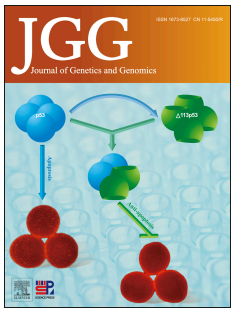
Journal Pre-proof



An efficient metatranscriptomic approach for capturing RNA virome and its application to SARS-CoV-2

Yao Meng, Liwen Xiao, Wenbing Chen, Fangqing Zhao, Xiang Zhao

PII: S1673-8527(21)00264-2

DOI: https://doi.org/10.1016/j.jgg.2021.08.005

Reference: JGG 948

To appear in: *Journal of Genetics and Genomics*

Received Date: 14 July 2021

Revised Date: 9 August 2021

Accepted Date: 16 August 2021

Please cite this article as: Meng, Y., Xiao, L., Chen, W., Zhao, F., Zhao, X., An efficient metatranscriptomic approach for capturing RNA virome and its application to SARS-CoV-2, *Journal of Genetics and Genomics*, https://doi.org/10.1016/j.jgg.2021.08.005.

An efficient metatranscriptomic approach for capturing RNA virome and its application to SARS-CoV-2

Coronavirus disease 2019 (COVID-19) is a new type of respiratory disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which has caused severe social and economic costs and has become a great threat to public health because of the high mortality rate. Recent studies indicated that respiratory pathogens co-infection and secondary infection were critical risk factors for the severity and mortality rates of COVID□19 (Zhou et al., 2020). Current laboratory-confirmed co-infections were those identified by pathogens culture, antigen detection, PCR detection and metatranscriptomic sequencing. Among them, metatranscriptomic sequencing has a unique advantage due to its ability to detect all types of pathogens. This method typically starts with RNA fragmentation, followed by first- and second-strand cDNA synthesis, and then moves on to library preparation specified by high throughput sequencing technologies of choice (Chiara et al., 2021). However, most preparation protocols involve multiple costly and complicated steps, which inevitably affects the efficacy in the sequencing of time-sensitive SARS-CoV-2 samples and in the identification of other pathogens. For this reason, an easy, cost-effective, and relatively unbiased protocol for SARS-CoV-2 sample sequencing is needed to address this issue.

A great advance in library preparation for whole genome sequencing (WGS) was the introduction of bacterial transposase Tn5, which is a member of the RNase H superfamily (Majorek et al., 2014). Its one-step tagmentation greatly simplified library preparation by fragmentating dsDNA and ligating synthetic oligonucleotides at both ends in 5 minutes. Recently, two studies reported that Tn5 can effectively fragment RNA/DNA heteroduplex and ligate the sequencing adaptors onto the hybrid without requiring second-strand synthesis, which was named SHERRY (Di et al., 2020) and TRACE-seq (Lu et al., 2020), respectively. SHERRY consists of three components: RNA reverse transcription, RNA/cDNA hybrid tagmentation, and PCR amplification, in which the final product is an indexed library that is used for sequencing. By optimizing the removal of bacterial rRNA and enriching virome sequences, we developed Tn5-mediated RNA sequencing with rRNA Removal (TRIM) based on SHERRY (Fig. 1A) and further compared its performance with that of other three traditional methods, including Illumina TruSeq® Stranded Total RNA Library Prep (TruSeq), NuGEN Trio RNA-Seq Library Preparation (Trio) and Sequence-independent, single-primer amplification (SISPA), respectively (Fig. S1A).

Samples collected from respiratory tract usually contain abundant host and bacterial ribosomal RNA (rRNA), which can impede the analysis of mRNA transcripts (Li et al., 2018). Consequently, removing rRNA prior to sequencing is required for optimal results. Commonly used commercial RNA-seq kits normally include a step to remove human rRNA, for example, probe capture process (Ribo-Zero™ technology) in TruSeq, and Insert Dependent Adaptor Cleavage (AnyDeplete technology) in Trio. In this study, TruSeq Stranded Total RNA Library Prep kit was used to remove human rRNA in all methods except Trio which has already included such a step after library construction. Bacterial rRNA were removed in TRIM using NEBNext® rRNA Depletion Kit (Bacteria). Then random hexamers and Oligo (dT)s were mixed to reverse transcribe the templates to obtain DNA/RNA hybrids. We compared the performance of the four methods on the sequencing of clinical samples. In total, 96 transcriptomic libraries were constructed from 24 samples using the four methods. Each library was individually sequenced on Illumina NextSeq 550 Sequencing System (Mid-Output Kit, paired-end 150 bp). Among the four methods, TRIM was the fastest and cost effective, which only needed 6 hours and $67 for building one library (Fig. 1B).

