CovProfile: profiling the viral genome and gene expressions of SARS-COV2

Yonghan Yu[1,†], Zhengtu Li[2,†], Yinhu Li[1,†], Le Yu[3,†], Wenlong Jia[1], Feng Ye[2,*] Shuai Cheng Li[1,*]

1 City University of Hong Kong Shenzhen Research Institute
2 State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou 510120, China
3 Beijing YuanShengKangTai (ProtoDNA) Genetech Co. Ltd., Beijing 100190, China
† The authors contributed equally to the work.

***yefeng@gird.cn, *shuaicli@gmail.com**

Abstract

SARS-COV2 is infecting more than one million people. Knowing its genome and gene expressions is essential to understand the virus. Here, we propose a computational tool to detect the viral genomic variations as well as viral gene expressions from the sequenced amplicons with Nanopore devices of the multiplex reverse transcription polymerase chain reactions. We applied our tool to 11 samples of terminally ill patients, and discovered that seven of the samples are infected by two viral strains. The expression of viral genes ORF1ab gene, S gene, M gene, and N gene are high among most of the samples. While performing our tests, we noticed the abundances of MUC5B transcript segments to be consistently higher than those of other genes in the host.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) has infected more than one million people, and has led to more than 80,000 deaths, at the time of preparation of this manuscript [1, 23, 25, 28]. The virus is an enveloped and single-stranded RNA betacoronavirus 30kb base-pairs (bps), , which belongs to the family Coronaviridae [6, 27]. Since the year 2000, we have witnessed and experienced three highly widespread pathogenic coronaviruses in human populations — the other two viruses are SARS-COV in 2002-2003, and MERS-CoV in 2012 [4]. All three viruses can lead to Acute Respiratory Distress Syndrome (ARDS) in the human host, which may cause pulmonary fibrosis,and lead to permanent lung function reduction or death [2, 20, 22]. The mechanisms of these severe cases remain unclear.

No drug is currently confirmed to be effective against SARS-COV-2. Some vaccine candidates are under intensive studies [12]. A nucleotide analog RNA-dependent RNA Polymerase (RdRP) inhibitor called remdesivir has demonstrated effects on SARS-COV-2 [19]. Hoffmann *et al.* suggested that TMPRSS2 inhibitor approved for clinical use can be adopted as a treatment option [9]. *In silico* methods have as well produced lists of potential drugs [7] and epitopes [16] worthy of further exploration. However, short of details on how the viral molecule compromises the cells of hosts, it is hard to prioritize our exploration of the huge array of potential drugs and therapies. Analysis of the mechanism can provide us insights on the viral transmission as well as revealing therapeutic targets.

However, there are many difficulties in analyzing the virus. Isolating the virus can be tedious and time-consuming. The samples usually carry low viral load, resulting in false negatives. To reduce the

false negatives, one approach is to amplify the viral sequences with multiplex reverse transcription polymerase chain reaction (RT-PCR), followed by sequencing the amplicons with third generation sequencing technologies such as the Nanopore devices. The approach has been applied to sequence the dengue [21] and Zika viruses [17], as well as in the SARS-COV-2 viruses [14].

Nanopore sequencing generate long reads, which enable the detection of large genomic structural changes, viral recombination and viral integration into host genome. They also supports alternative splicing analyses of microorganism studies [3]. However, The long reads suffers high error rates. Detection of the viral and host gene expressions using amplicons is further complicated by several factors. First, the sub-sequences can be amplified unevenly. Second, reads in the gene region amplified from the viral genome are indistinguishable from the gene transcripts. Third, chimeric reads are frequent.

Here, we developed a computational tool, CovProfile, to detect viral genomic variations, and profile viral gene expressions through data from multiplex RT-PCR Amplicon Sequencing on Nanopore GridION. To detect the genomic variations, we mapped the reads to SARS-CoV-2 reference genome (MN908947.3), and to profile the viral gene expressions, we adopt a regularization method to find out the primer efficiencies, We have tested the tool on the data obtained from the sputum samples in eleven severe COVID-19 patients. We found seven out of eleven patients are infected by two strains, and discovered no structural variations. We found that the expression of viral genes ORF1ab gene, S gene, M gene, and N gene are high among most of the samples. While performing our tests, we noticed the abundances of MUC5B transcript segments to be consistently higher than those of other genes in host. The function of this gene may be associated with the severe symptom of pulmonary fibrosis.

