MINERVA: A facile strategy for SARS-CoV-2 whole genome deep sequencing of clinical samples

Chen Chen[a,#], Jizhou Li[b,#], Lin Di[c,d#], Qiuyu Jing[e,#], Pengcheng Du[a,#], Chuan Song[a], Jiarui Li[a], Qiong Li[b], Yunlong Cao[c], X. Sunney Xie[c], Angela R. Wu[e,f], Hui Zeng[a,*], Yanyi Huang[c,g,h,i,*], Jianbin Wang[b,i,j,*]

[a] Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University and Beijing Key Laboratory of Emerging Infectious Diseases, Beijing 100015, China.

[b] School of Life Sciences, Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing 100084, China.

[c] Beijing Advanced Innovation Center for Genomics (ICG), Biomedical Pioneering Innovation Center (BIOPIC), Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China.

[d] School of Life Sciences, Peking University, Beijing 100871, China.

[e] Division of Life Science, Hong Kong University of Science and Technology, Hong Kong SAR, China.

[f] Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Hong Kong SAR, China

[g] College of Chemistry and Molecular Engineering, Beijing 100871, China

[h] Institute for Cell Analysis, Shenzhen Bay Laboratory, Guangdong 518132, China

[i] Chinese Institute for Brain Research (CIBR), Beijing 102206, China.

[j] Beijing Advanced Innovation Center for Structural Biology (ICSB), Tsinghua University, Beijing 100084, China.

[#] These authors contributed equally to this work.

* Corresponding authors: jianbinwang@tsinghua.edu.cn (J.W.), yanyi@pku.edu.cn (Y.H.) and zenghui@ccmu.edu.cn (H.Z.).

Abstract

The novel coronavirus disease 2019 (COVID-19) pandemic poses a serious public health risk. Analyzing the genome of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) from clinical samples is crucial for the understanding of viral spread and viral evolution, as well as for vaccine development. Existing sample preparation methods for viral genome sequencing are demanding on user technique and time, and thus not ideal for time-sensitive clinical samples; these methods are also not optimized for high performance on viral genomes. We have developed MetagenomIc RNA EnRichment VirAl sequencing (MINERVA), a facile, practical, and robust approach for metagenomic and deep viral sequencing from clinical samples. This approach uses direct tagmentation of RNA/DNA hybrids using Tn5 transposase to greatly simplify the sequencing library construction process, while subsequent targeted enrichment can generate viral genomes with high sensitivity, coverage, and depth. We demonstrate the utility of MINERVA on pharyngeal, sputum and stool samples collected from COVID-19 patients, successfully obtaining both whole metatranscriptomes and complete high-depth high-coverage SARS-CoV-2 genomes from these clinical samples, with high yield and robustness. MINERVA is compatible with clinical nucleic extracts containing carrier RNA. With a shortened hands-on time from sample to virus-enriched sequencing-ready library, this rapid, versatile, and clinic-friendly approach will facilitate monitoring of viral genetic variations during outbreaks, both current and future.

Introduction

As of April 24, 2020, the ongoing COVID-19 viral pandemic has affected more than 2.6 million people in over 200 countries and territories around the world, and has claimed more than 180 thousand lives. Closely monitoring the genetic diversity and distribution of viral strains at the population level is essential for epidemiological tracking, and for understanding viral evolution and transmission; additionally examining the viral heterogeneity within a single individual is imperative for diagnosis and treatment. The disease-causing pathogen, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was identified from early disease cases and its draft genome sequenced within weeks, thanks to the rapid responses from researchers around the world(*1*, *2*). The initial SARS-CoV-2 draft genome was obtained independently from the same early COVID-19 patient samples using various conventional RNA-seq sequencing library construction methods. Although these library construction methods successfully generated a draft genome, several drawbacks hinder the use of these methods for routine viral genome sequencing from the surge of clinical samples during an outbreak.

One direct library construction approach which was used to generate the SARS-CoV-2 draft genome(*1*, *2*) essentially captures each sample's entire metatranscriptome, in which SARS-CoV-2 is just one species among many. The abundance of SARS-CoV-2 in clinical swabs, sputum, and stool samples is often low(*3*, *4*), therefore this catch-all method requires deeper sequencing of each sample in order to obtain sufficient coverage and depth of the whole viral genome, which increases the time and cost of sequencing. Target enrichment with spiked-in primers can improve SARS-CoV-2 genome coverage(*5*), but the reliance on specific primers inherently limits this approach for the profiling of fast evolving viruses such as coronaviruses. The same limitation applies to multiplex RT-PCR-based strategies(*6*). Additionally, once the sample is subject to targeted amplification during the initial reverse transcription (RT) steps, its metatranscriptomic information is lost forever.

Currently, the most comprehensive strategy is the combination of metatranscriptomics profiling with post-library SARS-CoV-2 target enrichment(*6*). However, in most

conventional RNA-seq methods, the double-strand DNA ligation (dsDL) portion of the protocol is usually the most demanding on hands-on time and user technique(*7*). When superimposed on the target enrichment process, these labor intensive and lengthy protocols become impractical for routine use in the clinic, much less for the timely monitoring of viral genetics and evolution on large volumes of samples during an outbreak. Furthermore, due to the low molecular efficiency of dsDL, these protocols also require a high amount of input material, further restricting their application on clinical samples.

Summarily, although next generation sequencing platforms are high-throughput and have short turn-around time, library construction from samples – whether including targeted enrichment or not – remains a major bottleneck. To broadly apply viral sequencing on clinical samples, especially during outbreaks when biomedical resources are already limited, a rapid, simple, versatile, and scalable sample library construction method that does not compromise on performance is urgently needed.

Recently, we reported a new RNA-seq library construction strategy that aims to address some of these challenges: SHERRY avoids the problematic dsDL step in library construction by taking advantage of the newly discovered Tn5 tagmentation activity on RNA/DNA hybrids, to directly tag RNA/cDNA fragments with sequencing adapters (*7*). As such, SHERRY has minimal sample transfers and greatly reduced hands-on time, making it simple, robust, and suitable for inputs ranging from single cells to 200 ng total RNA. We now combine the advantages of SHERRY with a simplified post-library target enrichment protocol to enable rapid sequencing of high-quality viral genomes from clinical samples. MetagenomIc RNA EnRichment VirAl sequencing or MINERVA, is an easy-to-use, versatile, scalable, and cost-effective protocol that yields high-coverage high-depth viral genomes, while preserving the sample's rich metatranscriptomic profile. The hands-on time required from clinical sample to sequencing-ready library using conventional approaches without enrichment is 190 mins; MINERVA requires only 100 mins hands-on time, and if deep viral coverage is desired, an additional 90 mins for post-library enrichment, totaling 190 mins for the entire workflow (**Fig. S1**), making MINERVA practical for high-volume, routine clinical use. We applied MINERVA to various types of COVID-19

samples and successfully obtained up to 10,000-fold SARS-CoV-2 genome enrichment. This strategy will facilitate all studies regarding SARS-CoV-2 genetic variations in the current pandemic, and can also be applied to other pathogens of interest.

Results

Metagenomic RNA enrichment viral sequencing (MINERVA). To analyze both metagenomics and SARS-CoV-2 genetics from COVID-19 patient samples, we developed a two-stage metagenomic RNA enrichment viral sequencing strategy termed MINERVA (**Fig. 1**). First, we employed a SHERRY-based RNA-seq pipeline for metagenomic analysis. Since clinical samples may contain DNA, RNA, and possibly carrier RNA, MINERVA starts with ribosomal RNA (rRNA) removal and optional simultaneous carrier RNA removal, followed by DNase I treatment. The remaining RNA is then subject to standard SHERRY. Previously we observed 3' bias in SHERRY libraries; to address this, we used 10 ng mouse 3T3 cell total RNA as starting material, and tested whether adding random decamers (N10) during RT could improve coverage evenness (**Fig. S2**). Compared with the standard SHERRY protocol, which uses 1 μM T30VN primer during RT, the supplement of 1 μM N10 indeed improves gene body coverage evenness, presumably by improving the RT efficiency. When the N10 concentration was further increased to 10 μM, we observed almost no coverage bias in the gene body. The high N10 concentration can result in an increased rRNA ratio in the product, sometimes as high as 90%, but MINERVA employs rRNA removal as the first step prior to RT, thus negating this problem. We also performed enzyme titration with homemade and commercial Tn5 transposomes. Based on these N10 and Tn5 titration results, we used 10 μM N10 during RT and 0.5 μl V50 for each 20-μl tagmentation reaction in all following experiments. The whole procedure from nucleic acid to metagenomic sequencing-ready library, including wait time, takes 5.5 hours (**Fig. S1**).