We next compared the ratio of rRNA in the metatranscriptomic datasets generated by the four methods. Because of the rRNA removal step, the ratio of bacterial rRNA reads in TRIM was 1.307%, significantly lower than that of the other methods (TruSeq = 12.74%; Trio = 52.01%; SISPA = 66.83%; Fig. 1C), but the eukaryotic rRNA ratio in TRIM was higher (Fig. S1B). However, the mean of the human rRNA ratio in TRIM (Fig. S1C) only accounted for 52% of eukaryotic rRNA, suggesting that human rRNA was not the only major source of eukaryotic rRNA in TRIM. Fungi is another important source of eukaryotic rRNA in respiratory samples, the most common of which is Candida (Zhou et al., 2021). We directly mapped the reads to the Candida rRNA reference sequence and found that the ratio in TRIM was significantly higher than that in the other three methods (Fig. S1D). Taken together, compared to other methods, TRIM exhibited the highest efficiency of rRNA removal (Fig. S1E and S1F).

We further compared the difference in the composition ratio of other non-rRNA sequences among the four methods. The proportion of human mRNA in TRIM was significantly higher than that in Trio and SISPA, but exhibited no significant difference compared to that in TruSeq (Fig. 1D). Similar to that of Candida rRNA, we detected the highest ratio of viral RNA sequences in TRIM among the four methods (Fig. 1E). It is suggested that the removal of host and bacterial rRNA resulted in a significant increase in the proportion of low-abundance RNA in the library. In the comparison of

bacteria mRNA in the library (Fig. 1F), the proportion in the TRIM library (median = 26.14%) was lower than that of Trio RNA (median = 47.09%), but both of them were significantly higher than TruSeq (median = 8.383%) and SISPA (median = 7.211%).

To measure the performance on the detection of virus, we compared the genome coverage and completeness of SARS-CoV-2 among the four methods (Figs. 1G, S2A and S2B). As shown in Fig. 1G, a large number of SARS-CoV-2 reads were detected using TRIM (TRIM = 1446; Trio-RNA =167; TruSeq = 93; SISPA = 107; reads per million (RPM)), indicating the high ability of our method to detect specific virus. Moreover, the number of samples that covered more than 90% of viral genome was the highest in TRIM (TRIM = 10, TruSeq = 9, Trio RNA = 9, SISPA = 1; Fig. S2A). To further evaluate the difference among the four methods, a regression analysis was implemented. With the exception of SISPA, a significant linear relationship was observed between coverage rate and Ct value in the other three methods (*P* value for each method: TRIM < 0.0001, TruSeq < 0.0001, Trio < 0.0001, SISPA = 0.0623, Fig. 1H), suggesting the consistency between metatranscriptomic sequencing and experimental validations. For samples with a Ct value higher than 26, the whole genome of SARS-CoV-2 could not be obtained by the four methods, indicating that metatranscriptomic sequencing is not suitable for samples with very low viral load (Xiao et al., 2020). Hybrid capture sequencing for $29 \leq Ct \leq 34$ samples, however, can still obtain the whole genome (Xiao et al., 2020; Xu et al., 2020), in which metatranscriptomic library preparation is a necessary preprocessing step for the hybrid capture sequencing.

Co-infection of SARS-CoV-2 with other respiratory viruses frequently occurred to some COVID-19 patients (Kim et al., 2020) and led them to death (Hashemi et al., 2021). Compared to bacterial, fungal or human genomes, viral genomes were too small to be detected in raw sequencing data. It was reported that there were less than 10 viral sequences per 25 million reads in the respiratory samples without any prior nucleic acid processing (Wylie et al., 2015), which affects the detection of the pathogen and greatly increases the sequencing cost. By counting the viral RPMs of all samples in the same method, we compared the detection efficiencies of the four methods for SARS-CoV-2 and other viruses (Fig. 1I) and found that TRIM was able to detect the highest number of viral reads. Among them, most of the non-SARS-CoV-2 virus reads were identified as bacteriophages of *Streptococcus* (Fig. S3A), which was the most abundant bacteria in respiratory tract (Fig. S3B). In addition to the coronavirus like SARS-CoV-2, TRIM was able to detect the full-length transcripts of the RPMS1 fragment of Epstein-Barr