Method and Materials

Data

In this study, we collected Nanopore sequencing data for SARS-CoV-2 virus from the sputum samples of 11 COVID-19 patients. A total of 1,804,043 clean reads were generated, with an average of 164,004.00±90,848.28 (Mean±SD) reads per sample (Supplementary Figure 1). Aligning the reads to SARS-CoV-2 and human genomes respectively, the alignment ratio ranged from 20.14% to 99.74% on SARS-CoV-2 genome, and ranged from 0.11% to 72.12% on the human genome. Furthermore, each sample contained 71.24±35.36M data on average, while 61.29±42.13M and 8.87±9.34M data were aligned to SARS-CoV-2 and host genomes, respectively.

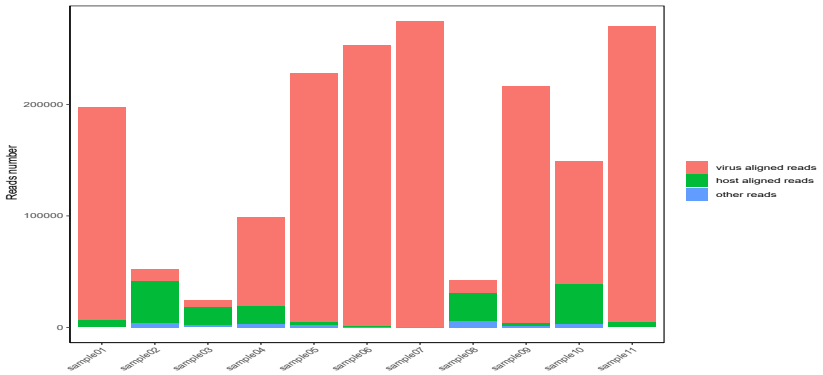

Figure 1. Compositions of the data sets. The heights of the columns represented the composition of dsta. The data came from the host, SARS-CoV-2 and other genomes respectively, which were represented by blue, orange and grey colors.

Remove false-positive chimeric reads

Chimeric nanopore reads formed by random connection of multiplex RT-PCR amplicons were split into multiple segments if the location of their bilateral ends matches to corresponding locations of PCR primer pair. These chimeric reads could be split into several segments with their bilateral ends exactly matching the sites where the multiplex PCR primers are located, suggesting that they should be processed accurately to avoid false positive identification of virus recombinations or host integrations. We position primers on both viral and host genome, and splits the identified chimeric reads into segments corresponding to PCR amplicons. This method rescues lots of sequencing data, providing more accurate alignment and higher coverage. Realignment were performed on the separated segments before analysis.

Mutation Detection

We aligned the clean data to SARS-CoV-2 genome (NCBI MN908947.3) with minimap2 [11] with default parameters for Oxford Nanopore reads. With the alignment results, the genomic variations with average quality larger than 10 and coverage depth larger than 20 were called with bcftools (version 1.8). To reduce the effects brought about by the PCR amplification, a variation is filtered if it was located within 10 bp upstream or downstream of the primer region, and occurred only in one sample.

Estimating viral load and primer efficiency

Assume that there are n primer pairs (s_i, t_i), with primer efficiency e_i, $1 \leq i \leq n$. Let the viral gene boundaries be denoted (l_g, r_g), where g indexes the genes. Denote the observed sequence depth captured by the primer pair (s_i, t_i) as d_i.

We want to estimate the viral gene expression. Let the number of viral copies be denoted v, the number of transcripts of gene g be denoted t_g, and the sequencing rate of amplified segment be denoted α. Then, we can have the following two approximations for d_i.

If the primer i is fully covered by gene g (Fig1. A), then as the number of amplicons increases exponentially with the number of replication cycles we have

$$d_i \approx \frac{1}{\alpha}(v + t_g)e_g^k. \tag{1}$$

where k is the number of replication cycles.