For target enrichment, we first quantified SARS-CoV-2 abundance in each metagenomic sequencing library using an N gene qPCR assay, and pooled eight libraries based on quantification results. Then we performed standard in-solution hybridization on the pooled library with biotinylated RNA probes covering the whole viral genome. The enrichment

147 procedure takes 19 hours; the entire MINERVA pipeline can be completed within 24 hours.
148
149 **MINERVA achieves better SARS-CoV-2 genome coverage compared to**
150 **conventional dsDL strategies.** To evaluate its performance on clinical samples, we
151 applied MINERVA on 85 pharyngeal swabs, sputum, or stool samples, collected from 72
152 COVID-19 patients. These patients were admitted to Ditan Hospital within a two-month
153 period from January to March 2020, presenting different symptom severity (**Fig. 2A and**
154 **Fig. S3**). Some patients were re-sampled longitudinally to investigate temporal and intra-
155 host viral heterogeneity. We also used these samples to benchmark MINERVA against
156 conventional dsDL strategies. Samples Y1-Y6 were excluded from dsDL processing as
157 this processing strategy is incompatible with samples containing carrier RNA. On average,
158 we sequenced 1-3 Gbp for each MINERVA library pre- and post-enrichment, and nearly
159 100 Gbp for each dsDL library (**Fig. 2B**).
160
161 The metagenomic composition of SHERRY and dsDL libraries were comparable: total
162 virus, fungus and bacteria ratios were highly concordant between the two methods (**Fig.**
163 **S4**); bacterial heterogeneity revealed by entropy is also correlated between the two
164 strategies. In both MINERVA and dsDL data, we detected low yet significant levels of
165 SARS-CoV-2 sequences. The viral ratio is between 10^{-7} and 10^{-1}. It is worth noting that
166 the SARS-CoV-2 sequence ratio is higher in MINERVA data than in dsDL data (**Fig. 2C**
167 **and 2D**), suggesting that MINERVA libraries capture more SARS-CoV-2 sequences. This
168 phenomenon was more prominent for low viral load samples. Though SARS-CoV-2
169 genome coverage and depth was not high in SHERRY results due to low viral ratio and
170 low sequencing depth, performing MINERVA subsequently can enrich the SARS-CoV-2
171 sequence ratio up to 10,000-fold (**Fig. 2E and S5**). As a result, MINERVA gives more
172 complete and deeper coverage of SARS-CoV-2 genomes (**Fig. 2F and 2G**), despite
173 sequencing dsDL libraries to two orders of magnitude more depth (**Fig. 2B**). We achieved
174 even higher quality of MINERVA data by scaling up the reaction volume (**Fig. S5**). Using
175 the same samples and the same amount of sequencing data, more input in a higher
176 reaction volume resulted in deeper SARS-CoV-2 genome coverage.
177

The superior quality of MINERVA data became clearer when we included clinical RT-qPCR results. Both dsDL and MINERVA libraries detect SARS-CoV-2 sequences for samples with various Ct values, but MINERVA produced more complete and deeper genome coverage than dsDL methods (**Fig. 2H and 2I**), and this advantage is more pronounced for low viral load samples, including two samples with negative qPCR results.

Carrier RNA, which is widely used in viral DNA/RNA extraction before RT-qPCR assays, severely impacts high-throughput sequencing analysis. Therefore, most RT-qPCR positive clinical samples are not amenable to further viral genetic studies. We explored the effect of adding polyT oligos during the rRNA removal step to simultaneously remove spike-in polyA RNA and carrier RNA. By incorporating this step in MINERVA, we successfully avoided the overwhelming representation of unwanted RNA sequences while retaining desired human and viral sequences (**Fig. 2J**).

MINERVA can be used to investigate multiple facets of COVID-19 biology. Among the three sample types, pharyngeal swabs, sputum, and stool samples, MINERVA showed the biggest improvement over dsDL for stool samples (**Fig. 2H**), which prompted us to investigate these samples further. By assessing the relative compositions of bacterial species and viral species in each sample type, we noted that the bacteria largely account for the difference between stool samples and the other two samples types, not viruses (**Fig. 3A and S6**). This is likely due to the unique relationship between the microbiome and its environment, subject to host-specific variations. In the stool samples, the bacterial abundance and diversity are highest, and are most likely to dominate and obscure signals coming from viral species. Since dsDL approaches are less sensitive and require more input, this may explain why MINERVA outperforms dsDL most evidently in stool samples.

Following this line of investigation based on metagenomic composition, we further performed principle component analysis of specifically the bacterial sequences and confirmed clear separation between stool sample and the other two sample types along PC1 (**Fig. 3B**). Stool samples contained more *Bacteroides*, whereas the pharyngeal and

sputum samples were rich in *Streptococcus* (**Fig. 3C and S7**). There also appears to be separation between samples by COVID-19 symptom severity along PC2 (**Fig. 3B**); specifically, we found samples from severe and critical condition patients to be abundant in *Comamonas* rather than *Streptococcus*, which is abundant in mild and severe condition samples (**Fig. 3C**). SARS-CoV-2 mapping ratio of SHERRY libraries further confirmed that virus is more abundant in samples from critically ill patients (**Fig. 3D and S8**).

As a novel virus, little known about the evolutionary features of SARS-CoV-2. Recent studies have identified genetic variations and raised the possibility that multiple variants could co-exist in the same host individual. We evaluated the performance of MINERVA in detecting SARS-CoV-2 genetic variations. Aided by deep coverage of the viral genome (**Fig. 3E**), MINERVA detects genetic variations more effectively than dsDL. Using 85 samples, we constructed a SARS-CoV-2 mutational profile (**Fig. 3E**), which was distinct from the Guangdong profile(*8*). A few mutation sites, including the two linked to S and L strains, were found in multiple samples. Surprisingly, despite high linkage between these two positions, there were 6 samples in which we detected only the T28144C mutation but not the C8782T mutation. Further investigation showed that the T28144C mutant fractions in these 6 samples were all <87%, in contrast to >93% in the other 7 double-mutant samples. These results, together with the intermediate mutation fractions in many other sites, support the co-existence of multiple variants in the same individual. Further investigation is required to understand this phenomenon.

In summary, MINERVA effectively converts SARS-CoV-2 genomes into sequencing libraries with a simple and quick experimental pipeline, and subsequent target enrichment can further improve SARS-CoV-2 genome coverage and genetic variation detection. MINERVA can facilitate the study of SARS-CoV-2 genetics, and be easily implemented to fight future RNA pathogen outbreaks.

Discussion

As of today, our knowledge of SARS-CoV-2 is still preliminary and much of it extrapolated from past studies of other beta coronaviruses such as SARS and MERS. However, the

epidemiology, physiology, and biology of COVID-19 are evidently unique(*9*). To speed up our investigation of this virus and the disease it causes, a practical protocol for viral genome research of clinical samples is urgently needed. Currently, methods for transforming clinical samples into sequencing libraries are laborious and painstaking, while clinical personnel at the frontlines are already strained for time and energy. MINERVA minimizes the need for expert technique and hands-on operation; we believe it will be pivotal in accelerating clinical research of SARS-CoV-2.

Recent evolutionary tracing studies suggest the emergence of multiple novel, evolved subtypes of SARS-CoV-2(*10*), including the S/L-subtypes(*11*) and the A/B/C-variants(*12*). New variants will likely continue to emerge as the virus mutates, and to uncover them requires deep, complete coverage of viral genomes from a large number of patients. With the existence of asymptomatic carriers(*13*) and possible recurrent infections in the same individual(*14*), longitudinal re-sampling of patients is also important to uncover intra-host viral heterogeneity, but as viral load decreases with time(*15*), the sensitivity of the sample processing method becomes critical. These studies all require processing large volumes of clinical samples with a highly robust and scalable method that does not compromise on sensitivity. We have demonstrated that MINERVA libraries from clinical samples can generate deep and complete coverage of SARS-CoV-2 genomes that can be used for evolutionary tracing and variant characterization research.

It is well-established now that SARS-CoV-2 can infect multiple organ systems, tissue compartments, and cell types(*3*, *4*, *16*, *17*). In our profiling of COVID-19 clinical samples from multiple body sites of the same patient, we found that the viral load and viral subtypes vary across different body sites, possibly affected by interactions between microbial and other viral species as well as overall metagenomic diversity present in different microenvironments of each body site. The effects of metatranscriptomic diversity and inter-compartment heterogeneity on SARS-CoV-2 biology and COVID-19 symptom severity are also not understood. In particular, it is difficult to obtain high-quality unbiased metatranscriptomes using conventional library construction methods from low-quantity samples, as well as samples such as stool in which bacteria dominate the

metatranscriptomes, as conventional methods are not sufficiently sensitive. The versatility of MINERVA as a two-part protocol integrating SHERRY and post-library virus enrichment provides flexibility for sample processing that uses one standard sample pipeline for both highly sensitive metatranscriptomic analysis and targeted deep sequencing of specific transcripts.

MINERVA was not created to be a rapid diagnostic assay; rather, we hope its ease-of-use, versatility, scalability, sensitivity, and cost-effectiveness will drive adoption of routine sequencing of COVID-19 clinical samples, and thereby facilitate multiple areas of much-needed SARS-CoV-2 and COVID-19 research for clinicians and researchers.

Author contributions

C.C., Y.C., X.S.X., H.Z., Y.H. and J.W. conceived the project; J.L., P.D., Q.L. and C.S. conducted experiments; C.C., L.D., Q.J., J.L., Y.H. and J.W. analyzed the data; C.C., J.L., L.D., Q.J., A.R.W., Y.H. and J.W. wrote the manuscript with the help from all other authors.

Conflict of interest statement

The authors declare no conflict of interest.

Acknowledgement

We thank Ms. Chenyang Geng and BIOPIC sequencing platform at Peking University for the assistance of high-throughput sequencing experiments, and Ms. Amelia Huang for the assistance of figure preparation. This work was supported by National Natural Science Foundation of China (21675098, 21927802, 21525521), Ministry of Science and Technology of China (2018YFA0800200, 2018YFA0108100, 2018YFC1002300), 2018 Beijing Brain Initiation (Z181100001518004), Beijing Advanced Innovation Center for Structural Biology, Beijing Advanced Innovation Center for Genomics, HKUST's start-up and initiation grants (Hong Kong University Grants Committee), Hong Kong Research Grants Council Theme-based Research Scheme (RGC TBRS T12-704/16R-2) and Collaborative Research Fund (RGC CRF C6002-17G), Hong Kong RGC Early Career

Support Scheme (RGC ECS 26101016), Hong Kong Epigenomics Project (LKCCFL18SC01-E), and HKUST BDBI Labs.

References

1. L.-L. Ren *et al.*, Identification of a novel coronavirus causing severe pneumonia in human. *Chinese Medical Journal*. Online Publication (2020). doi: 10.1097/CM9.0000000000000722.

2. R. Lu *et al.*, Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. **395**, 565–574 (2020). doi: 10.1016/S0140-6736(20)30251-8.

3. Wölfel *et al.*, Virological assessment of hospitalized patients with COVID-2019. *Nature*, Online Publication (2020). doi: 10.1038/s41586-020-2196-x.

4. W. Wang *et al.*, Detection of SARS-CoV-2 in Different Types of Clinical Specimens. *JAMA*, Online Publication (2020). doi: 10.1001/jama.2020.3786.

5. X. Deng, A Genomic Survey of SARS-CoV-2 Reveals Multiple Introductions into Northern California without a Predominant Lineage. *medRxiv*, Online Publication (2020). doi: 10.1101/2020.03.27.20044925.