virus (EBV) and the partial HA fragment of Influenza B virus (FluB) in two different samples (sample15 with FluB and sample02 with EBV). In contrast, the other three methods failed to detect Influenza B virus reads, and only TruSeq detected a small number of EBV reads (Fig. 1I). Co-infection of SARS-CoV-2 with both EBV and FluB were reported in previous studies (Cuadrado-Payán et al., 2020; Soler Rich et al., 2020). EBV is a DNA virus infected more than 90% of adults, due to the DNase I treatment in the pre-processing step, RNA-Seq cannot detect the EBV genome directly. When the EBV transcriptome is detected by RNA-Seq, it means that EBV is in an activated state. A previous study found that activated EBV can upregulate the ACE2 expression, which will make cells more susceptible to SARS-CoV-2 infection (Verma et al., 2021).

Identifying the pathogenic bacteria was a challenge in SARS-CoV-2 co-infection studies. Several studies from Wuhan found that secondary bacterial infections occurred in 50% of early deaths (Chen et al., 2020; Zhou et al., 2020), but whether there is a direct relationship between bacterial infections and deaths needs further confirmation. At the same time, clinical treatment of SARS-CoV-2 is characterized by antibiotic abuse. Two meta-analyses found that bacterial co-infections occurred in only 7% of hospitalized patients, but more than 90% of patients received broad-spectrum antibiotics (Lansbury et al., 2020; Youngs et al., 2020). Conducting pathogenic testing to identify the causative organism could help in more rational clinical use of antibiotics. Although the Trio RNA-seq library had the highest proportion of bacterial transcriptome (Fig. 1F), TRIM can detect a greater number of bacteria, which was 2.9–5.9 times higher than the other three methods (Fig. 1J). In addition, the consistency in the composition and the abundance of bacteria among the four methods indicates that TRIM exhibits a great performance in the detection of potential bacterial or viral pathogens (Fig. S3B and S3C).

In summary, our study demonstrates that TRIM is efficient to detect co-infection pathogens than traditional methods, and also shows great advantages on cost and time usage, which should have wide applications in clinical research and diagnosis of SARS-CoV-2 co-infection, helping to determine the disease and to control the use of antibiotics, and promoting research on the mechanism of viral infection and microbial interactions.

Conflict of interest

All authors declare no conflict of interests.

Acknowledgments

This work was supported by grants from National Natural Science Foundation of China (32025009, 91951209, 31722031). The study was approved by the ethics review committee of the Chinese Center for Disease Control and Prevention (IVDC2021-001).

References

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y.*, et al.*, 2020. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The Lancet. 395, 507-513.

Chiara, M., D'Erchia, A.M., Gissi, C., Manzari, C., Parisi, A., Resta, N., Zambelli, F., Picardi, E., Pavesi, G., Horner, D.S.*, et al.*, 2021. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. Brief Bioinform. 22, 616-630.

Cuadrado-Payán, E., Montagud-Marrahi, E., Torres-Elorza, M., Bodro, M., Blasco, M., Poch, E., Soriano, A.,Piñeiro, G.J., 2020. SARS-CoV-2 and influenza virus co-infection. The Lancet. 395.

Di, L., Fu, Y., Sun, Y., Li, J., Liu, L., Yao, J., Wang, G., Wu, Y., Lao, K., Lee, R.W.*, et al.*, 2020. RNA sequencing by direct tagmentation of RNA/DNA hybrids. Proc. Natl. Acad. Sci. U. S. A. 117, 2886-2893.

Hashemi, S.A., Safamanesh, S., Ghasemzadeh-Moghaddam, H., Ghafouri, M.,Azimian, A., 2021. High prevalence of SARS-CoV-2 and influenza A virus (H1N1) coinfection in dead patients in Northeastern Iran. J. Med. Virol. 93, 1008-1012.

Kim, D., Quinn, J., Pinsky, B., Shah, N.H.,Brown, I., 2020. Rates of Co-infection Between SARS-CoV-2 and Other Respiratory Pathogens. JAMA 323, 2085-2086.