Similarly, if primer pair i spans across gene boundaries; that is, it belongs to none of the gene transcripts (Fig1. B), then we have

$$d_i \approx \frac{1}{\alpha}ve_g^k. \tag{2}$$

If some reads share two ends, where one end is a primer and the other end is at some gene boundary (Fig 1D, Fig 1E), it may indicate that these amplicons are the result of linear amplification. Denote the number of reads $r_{i,g}$ with two endpoints as (t_i, e_g), and the number of reads $r_{g,i}$ with two endpoints as (l_i, s_g), then

$$r_g \approx \frac{1}{2\alpha}kt_ge_g \tag{3}$$

If a gene g is within some primer target region (Fig2. G), then the sequenced fragment can be expected to exactly cover the gene. Let the number of observed reads span exactly the whole gene g as r_g, then

$$r_g \approx \frac{1}{\alpha}t_g. \tag{4}$$

To find the expressions of the viral genes, we want to minimize the following objective function based on the entities defined above

$$\sum(\log(v + t_i) + k\log e_i - \log\alpha - \log d_i)^2 +$$

$$\sum(\log v + k \log e_i - \log \alpha - \log d_i)^2 +$$
$$\sum(\log k + \log t_i + \log \frac{e_i}{2} - \log r_i \log \alpha)^2 +$$
$$\sum \lambda_1 \log^2 e_i + \lambda_2 \log^2 v,$$

where λ_1 and λ_2 are two regularization parameters, and k is a weight factor (we set k to 21 according to our experiments). The viral load v, primer efficiencies e_i, and expressions t_g which minimize the objective equation are our estimation for these parameters.

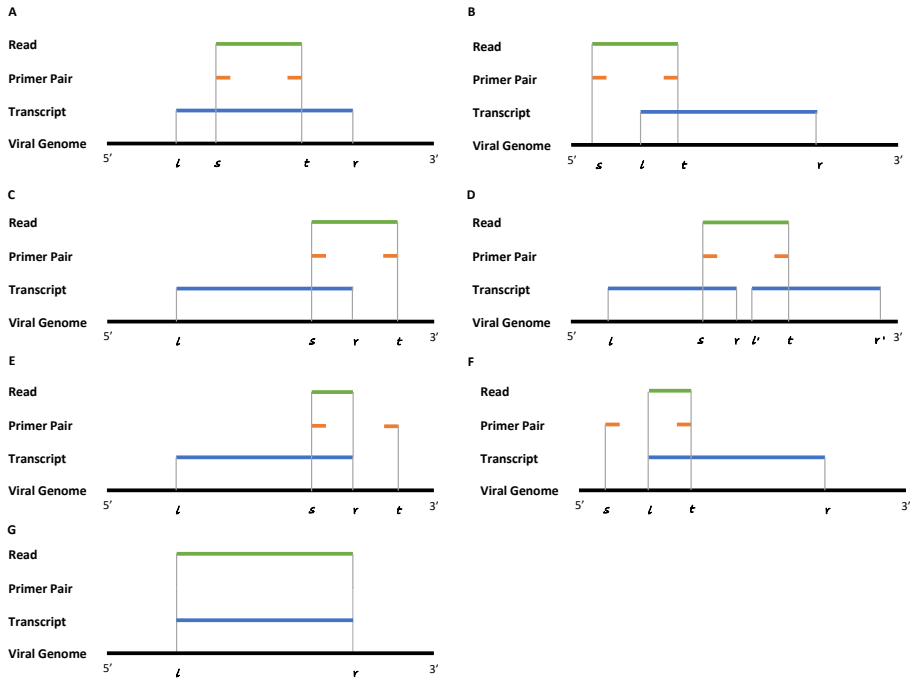

Figure 2. Relations between viral gene, primer pair and sequence. **A**: Primer pair spans fully on gene with sequence aligned with primers. **B**: Primer pair spans across gene boundaries at 5' with sequence aligned with primers. **C**: Primer pair spans across gene boundaries at 3' with sequence aligned with primers. **D**: Primer pair spans across two gene with sequence aligned with primers. **E**: Primer pair spans across gene boundaries at 5' with sequence aligned with gene boundaries at 3' and primer at 5'. **F**: Primer pair spans across gene boundaries at 3' with sequence aligned with gene boundaries at 5' and primer at 3'. **G**: Sequence aligned with gene boundaries at both 5' and 3'.

Calculate viral gene expression

To calculate the viral gene expression, the duplication generated during PCR were removed based on the estimated primer efficiencies. The gene expression can be calculated as the sum of the base depth of each gene divided by its length.