6. M. Xiao, Multiple approaches for massively parallel sequencing of HCoV-19 (SARS-CoV-2) genomes directly from clinical samples. *bioRxiv*, Online Publication (2020). doi: 2020.03.16.993584.

7. L. Di *et al.*, RNA sequencing by direct tagmentation of RNA/DNA hybrids. *Proc Natl Acad Sci USA*. **117**, 2886–2893 (2020). doi: 10.1073/pnas.1919800117.

8. J. Lu *et al.*, Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell*. Online Publication (2020). doi: 10.1016/j.cell.2020.04.023.

9. A. S. Fauci, H. C. Lane, R. R. Redfield, Covid-19 — Navigating the Uncharted. *N Engl J Med*. **382**, 1268–1269 (2020). doi: 10.1056/NEJMe2002387.

10. D. F. Gudbjartsson *et al.*, Spread of SARS-CoV-2 in the Icelandic Population. *N Engl J Med*, Online Publication (2020). doi: 10.1056/NEJMoa2006100.

11. X. Tang *et al.*, On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev*. Online Publication (2020). doi: 10.1093/nsr/nwaa036.

12. P. Forster, L. Forster, C. Renfrew, M. Forster, Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci USA*. Online Publication (2020). doi: 10.1073/pnas.2004999117.

13. Y. Bai *et al.*, Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA*. **323**, 1406–1407 (2020). doi: 10.1001/jama.2020.2565.

14. J. An *et al.*, Clinical characteristics of the recovered COVID-19 patients with re-detectable positive RNA test. *medRxiv*, Online Publication (2020). doi: 2020.03.26.20044222

15. X. He *et al.*, Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, Online Publication (2020). doi: 10.1038/s41591-020-0869-5.

16. B. E. Young *et al.*, Epidemiologic Features and Clinical Course of Patients Infected With SARS-CoV-2 in Singapore. *JAMA*. **323**, 1488–1494 (2020). doi: 10.1001/jama.2020.3204.

17. C. Chen *et al.*, SARS-CoV-2–Positive Sputum and Feces After Conversion of Pharyngeal Samples in Patients With COVID-19. *Annals of Internal Medicine*, Online Publication (2020). doi: 10.7326/M20-0991.

Figure Caption

Figure 1. Scheme of MINERVA. RNA extracted from pharyngeal swabs, sputum and stool samples undergo rRNA and DNA removal before a SHERRY processing pipeline metagenomic sequencing library construction. Multiple libraries were then pooled for SARS-CoV-2 sequence enrichment.

Figure 2. Direct comparison between MINERVA and the conventional dsDL strategy. (A) COVID-10 sample profiles. (B) Amount of sequencing data for different libraries. (C) SARS-CoV-2 mapping ratio statistics of SHERRY and dsDL libraries. (D) Comparison of SARS-CoV-2 mapping ratios between SHERRY and dsDL libraries. (E) Comparison of SARS-CoV-2 mapping ratios between SHERRY and MINERVA libraries. (F and G) SARS-CoV-2 genome coverage and depth statistics of MINERVA and dsDL libraries. (H and I) Comparison of SARS-CoV-2 sequencing results between MINERVA and dsDL libraries. (J) Removal of carrier RNA.

Figure 3. MINERVA could facilitate COVID-10 and SARS-CoV-2 research. (A) Total viral and bacterial ratios in different sample types. (B) Principle component analysis of bacteria composition. (C) Bacterial composition features of each sample. (D) SARS-CoV-2 map ratio of SHERRY libraries from patients with different severity. (E) SARS-CoV-2 mutation profiles.

Figure S1. Comparison of workflow between MINERVA and the conventional dsDL strategy.

Figure S2. Optimization of SHERRY protocol. (A-C) Effect of N10 primer during reserve transcription and Tn5 amount on detected gene number, ribosomal rate and insert size. (D-F) Effect of N10 primer during reserve transcription and Tn5 amount on gene body coverage evenness.

Figure S3. Sample collection profile.

Figure S4. Comparison between SHERRY and dsDL libraries on total viral ratio (A), total fungal ratio (B), total bacterial ratio (C), and bacterial entropy (D).

Figure S5. SARS-CoV-2 genome sequencing results of MINERVA libraries. (A) SARS-CoV-2 mapping ratio statistics of MINERVA libraries. (B and C) SARS-CoV-2 genome coverage and depth statistics of SHERRY libraries. (D) Effect of sample input and reaction volume on SARS-CoV-2 genome sequencing depth. (E) Results of carrier RNA removal.

Figure S6. Metagenomic profiles and the relationship between SARS-CoV-2 map ratio and RT-qPCR Ct values.

Figure S7. Individual bacterial ratio in samples.

Figure S8. SARS-CoV-2 map ratio of dsDL libraries from patients with different severity.

Material and Methods

Ethics approval

This study was approved by the Ethics Committee of Beijing Ditan Hospital, Capital Medical University (No. KT2020-006-01).

Optimization of SHERRY protocol

We used the total RNA extracted from 3T3 cells to optimize experimental protocols. RNA extraction was performed using RNeasy Mini Kit (Qiagen, Cat.No.74104). DNA was then removed through DNase I (NEB, Cat.No.M0303) digestion. The resulting total RNA was concentrated by RNA Clean & Concentrator-5 kit (Zymo Research, Cat R1015), and its quality was assessed by the Fragment Analyzer Automated CE System (AATI). Its quantification was done by Qubit 2.0 (Invitrogen). To optimize the SHERRY protocol, different amount of random decamer (N10) (0, 10, or 100 pmol) was used to set up reverse transcription reactions. Titration of Tn5 transposome (0.2, 0.5, or 1.0 μl Vazyme V50; 0.05 or 0.25 μl home-made pTXB1) was performed in tagmentation procedure. In all tests, 10 ng 3T3 total RNA was used, and all reagents except for N10 or Tn5 transposome remain unchanged. All libraries were sequenced on Illumina NextSeq 500 with 2x75 paired-end mode. Clean data was aligned to GRCm38 genome and known transcript annotation using Tophat2 v2.1.1. Ribosome-removed aligned reads were proceeded to calculate FPKM by Cufflinks v2.2.1 and gene body coverage by RSeQC v.2.6.4.

Patients and clinical samples

From January 23, 2020 to March 20, 2020, 72 patients were enrolled in this study according to the 7th guideline for the diagnosis and treatment of COVID-19 from the National Health Commission of the People's Republic of China. All patients, diagnosed with COVID-19, were hospitalized in Beijing Ditan Hospital and classified into four severity degrees, mild, moderate, severe, and critical illness, according to the guideline. We collected 85 samples (38 pharyngeal swabs, 34 sputum samples, and 13 stool samples) from these patients.

427

428 **RNA extraction and rRNA removal**

429 For all the clinical samples, nucleic acids extraction was performed in a BSL-3 laboratory.
430 Samples were deactivated by heating at 56°C for 30 min before extraction. Total RNA
431 was extracted using QIAamp Viral RNA Mini Kit (Qiagen) following the manufacturer's
432 instructions. In most samples (79 out of 85) we specifically omitted the use of carrier RNA
433 due to its interference on the most prevalent sample preparation protocols for high-
434 throughput sequencing. After nucleic acids extraction, rRNA was removed by rDNA probe
435 hybridization and RNase H digestion, followed by DNA removal through DNase I digestion,
436 using MGIEasy rRNA removal kit (BGI, Shenzhen, China). The final elution volume was
437 12-20 μl for each sample. For carrier RNA removal tests, 1.7 μg polyA carrier RNA was
438 spiked into 18 μl of elute from QIAamp Viral RNA Mini Kit. To remove the carrier RNA from
439 these spike-in samples and other samples extracted with carrier RNA, 2 μg poly(T) 59-
440 mer (T59) oligo was added during the rDNA hybridization step.

441

442 **dsDL Metagenomic RNA library construction and sequencing**

443 The libraries were constructed using MGIEasy reagents (BGI, China) following
444 manufacture's instruction. The purified RNA, after rRNA depletion and DNA digestion,
445 underwent reverse transcription, second strand synthesis, and sequencing adaptor
446 ligation. After PCR amplification, DNA was denatured and circularized before being
447 sequenced on DNBSEQ-T7 sequencers (BGI, China).

448

449 **MINERVA library preparation**

450 Totally, 2.7 μl RNA from rRNA and DNA removal reaction was used for standard SHERRY
451 reverse transcription, with the following modifications: 1) 10 pmol random decamer (N10)
452 was added to improve coverage; 2) initial concentrations of dNTPs and oligo-dT (T30VN)
453 were increased to 25 mM and 100 μM, respectively. For 5.4 μl and 10.8 μl input, the entire
454 reaction was simply scaled up 2 and 4 folds, respectively. The RNA/DNA hybrid was
455 tagmented in TD reaction buffer (10 mM Tris-Cl pH 7.6, 5 mM $MgCl_2$, 10% DMF)
456 supplemented with 3.4% PEG8000 (VWR Life Science, Cat.No.97061), 1 mM ATP (NEB,
457 Cat.No. P0756), and 1U/μl RNase inhibitor (TaKaRa, Cat.No. 2313B). The reaction was

