RNA.

In addition to utilizing capped model RNA as a spike-in control, we recommend the inclusion of a negative control involving spike-in with uncapped RNA. To further enhance the reliability of putative capped RNA reads, it is necessary to implement a filtering analysis to exclude any reads containing the spike-in model sequence. Moreover, an alternative

https://doi.org/10.1021/acs.analchem.3c05281
Anal. Chem. XXXX, XXX, XXX−XXX

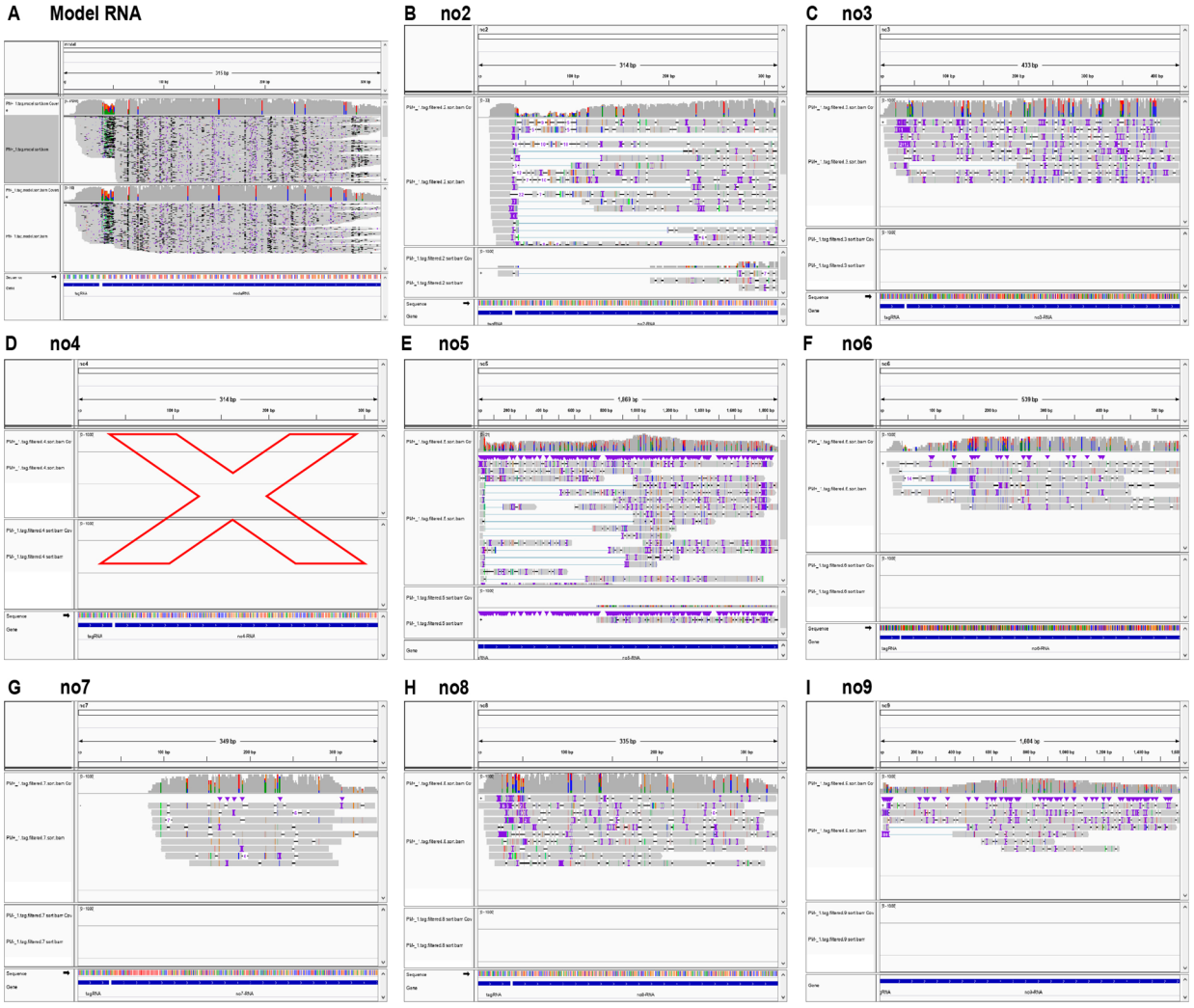

Figure 2. Genome mapping and visualization of (A) tagged model dpCoA-RNA and (B−I) other identified dpCoA-RNA candidates, nos. 2−9, in both PM− and PM+ groups.