Lansbury, L., Lim, B., Baskaran, V.,Lim, W.S., 2020. Co-infections in people with COVID-19: a systematic review and meta-analysis. J. Infect. 81, 266-275.

Li, F., Kaczor-Urbanowicz, K.E., Sun, J., Majem, B., Lo, H.C., Kim, Y., Koyano, K., Rao, S.L., Kang, S.Y., Kim, S.M.*, et al.*, 2018. Characterization of Human Salivary Extracellular RNA by Next-generation Sequencing. Clin. Chem. 64, 1085-1095.

Lu, B., Dong, L., Yi, D., Zhang, M., Zhu, C., Li, X.,Yi, C., 2020. Transposase-assisted tagmentation of RNA/DNA hybrid duplexes. Elife 9.

Majorek, K.A., Dunin-Horkawicz, S., Steczkiewicz, K., Muszewska, A., Nowotny, M., Ginalski, K.,Bujnicki, J.M., 2014. The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. Nucleic. Acids. Res. 42, 4160-4179.

Soler Rich, R., Rius Tarruella, J.,Melgosa Camarero, M.T., 2020. Expanded mesenchymal stem cells: a novel therapeutic approach for SARS-CoV-2 pneumonia (COVID-19). Concepts regarding a first case in Spain. Med. Clin. (Engl Ed) 155, 318-319.

Verma, D., Church, T.M.,Swaminathan, S., 2021. Epstein-Barr Virus Lytic Replication

Induces ACE2 Expression and Enhances SARS-CoV-2 Pseudotyped Virus Entry in Epithelial Cells. J. Virol. 95, e0019221.

Wylie, T.N., Wylie, K.M., Herter, B.N.,Storch, G.A., 2015. Enhanced virome sequencing using targeted sequence capture. Genome. Res. 25, 1910-1920.

Xiao, M., Liu, X., Ji, J., Li, M., Li, J., Yang, L., Sun, W., Ren, P., Yang, G., Zhao, J.*, et al.*, 2020. Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. Genome. Med. 12, 57.

Xu, Y., Kang, L., Shen, Z., Li, X., Wu, W., Ma, W., Fang, C., Yang, F., Jiang, X., Gong, S.*, et al.*, 2020. Dynamics of severe acute respiratory syndrome coronavirus 2 genome variants in the feces during convalescence. J. Genet. Genomics 47, 610-617.

Youngs, J., Wyncoll, D., Hopkins, P., Arnold, A., Ball, J.,Bicanic, T., 2020. Improving antibiotic stewardship in COVID-19: Bacterial co-infection is less common than with influenza. J. Infect 81, e55-e57.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X.*, et al.*, 2020. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet. 395, 1054-1062.

Zhou, X., Wang, H., Ji, Q., Du, M., Liang, Y., Li, H., Li, F., Shang, H., Zhu, X., Wang, W.*, et al.*, 2021. Molecular deconvolution of the neutralizing antibodies induced by an inactivated SARS-CoV-2 virus vaccine. Protein Cell. https://doi.org/10.1007/s13238-021-00840-z