458 incubated at 55°C for 30 min. 20 µl tagmentation product was mixed with 20.4 µl Q5 High-
459 Fidelity 2X Master Mix (NEB, Cat.No. M0492L), 0.4 µl SuperScript II reverse transcriptase,
460 and incubated at 42°C for 15 min to fill the gaps, followed by 70°C for 15 min to inactivate
461 SuperScript II reverse transcriptase. Then index PCR was performed by adding 4 µl 10
462 µM unique dual index primers and 4 µl Q5 High-Fidelity 2X Master Mix, with the following
463 thermo profile: 98°C 30 s, 18 cycles of [98°C 20 s, 60°C 20 s, 72°C 2 min], 72°C 5 min.
464 The PCR product was then purified with 0.8x VAHTS DNA Clean Beads (Vazyme, Cat.
465 No. N411). These SHERRY libraries were sequenced on Illumina NextSeq 500 with 2x75
466 paired-end mode for metagenomic analysis.
467 For preparing MINERVA libraries through SARS-CoV-2 enrichment, 1µl SHERRY
468 metagenomic library was first quantified with N gene using quantitative PCR (F:
469 GGGGAACTTCTCCTGCTAGAAT, R: CAGACATTTTGCTCTCAAGCTG) after 1:200
470 dilution, then multiple libraries were pooled together based on qPCR results and
471 processed with TargetSeq One Cov Kit (iGeneTech, Cat.No.502002-V1) following
472 manufacturer's instruction. The iGeneTech Blocker was replaced by the IDT xGen
473 Universal Blockers (NXT). These MINERVA libraries were sequenced on Illumina
474 NextSeq 500 with 2x75 paired-end mode for deep SARS-CoV-2 analysis.
475
476 **Data processing**
477 For metagenomic RNA-seq data, raw reads were quality controlled using BBmap (version
478 38.68) and mapped to the human genome reference (GRCh38) using STAR (version
479 2.6.1d) with default parameters. All unmapped reads were collected using samtools
480 (version 1.3) for microbial taxonomy assignment by Centrifuge (version 1.0.4). Custom
481 reference was built from all complete bacterial, viral and any assembled fungal genomes
482 downloaded from NCBI RefSeq database (viral and fungal genomes were downloaded
483 on February 4th, 2020, and bacterial genomes were downloaded on November 14th,
484 2018). There were 11,174 bacterial, 8,997 viral, and 308 fungal genomes respectively.
485 Bacterial Shannon diversity (entropy) was calculated at species level, and the species
486 abundance was measured based on total reads assigned at the specific clade normalized
487 by genome size and sequencing depth. Bacterial genus composition was analyzed based
488 on reads proportion directly assigned by Centrifuge. For dsDL sequencing data, sub-

sampling was performed for each sample to obtain ~12M paired-end nonhuman reads, which is the median of SHERRY datasets. Same workflow was performed as described above for the removal of human reads and microbial taxonomy assignment.

For SARS-CoV-2 genome analysis, raw reads were trimmed to remove sequencing adaptors and low-quality bases with Cutadapt v1.15. BWA 0.7.15-r1140 was used to align reads to the SARS-CoV-2 reference genome (NC_045512.2). Then we removed duplicates from the primary alignment with Picard Tools v2.17.6. We used mpileup function in samtools v1.10 to call SNP and InDel with parameter -C 50 -Q 30 -q 15 -E -d 0. We called mutation if the depth >=10 and strand bias >0.25. The strand bias is defined as the value that minimum of positive strand depth and negative strand depth divided by the maximum.

Data deposition

The datasets generated during this study (related to Figure 2 and 3) are not publicly available due to ethical concerns. Partial access to the datasets is available from the corresponding author on reasonable request.