Capture gene product from host

To capture the gene products of the host, reads and primers were aligned to the hg38 transcript database derived from UCSC refGene annotation. Primer alignment were performed with BLAT [10]. Regions where the depth of reads alignment are equal or larger to 5 are screened. Reads with endpoints matching the screened primer alignment were treated as PCR duplications. Due to mismatches in primer to

transcript alignment, primer efficiencies may change when capturing gene product. Thus, the recognized PCR duplication were removed and one copy of those reads were kept. The gene expression from host can then be estimated as the sum of the base depth of each transcript divided by its length.

Results

Distribution of SNPs in SARS-CoV-2 genome

After aligning the sequencing data to the SARS-CoV-2 genome, we detected the mutations of SARS-CoV-2 in the 11 samples3. Based on the mutations of SARS-CoV-2 in 11 samples, we discovered 38 SNPs (Single Nucleotide Polymorphisms) in total; 35 of them are located on the genetic regions, including gene ORF1ab, S, N and M. Gene ORF1ab contained 23 SNPs in 11 samples, and 17 of them were non-synonymous mutations. Non-synonymous mutations also occupied a large part of SNPs in the gene S, N and M: 5 of 8 SNPs in gene S, 1 of 3 SNPs in gene N, and 1 of 1 SNPs in gene M were non-synonymous mutations. Since most of the SNPs were located in only 1 sample, we deduced that the functions of the discovered SNPs need to be further verified according to lab experiments. Surprisingly, we also found that 7 among 11 samples were heterogeneous at loci 8782. This may imply that the hosts have two different virus strains. However, we did not find creditable InDels (Insertions and Deletions), or structural variations, viral-host integration during the mutation analysis.

Evaluations on the amplification efficient for the primers

Since 98 pairs of primers from two pools were designed for the enrichment of SARS-CoV-2 genome before sequencing, we discovered that different parts of SARS-CoV-2 genome exhibited different coverage depths. We then calculated the amplification efficiencies for the primers with our proposed method. We discovered that the primers located at the 5' and 3' ends of the SARS-CoV-2 genome exhibited higher amplification efficiency (>1.2, Figure 4), including primer01 and primer84-primer93. With the genomic locations of the primers, we know that the genomic sequences within the ranges 24-410 and 25,280-28,464 can be amplified with more sequencing data. On the other hand, the remaining primers exhibited slightly lower efficiencies, within the range of 1.0-1.2. For primers 64, 70, 83 and 85, their amplification efficiencies were as low as 1, which indicates that the target regions of these primers were not amplified. Based on the primer efficiencies, the effect of data bias which was induced by the PCR experiments can be attenuated from the original data, and they can be further applied for the normalization of viral gene expression under different primers.

Normalized coverage depth on SARS-CoV-2 genome

Due to the amplification discrepancy among the primers, the sequencing data was unevenly distributed along the SARS-CoV-2 genome. After aligning the sequencing data to the SARS-CoV-2 genome, we found that the coverage depths of some genomic regions reached over 8000x in some samples, but some regions have coverage depths less than 10x (Figure 5). In order to minimize the data imbalance induced by PCR amplification, we performed data normalization based on the primer efficiency and their locations on SARS-CoV-2 genome, so that the normalized data could reflect the actual virus reproduction in the host. Based on the algorithm illustrated in the method, we obtained the normalized sequencing depth of 11 sample data.

After data normalization, the overall coverage depth along the SARS-CoV-2 genome was relatively even, and less than 300x. In addition, we discovered that the ORF1ab gene, S gene, M gene, and N gene exhibited higher coverage depth in the samples, which might indicate higher expression (Figure 5). These genes are related to host cell recognition, amplification and virus particle assembly, so we speculate that the results might explain the severe SARS-CoV-2 virus infection in the patients. Since gene E, ORF6, ORF7a, ORF7b, ORF8 and ORF10 are less than 400bp in length, their transcripts might not have been captured by the primers, and hence their expressions cannot be estimated.