Figure legend

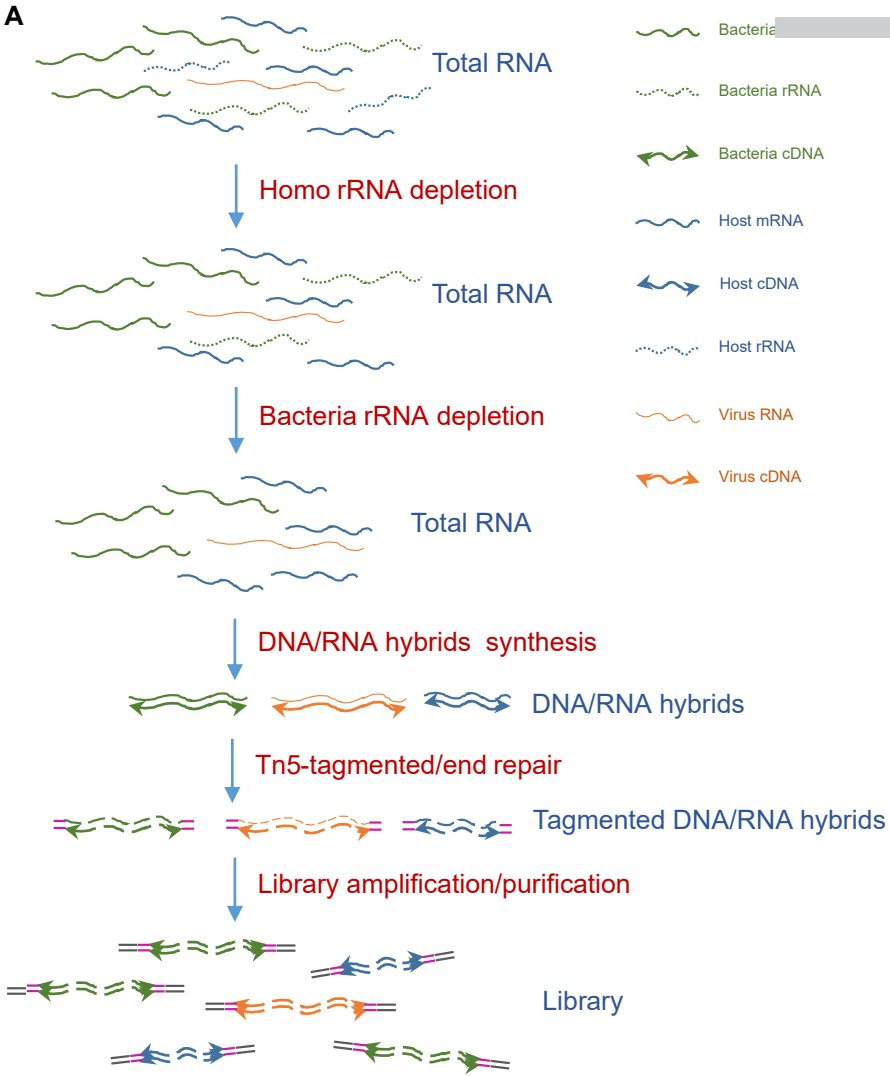
Fig. 1. Comparison of the four methods in the detection of viruses and bacteria. **A**: Workflow of library construction of TRIM. **B**: Protocol and detailed consumption in the library construction of the four methods. **C**–**F**: Comparison of reads mapping ratio of the four methods. Ratio of reads mapped to Bacteria ribosomal RNA (**C**), Human mRNA (**D**), Virus total RNA (**E**), Bacteria mRNA (**F**). **G**: The number of reads mapped to SARS-CoV-2 reference genome (NC45512.2), shown as reads per million (RPM). **H**: Simple linear regression analysis between SARS-Co*V-2 genome coverage rate and Ct value. **I**: Cumulative reads of 24 samples mapped to common respiratory viruses. Other viruses correspond to the viruses except for SARS-CoV-2 in the nr database. The abundance of EBV and FluB was detected using a custom viral database. **J**: Richness of bacterial genus detected from four methods. Boxplots show median (horizontal line in the box), lower and upper quantile (box), lower and upper 1.5 times interquartile range (whiskers). Wilcoxon rank-sum test, *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.*

Yao Meng[1]
Beijing Institutes of Life Science, Chinese Academy of Sciences,
Beijing 100101, China
Shaanxi Provincial Center for Disease Control and Prevention,
Xi'an 710054, China

Liwen Xiao[1]*, Wenbing Chen, Fangqing Zhao**
Beijing Institutes of Life Science, Chinese Academy of Sciences,
Beijing 100101, China
University of Chinese Academy of Sciences,
Beijing 100101, China

*Xiang Zhao**
State Key Laboratory for Manufacturing Systems Engineering,
Xi'an Jiaotong University, Xi'an 710054, China

[1] *These authors contributed equally to this work.*
**Corresponding authors.*
E-mail addresses: zhfq@biols.ac.cn (F. Zhao) zhaoxiang@cnic.org.cn (X. Zhao)



B

Protocol	Vendor	Catalog #	Duration of the procedure	Hands-On Time	Price/Rxn
TruSeq® Stranded Total RNA Library Prep	Illumina	20020613	480 min	145 min	$137.00
Trio RNA-Seq Library Preparation Kit	Tecan Genomics	0506-96	645 min	145 min	$175.00
Sequence-independent, single-primer amplification (SISPA)	-	-	515 min	138 min	$30
Tn5-mediated RNA sequencing with rRNA Removal (TRIM)	-	-	240[a]-360 min	90 min	$27[a]~$67

[a]Time and money cost without rRNA depletion step