Figure 1

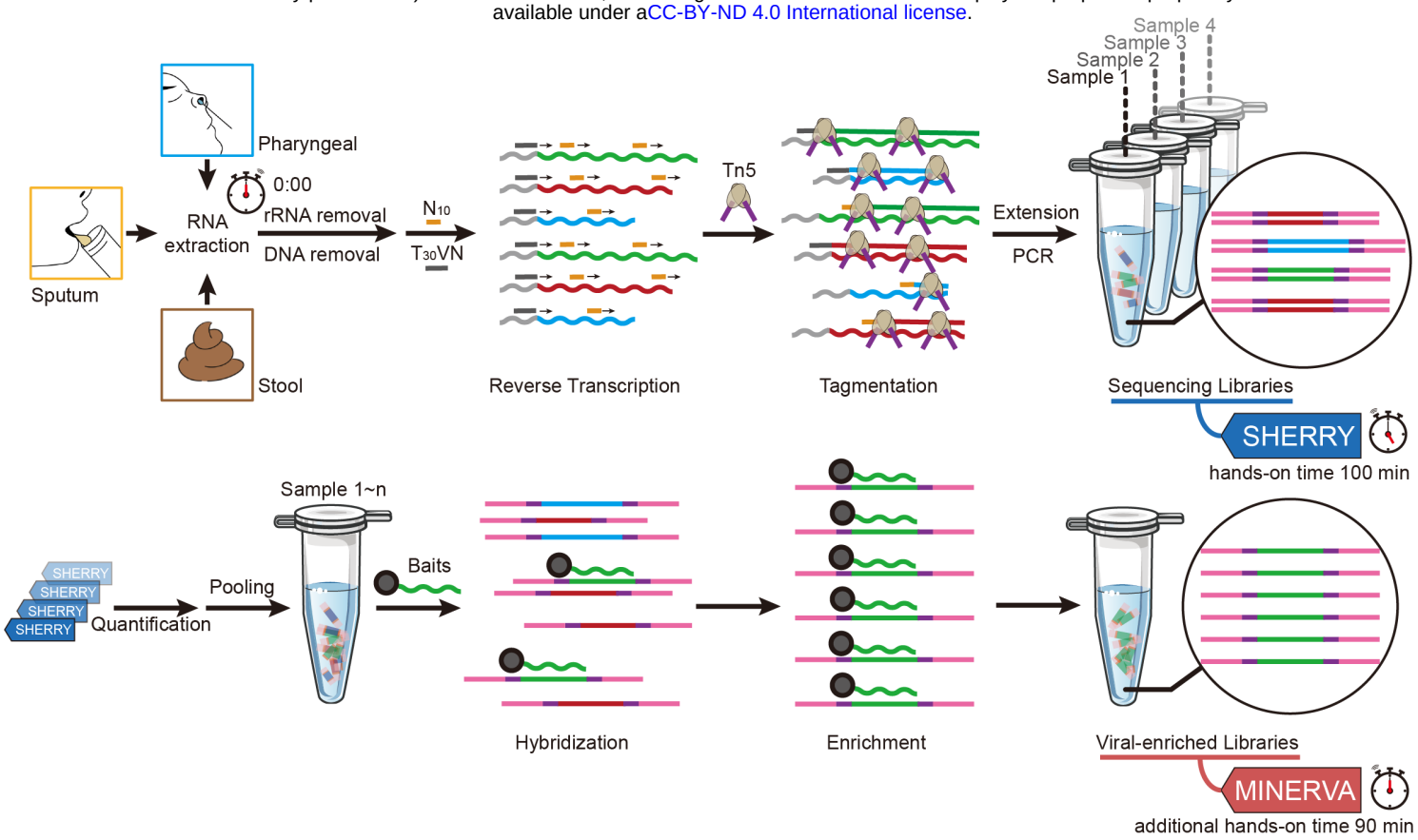


Figure 2

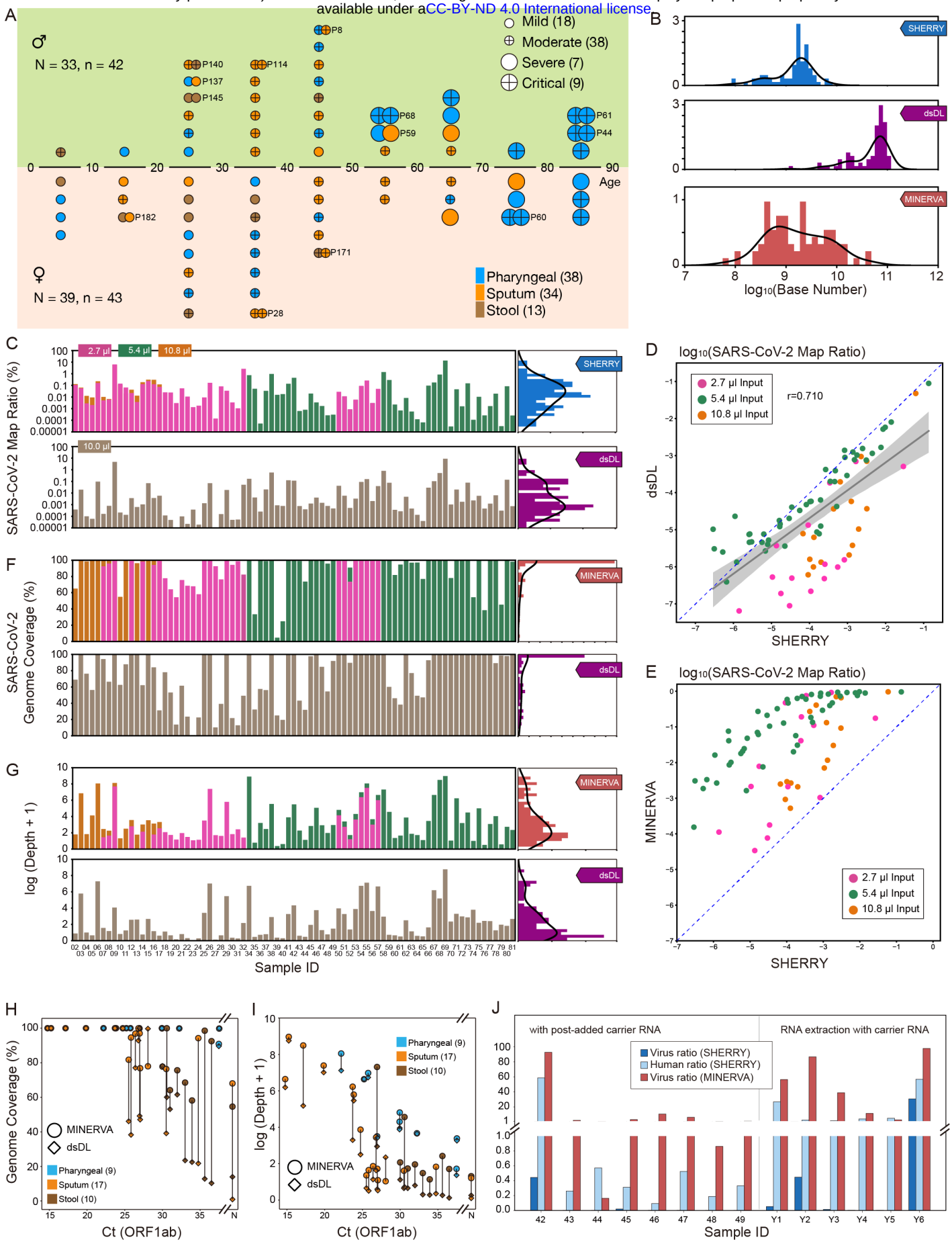
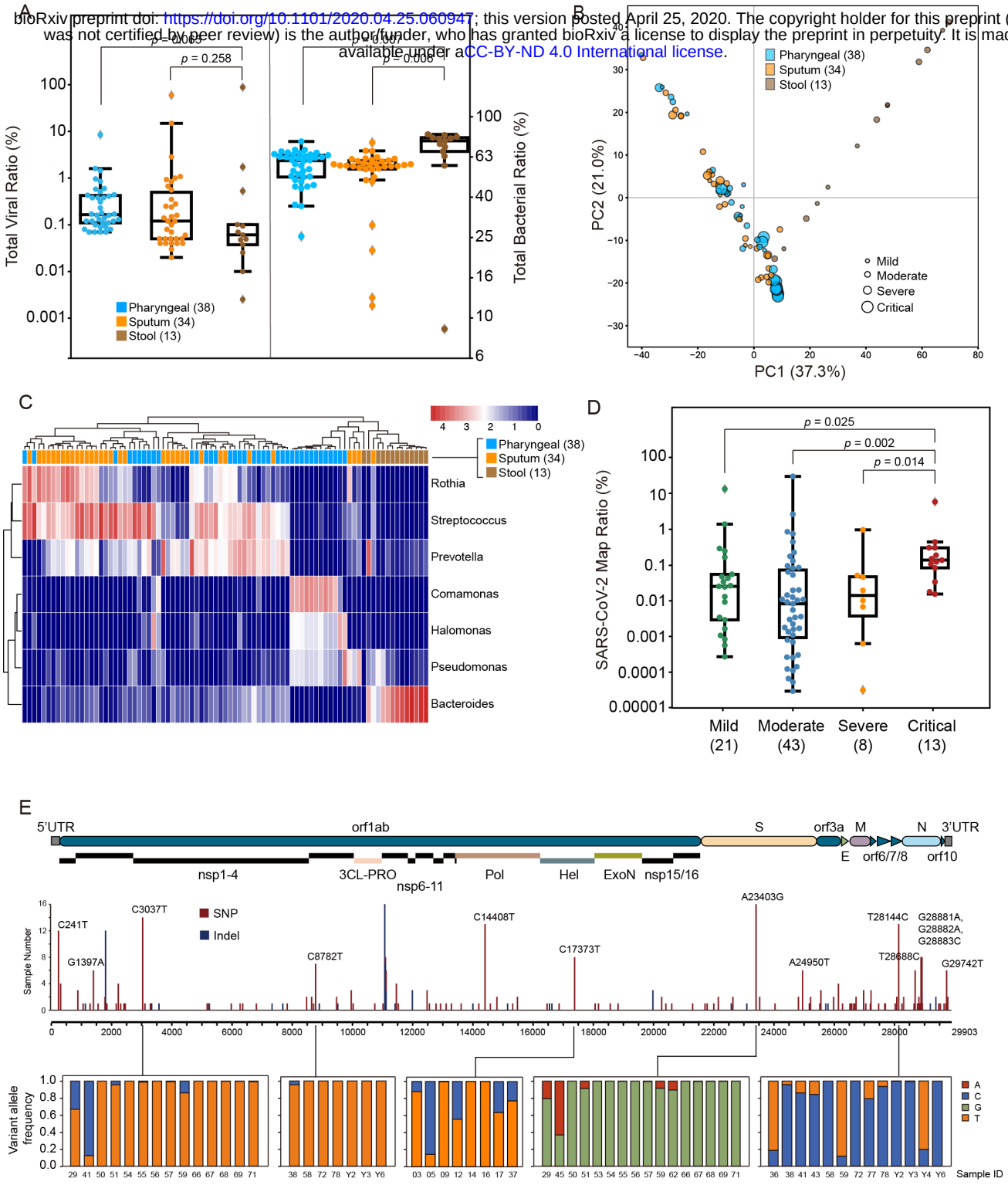
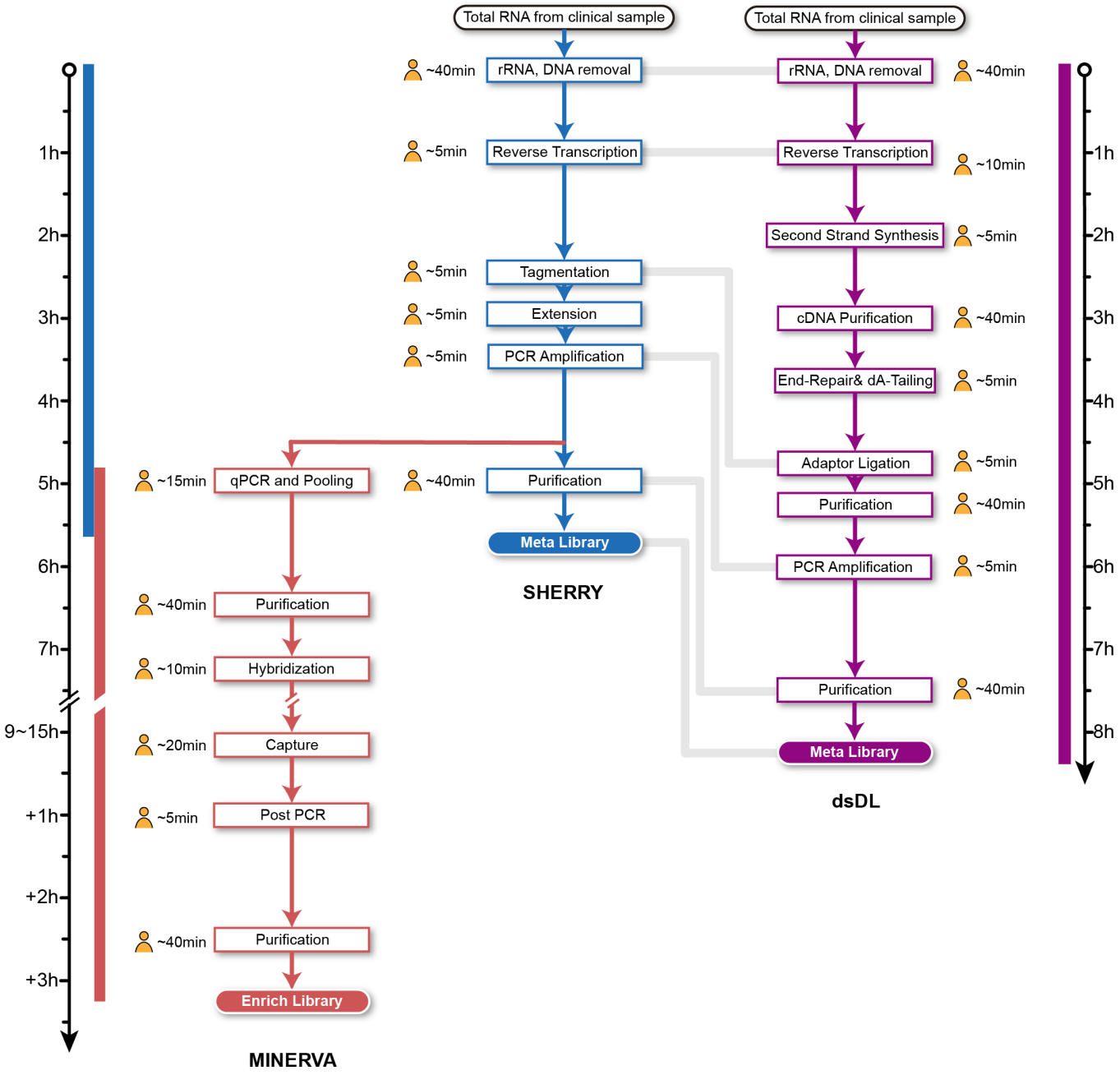