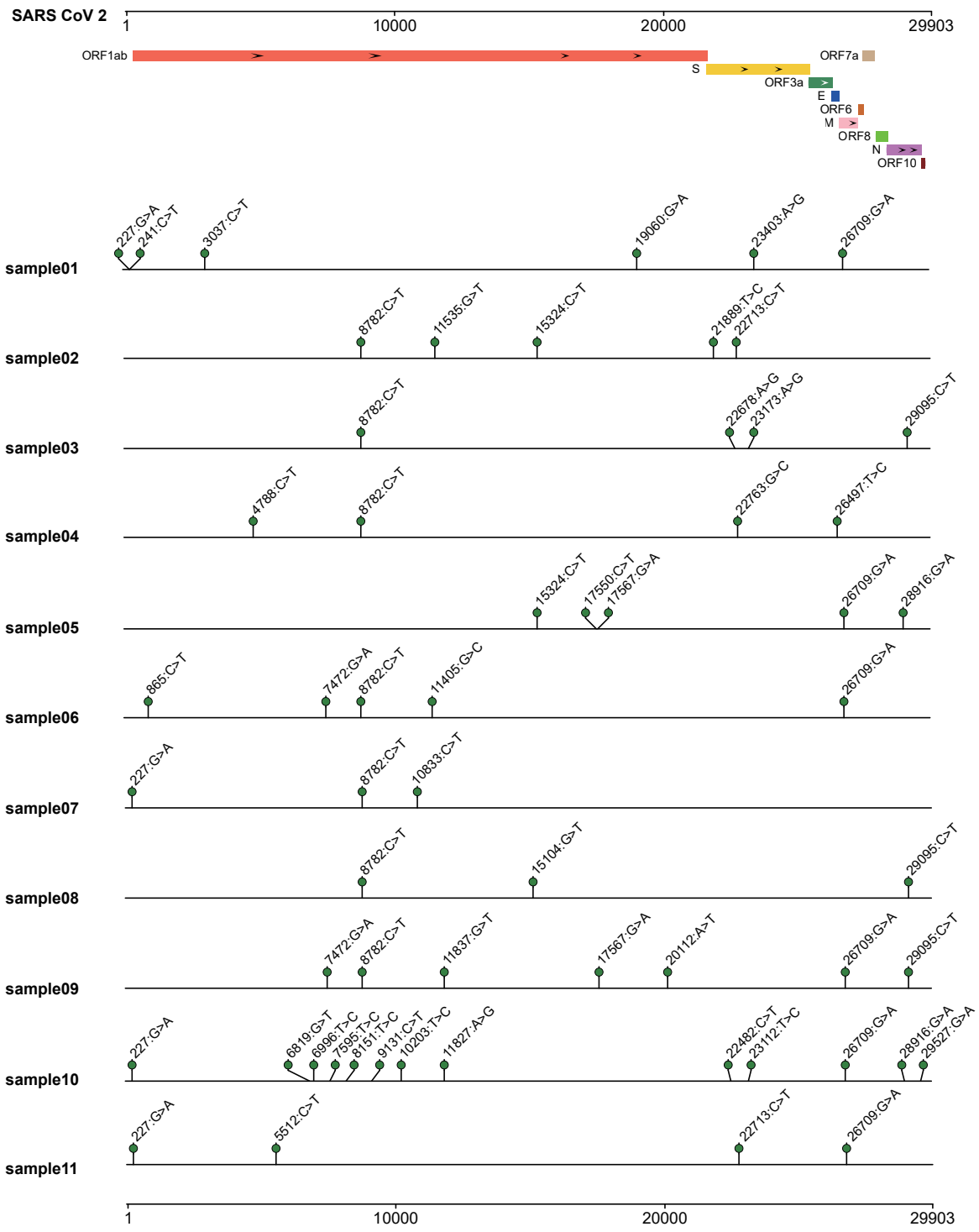

Figure 3. Distributions of High-frequency SNPs among the 11 COVID-19 samples. The locations of genes in SARS-CoV-2 were exhibited on top of the plot, and the distributions of SNPs were specified for the 11 samples.