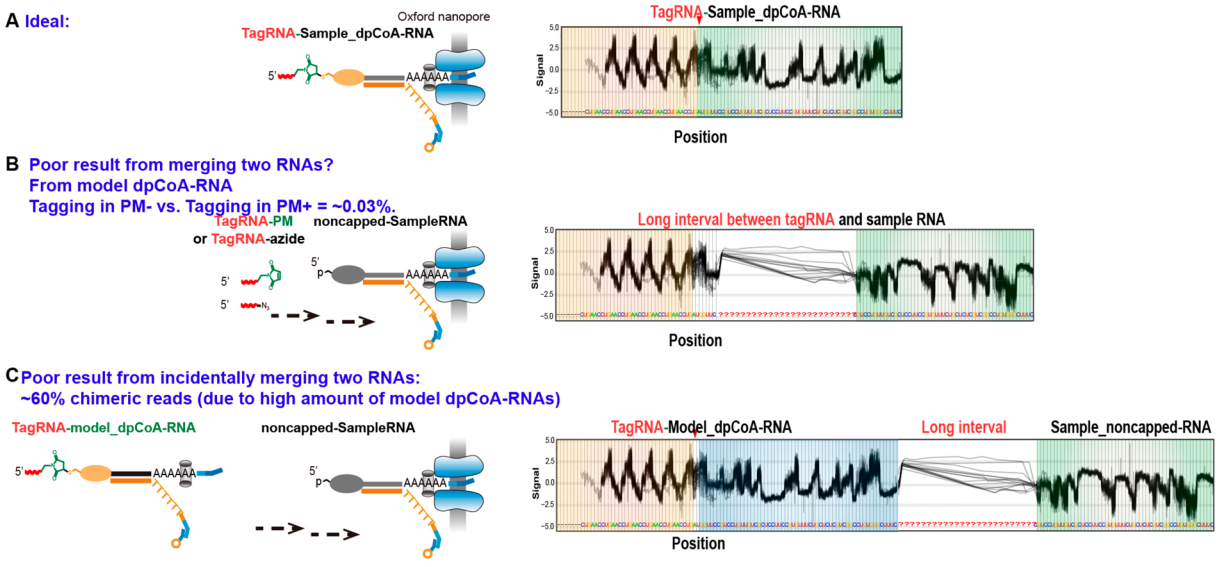

Figure 3. Results of nanopore's base-calling events and putative causes of false positives. (A) Ideal nanopore sequencing of tagged dpCoA-RNA. (B) Poor results arising from merging tagRNA-PM/tagRNA-azide and noncapped RNAs. (C) Poor results from incidentally merging tagged model dpCoA-RNA and noncapped RNAs. Please note the signal in panels B and C is not real data, which are adjusted from the real signal in panel A and only used for interpretation of possible chimeric RNA reads.

https://doi.org/10.1021/acs.analchem.3c05281
Anal. Chem. XXXX, XXX, XXX−XXX

strategy could involve the use of distinct tag RNA sequences to separately label model RNA and sample RNAs, providing flexibility in the experimental design.

■ ASSOCIATED CONTENT

Data Availability Statement

The software packages and codes used are available on GitHub (https://github.com/rocketjishao/tagSeq).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.3c05281.

Supporting Data 1 (ZIP)
Supporting Data 2 (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Zhu Yang; orcid.org/0000-0001-5934-1617
Yiji Xia
Zongwei Cai; orcid.org/0000-0002-8724-7684

Authors

Xiaojian Shao
Hailei Zhang
Zhou Zhu
Fenfen Ji
Zhao He

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.3c05281

Author Contributions

‡X.S. and H.Z. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors wish to thank the donation from the Kwok Chung Bo Fun Charitable Fund for the establishment of the Kwok Yat Wai Endowed Chair of Environmental and Biological Analysis and the Research Grants Council of Hong Kong (GRF grant no. C2009-19GF to Prof. Xia and Prof. Cai).

■ REFERENCES

(1) Shao, X.; Zhang, H.; Zhu, Z.; Ji, F.; He, Z.; Yang, Z.; Xia, Y.; Cai, Z. *Anal. Chem.* **2023**, *95*, 11124−11131.
(2) Shao, X.; Zhang, H.; Yang, Z.; Zhong, H.; Xia, Y.; Cai, Z. *Nat. Protoc.* **2020**, *15*, 2813−2836.
(3) Zhang, H.; Zhong, H.; Zhang, S.; Shao, X.; Ni, M.; Cai, Z.; Chen, X.; Xia, Y. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 12072−12077.
(4) Zhang, H.; Zhong, H.; Wang, X.; Zhang, S.; Shao, X.; Hu, H.; Yu, Z.; Cai, Z.; Chen, X.; Xia, Y. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, e2026183118.
(5) Hu, H.; Flynn, N.; Zhang, H.; You, C.; Hang, R.; Wang, X.; Zhong, H.; Chan, Z.; Xia, Y.; Chen, X. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, e2025595118.
(6) Rozewicki, J.; Li, S.; Amada, K. M.; Standley, D. M.; Katoh, K. *Nucleic Acids Res.* **2019**, *47*, W5−W10.
(7) White, R.; Pellefigues, C.; Ronchese, F.; Lamiable, O.; Eccles, D. *F1000Res* **2017**, *6*, 631.
(8) Martin, S.; Leggett, R. M. *BMC Bioinformatics* **2021**, *22*, 124.
(9) Payne, A.; Holmes, N.; Rakyan, V.; Loose, M. *Bioinformatics* **2019**, *35*, 2193−2198.

https://doi.org/10.1021/acs.analchem.3c05281
Anal. Chem. XXXX, XXX, XXX−XXX