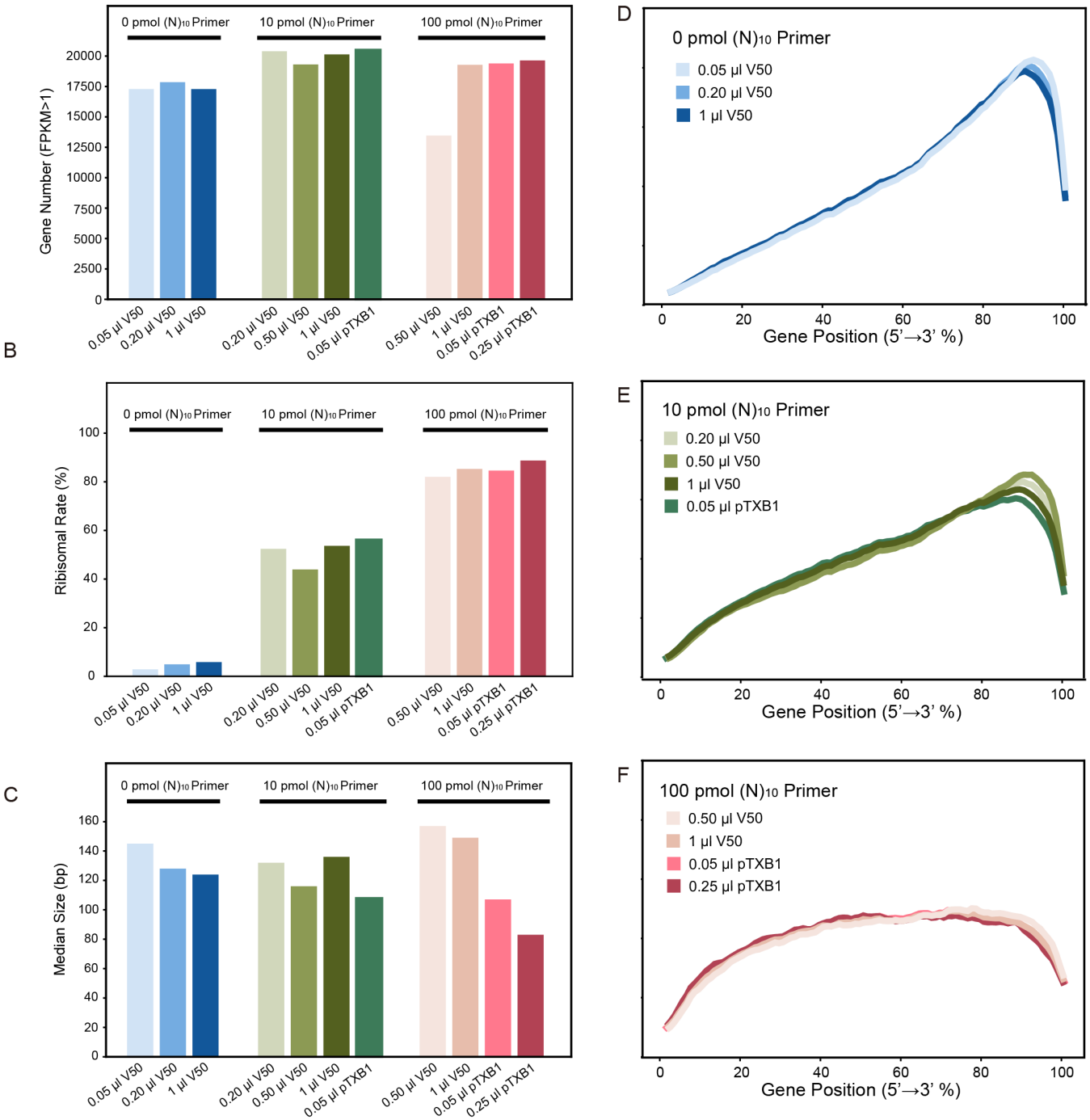
Figure 3



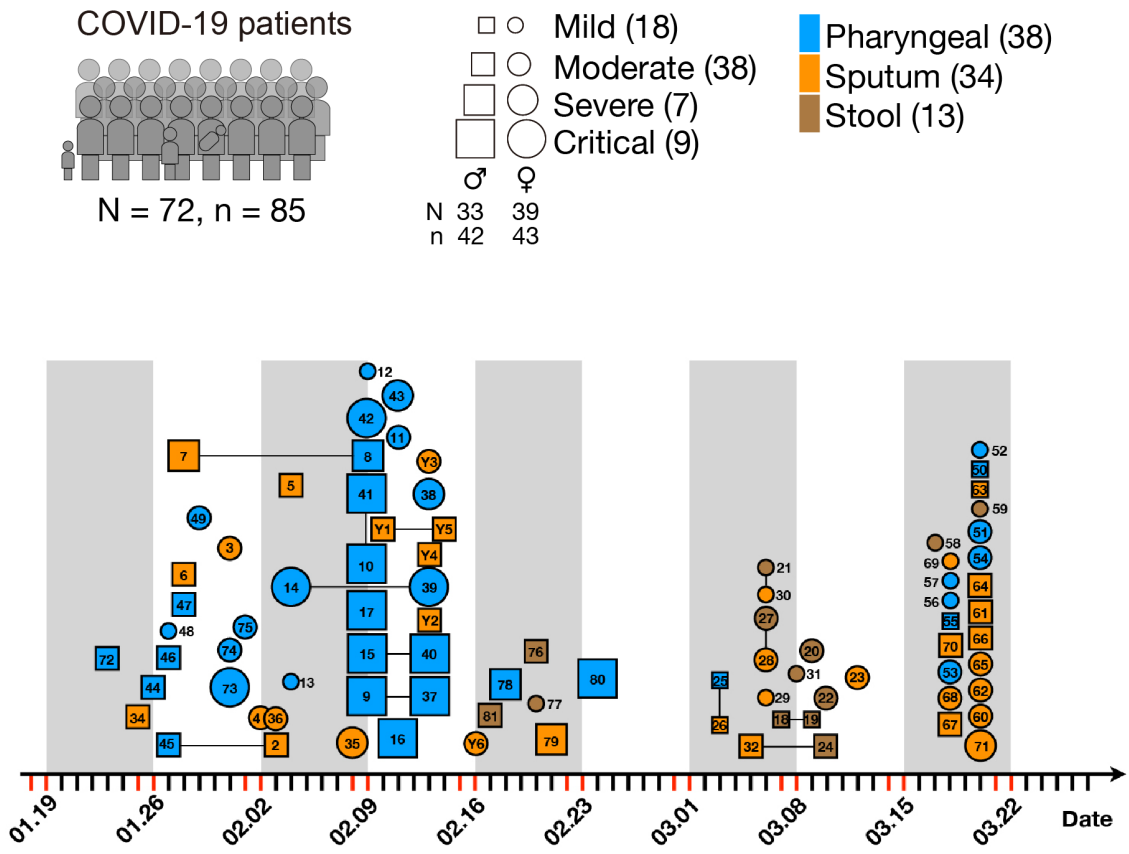
Supplementary Figure 1



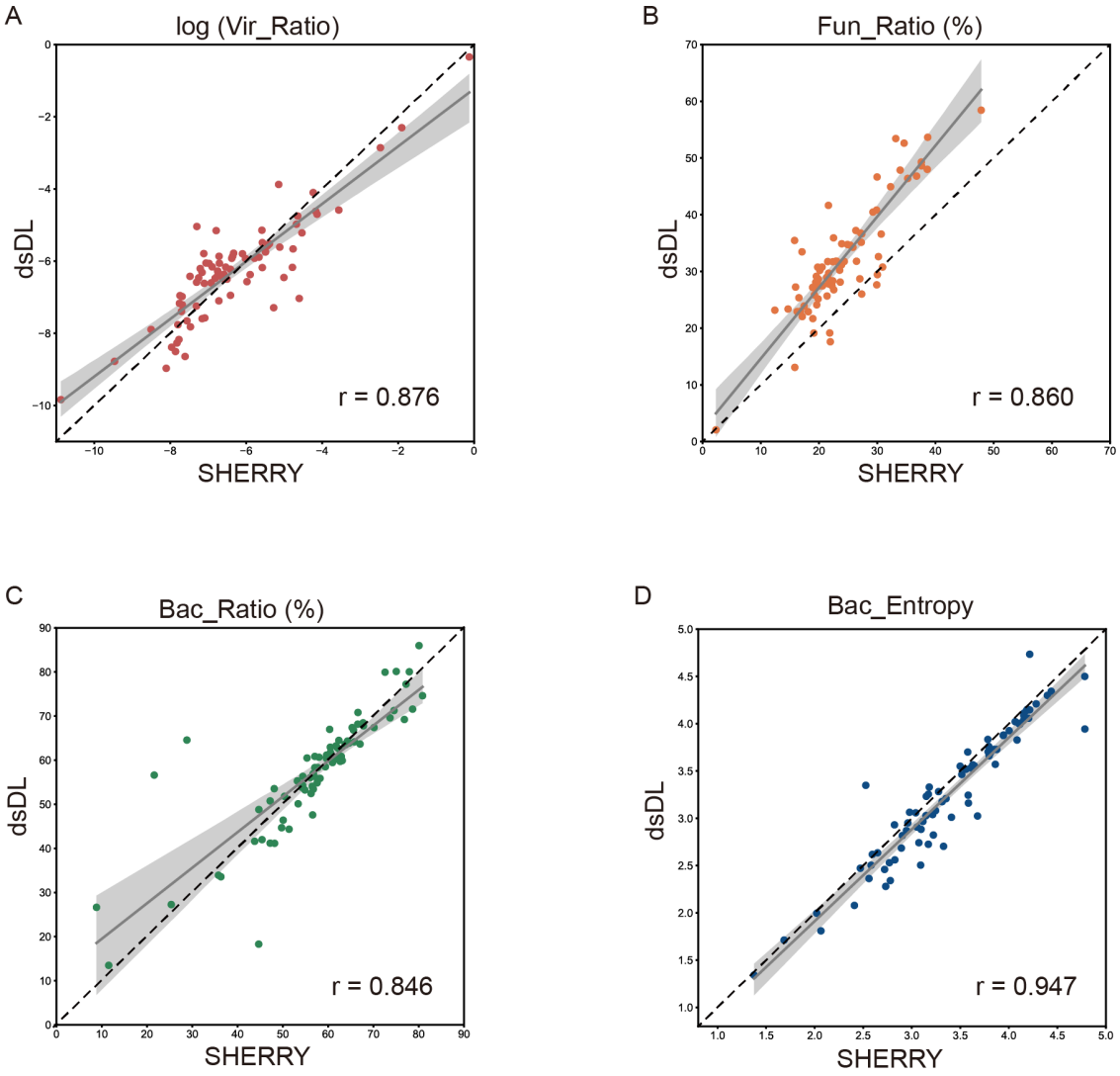
Supplementary Figure 2



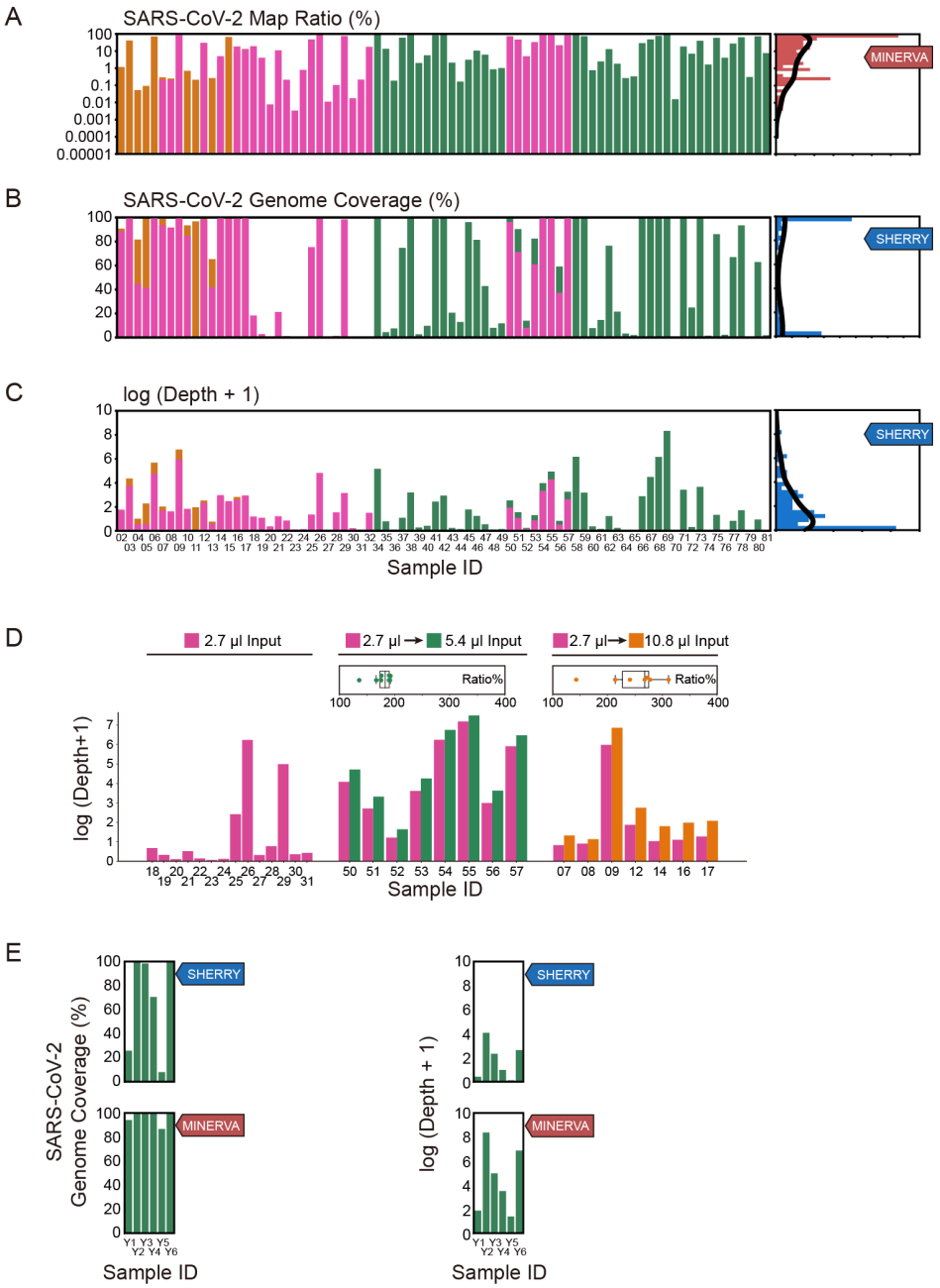
Supplementary Figure 3



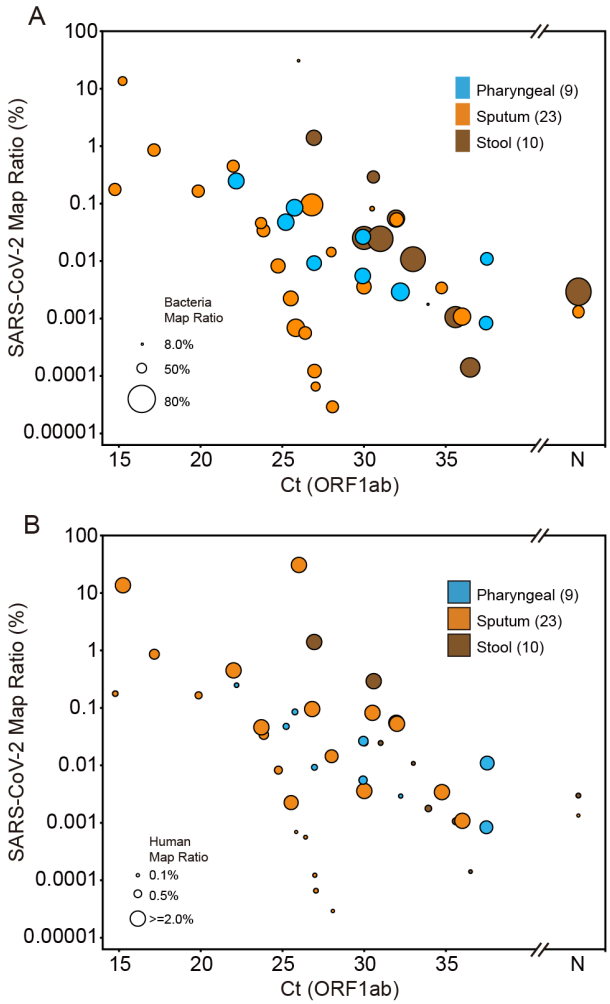
Supplementary Figure 4



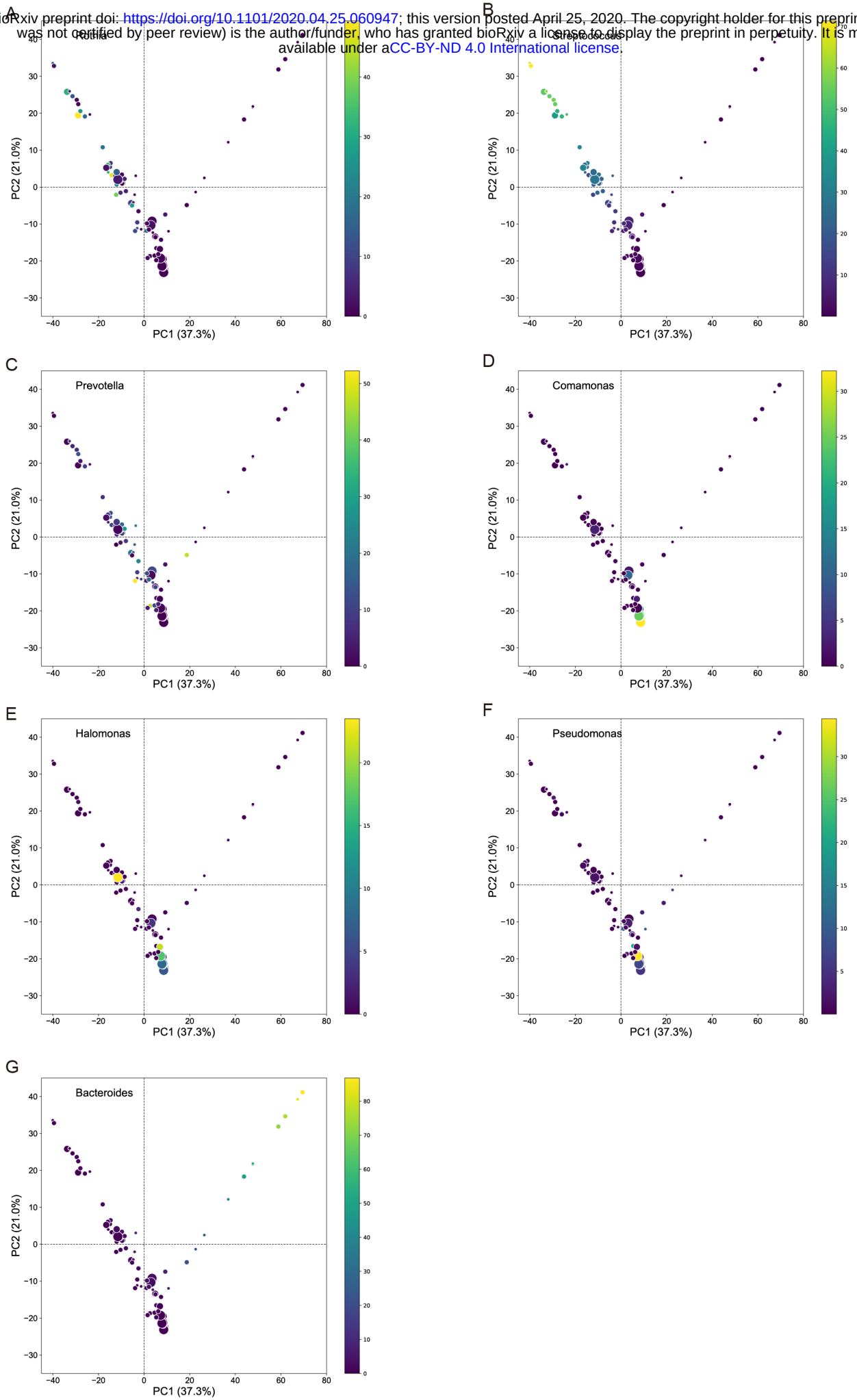
Supplementary Figure 5



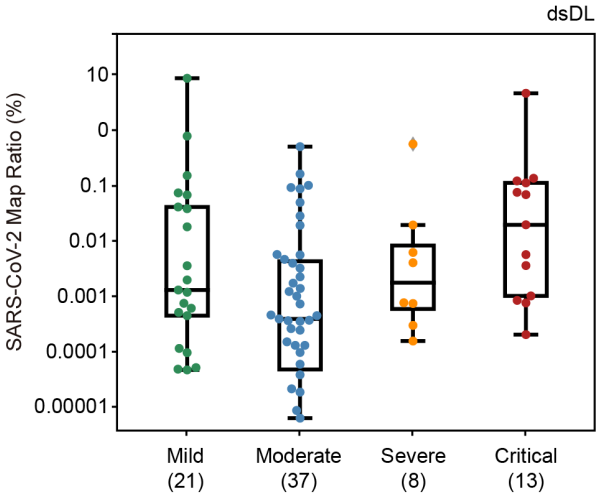
Supplementary Figure 6



Supplementary Figure 7



Supplementary Figure 8



Supplementary Appendix - Protocol

Protocol: MINERVA – a rapid library construction method to sequence SARS-CoV-2 gRNA

Version 1.1

April 23, 2020

Materials

REAGENTS

- RNaseZap (Ambion, Cat. No. AM9780)
- DNA-OFF (Takara Bio, Cat. No. 9036)
- QIAamp Viral RNA Mini Kit (Qiagen, Cat. No. 52906)
- MGIEasy rRNA removal kit (BGI, Cat. No. 1000005953)
- DNase I (RNase-free) (NEB, Cat.No.M0303)
- Tris-HCl (1M, pH 7.6; ROCKLAND, Cat. No. MB-003)