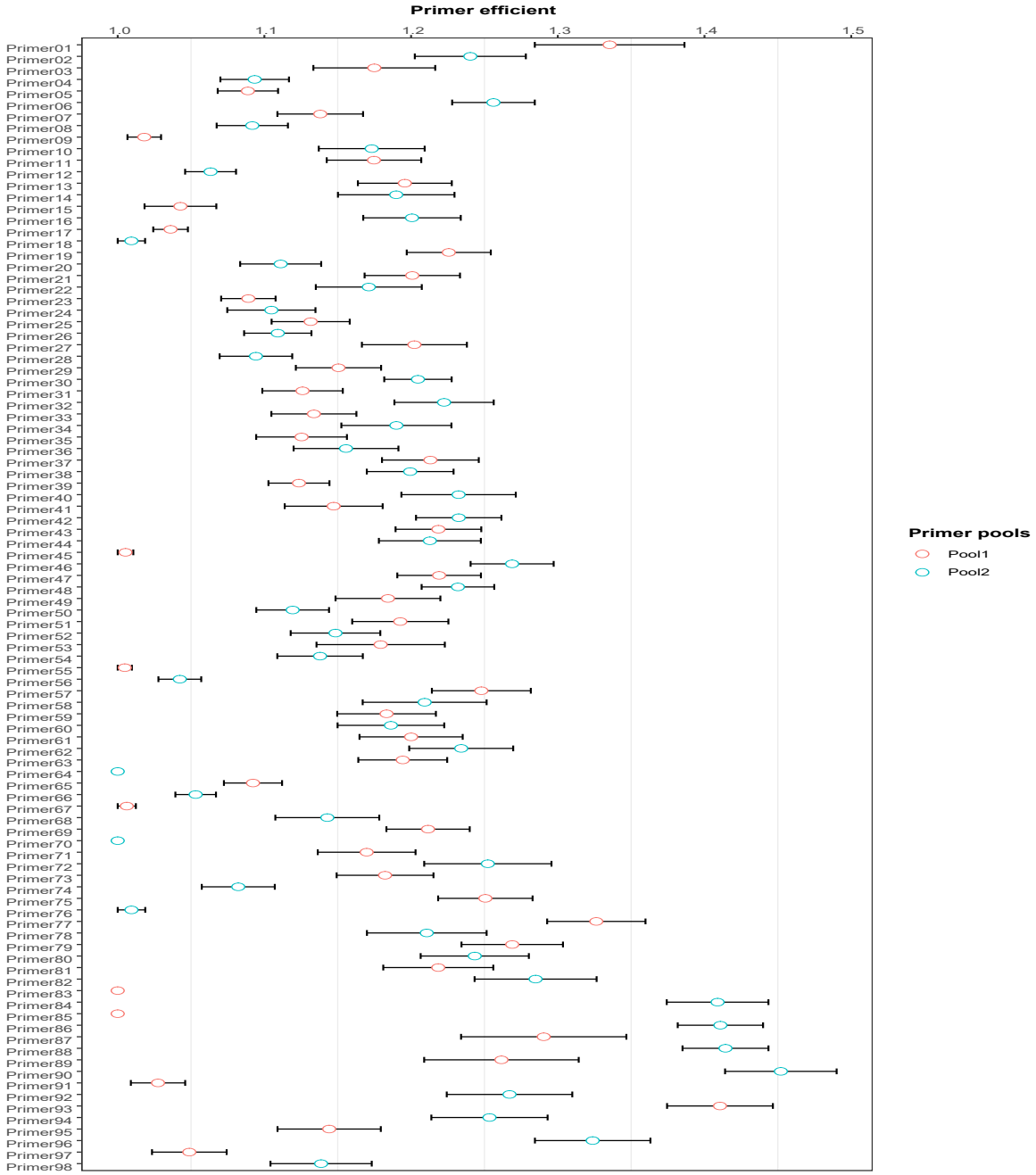

Figure 4. Amplification efficient for the 98 pairs of primers. The amplification efficient was calculated for the primers, and their distributions were detected in 11 samples. The primers in pool1 and pool2 were marked with red and blue colors respectively.