- $MgCl_2$ (1 M; Invitrogen, Cat. No. AM9530G)
- N,N-Dimethylformamide (for molecular biology, ⩾99%; Sigma, Cat. No. D4551)
- DPEC-treated water (Invitrogen, Cat. No. AM9915G)
- Recombinant RNase Inhibitor (40 U/μl; Takara, Cat. No. 2313)
- Deoxynucleotide (dNTP) Solution Set (NEB, Cat. No. N0446S)
- Superscript II reverse transcriptase (Invitrogen, Cat. No. 18064014)
- DTT (0.1M; Invitrogen, Cat. No. 18064014)
- Betaine solution (5 M; Sigma, Cat. No. B0300)
- TruePrep DNA Library Prep Kit V2 for Illumina (Vazyme, Cat. No. TD501)
- PEG8000 (VWR Life Science, Cat.No.97061)
- ATP (10 mM; NEB, Cat. No. P0756)
- Q5 High-Fidelity 2x Master Mix (NEB, Cat. No. M0492)
- ChamQ SYBR qPCR master mix (Vazyme, Cat. No. Q311-02)
- VAHTS DNA Clean Beads (Vazyme, Cat. No. N411)
- Ethanol (200 proof, for molecular biology; Sigma, Cat. No. E7023)
- TargetSeq One Cov Kit (iGeneTech, Cat. No. 502002-V1)
- xGen Universal Blockers (IDT, Cat.No. 1079586)
- All oligos were acquired from Sangon.

SUPPLIES

- Millex-GP Syringe Filter Unit (0.22 μm, polyethersulfone; Millipore, Cat. No. SLGP033RB)
- 0.2 mL Thin Wall PCR Tubes (Axygen, Cat. No. PCR-02-C)

EQUIPMENT

- Thermo cycler
- Magnetic stand
- Vortexer
- Real-Time PCR machine
- A compatible Illumina DNA sequencing instrument

REAGENT SETUP

- **SARS-Cov-2 infected samples**

SARS-Cov-2 infected samples, including pharyngeal swabs, sputum samples or stool samples are collected following clinical guidelines. All of the samples, or their viral transfer media, must be deactivated at 56 °C for 30 min before nucleic acid extraction.

IMPORTANT NOTE: All the procedure should be operated in a BSL-3 laboratory before the total RNA is ready.

- **Total RNA**

  Total RNA can be extracted from SARS-Cov-2 infected samples by QIAamp Viral RNA Mini Kit, while omitting the addition of carrier RNA if possible. Then use MGIEasy rRNA removal kit to remove the ribosomal RNA, and DNase I to remove the DNA. The final elution volume is 12-20 μl for each sample.

  IMPORTANT NOTE: All the procedure should be operated in a BSL-3 laboratory before the total RNA is ready.

- **Oligo dT primer**

  Oligo dT primer (5'-T_{30}VN-3') anneals to all the RNAs containing a poly(A) tail. 'N' is any base and 'V' is either A, C or G. Dissolve the oligonucleotide in DPEC-treated water, to a final concentration of 100 μM. The oligo solution can be stored at −20 °C for at least 6 months.

- **Random decamer primer**

  Random decamer primer (5'-N_{10}-3') anneals to the RNA randomly. 'N' represents any base. Dissolve the oligonucleotide in DPEC-treated water, to a final concentration of 100 μM. The oligo solution can be stored at −20 °C for at least 6 months.

- **dNTP mix**

  Combine equal volume of dATP, dTTP, dCTP, dGTP in Deoxynucleotide (dNTP) Solution Set.

- **5X TD Buffer**

  50 mM Tris-HCl, 25 mM $MgCl_2$, 50% N,N-Dimethylformamide (DMF). This buffer can be stored at 4℃ for at least 6 months.

- **N-ch-F primer**

  N-ch-F primer (5'-GGGGAACTTCTCCTGCTAGAAT-3') is the forward primer used to amplify the N region of SARS-Cov-2 in qPCR. Dissolve the oligonucleotide in DPEC-treated water, to a final concentration of 100 μM. The oligo solution can be stored at −20 °C for at least 6 months.

- **N-ch-R primer**

  N-ch-R primer (5'-CAGACATTTTGCTCTCAAGCTG-3') is the reverse primer that used to amplify the N region of SARS-Cov-2 in qPCR. Dissolve the oligonucleotide in DPEC-treated water, to a final concentration of 100 μM. The oligo solution can be stored at −20 °C for at least 6 months.

Procedure

1) Clean the work space, including the hood and pipettes, with DNA-off and RNase-Zap. Filter the self-made buffer with 0.22 μm filter. Use a thermal cycler with a heated lid set to 105 °C for all incubations throughout this protocol. All reaction mixes should be set up on ice.

IMPORTANT NOTE: All the procedure should be operated in a BSL-3 laboratory before the total RNA is ready.

Reverse transcription

2) Prepare preRT mix on ice by adding 0.2 µl RNase Inhibitor (40U/µl), 0.2 µl Oligo dT (100 µM), 2 µl $(N)_{10}$ random primer (100 µM), 0.8 µl dNTP mix (25 mM) to 5.4 µl total RNA.

NOTE: If the volume of the RNA sample is more or less than 5.4 µl, please scale up or down the whole reaction volume proportionally.

3) Mix the reaction gently and thoroughly without bubbles and incubate it at 72°C for 3 min, then quickly cool it on ice.

4) Prepare RT mix by combining the reagents in the table below.

Component	Volume (µl)	Final concentration
SuperScript II reverse transcriptase (200U/µl)	1.00	200 U
RNase Inhibitor (40U/µl)	0.50	20 U
SuperScript II first strand buffer (5x)	4.00	1x
DTT (0.1M)	1.00	5 mM
Betaine (5M)	4.00	1 M
$MgCl_2$ (1M)	0.12	6 mM
DPEC water	0.78	-
Total Volume	11.40	-

5) Add 11.40 µl RT mix to the samples from step 3 and gently pipette without bubbles. Incubate the reaction at 42°C for 1.5 hour for reverse transcription, and followed by 70°C 15min to inactivate SuperScript II reverse transcriptase. Then put the sample on ice.

Tagmentation

6) Dissolve PEG8000 powder to DPEC-treated water with concentration of 40% (w/w), and filter the solution by 0.22 µm filter.

7) Prepare the tagmentation mix by combining the reagents listed below.

Component	Volume (µl)	Final concentration
V50 (TruePrep DNA Library Prep Kit V2)	1.00	-
5xTD buffer	8.00	1x
RNase Inhibitor (40U/µl)	1.00	40 U
40% PEG8000	3.40	3.4%
ATP (10 mM)	4.00	1.00 mM
DPEC water	2.60	-
Total Volume	20.00	-

8) Add the tagmentation mix to the sample from step 5. Pipette the reaction gently and thoroughly. Incubate the reaction at 55°C for 30 min, then cool it on ice.

Amplification and Sample Pooling

9) Add 0.8 µl SuperScript II reverse transcriptase and 40.8 µl Q5 High-Fidelity 2x Master Mix to the tagmentation products.

10) Mix the reaction well without forming bubbles. Incubate the reaction at 42°C for 15 min to fill the 9 bp gap left by Tn5 transposome, followed by 70°C 15 min to inactivate SuperScript II reverse transcriptase.

11) Prepare the PCR mix by combining 4 µl N6xx index primer (10 µM), 4 µl N8xx index primer (10 µM) and 8 µl Q5 High-Fidelity 2x Master Mix. The final concentration of each index primer is around 0.4 µM.

12) Add PCR mix to the sample from step 10 and pipette thoroughly. Perform PCR as detailed below.

Cycle	Denature	Anneal	Extension	Hold
1	98°C, 30 s	-	-	
2-19	98°C, 20 s	60°C, 20 s	72°C, 2 min	
20	-	-	72°C, 5 min	
21	-	-	-	4°C

13) For targeted SARS-CoV-2 deep sequencing, continue with **qPCR and Sample Pooling.** For metagenomic sequencing, pool 8-16 libraries in equal volumes and go to **Step 18**.

qPCR and Sample Pooling

14) Take 1 µl PCR product and dilute it 200-fold with nuclease-free water.

15) Prepare qPCR mix by adding 0.05 µl N-ch-F primer (100 µM), 0.05 µl N-ch-R primer (100 µM), 5 µl ChamQ SYBR qPCR master mix and 3.9 µl nuclease-free water to 1 µl diluted template.

16) The program of qPCR is performed at 95°C for 60s, followed by 40 cycles of [95°C 5s, 60°C 15s].

17) Pool 8-16 libraries together according to their Ct values as following:

Ct	Volume taken (µl)
>28	50.0
24-28	16.0
20-24	3.0
<20	0.5

Purification

18) Place VATHS DNA clean beads at room temperature for 15 min, then vortex violently.

19) Add equal volume of beads (1x) to the pooled samples and vortex violently. Incubate the mixture at room temperature for 5 min.

20) Transfer the tube to compatible magnetic stand until the solution is clear.

21) Carefully remove the solution without disturbing beads.

22) Wash beads with 200 µl 80% ethanol (freshly prepared) and incubate for 30 s, then remove the ethanol. Repeat this step one more time.

23) Dry the beads on magnetic stand with cap open until the color of beads gets light. Add 52

μl nuclease-free water and close the cap, vortex violently to wash DNA off.

24) Incubate the tube at room temperature for 5 min off the magnet stand.

25) Quickly spin down the tube then place it on magnetic stand until the solution is clear.

26) Carefully aspirate 50 μl supernatant to a clean tube without disturbing beads. The library can be restored at -20°C for 6 months.

IMPORTANT NOTE: The equal-volume mixed libraries are now ready for sequencing.

SARS-Cov-2 Sequence Enrichment

27) Subject the purified Ct-adjusted library pool to one round of SARS-Cov-2 sequence capture following the instruction of TargetSeq One Cov Kit. Replace the iGeneTech Blocker with the IDT xGen Universal Blockers.

High-Throughput Sequencing

28) Both the metagenomic library and the SARS-Cov-2 targeted library can be sequenced on any Illumina platform with paired-end mode. 10-20 million reads are typically collected.