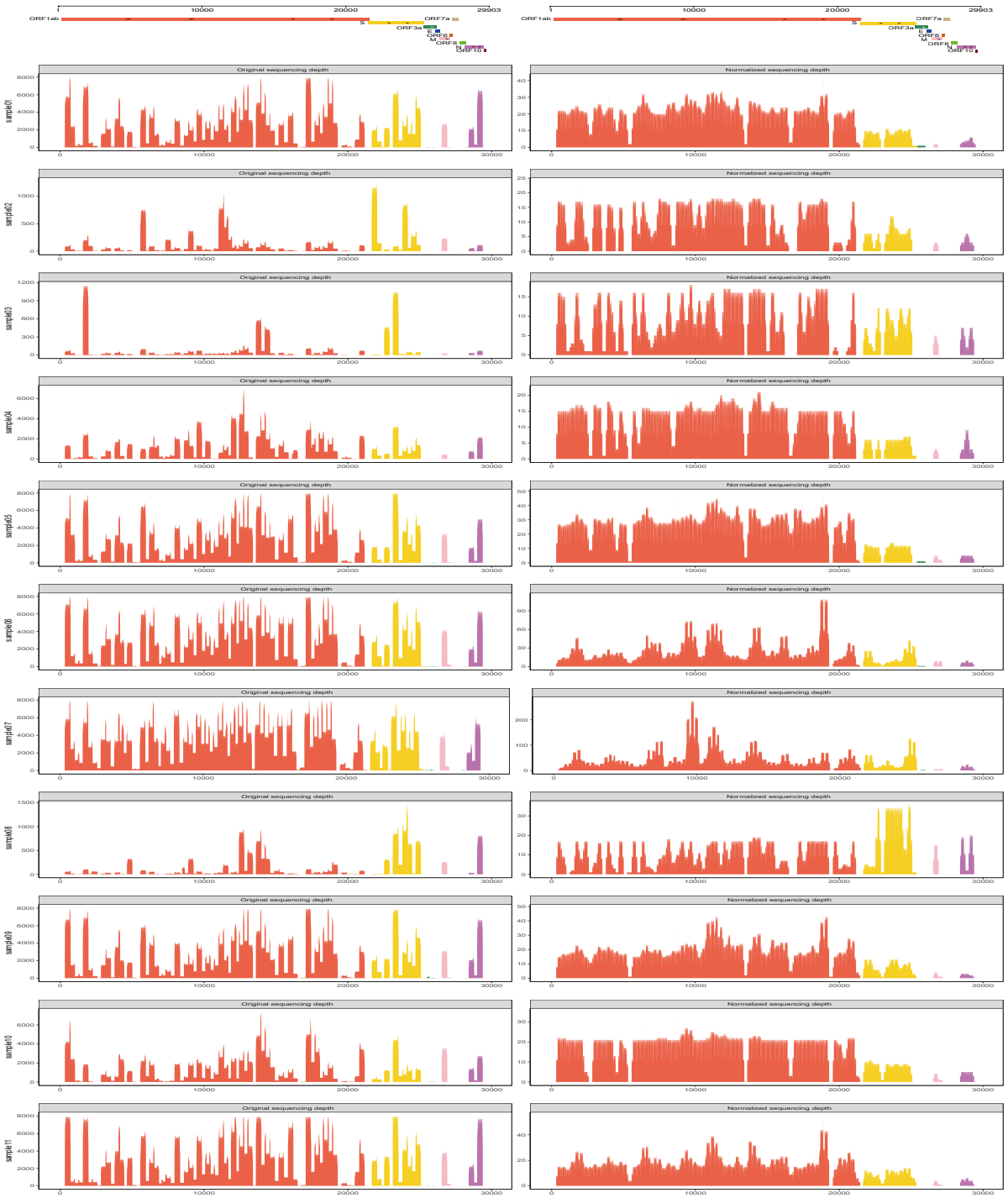

Figure 5. Original and normalized coverage depth for the 11 samples. The locations of genes in SARS-CoV-2 were exhibited on top of the plot, and they were marked by different colors. Each sample contained two plots, which represented the original (left) and normalized coverage (right) depths for the bases along the SARS-CoV-2 genome respectively. The coverage depth were colored by the genes in SARS-CoV-2.

High expression genes in host

Previous studies on the sequencing data have shown that it contains the human genome, so we detected the corresponding genes and their expression levels. Aligning these data to human transcript database, we identified many segments of human gene transcripts. Subsequently, the coverage depths of these genes were normalized based on the primer efficiencies. Only several genes are listed within the top 5% in more than five samples according to the normalized depth. They include the genes MUC5B, TCAF12, KRT12 and *etc*. MUC5B is the only gene consistently within the top 5% across all the samples. More than 50% of the MUC5B gene region is covered (Figure 5). However, the region with high coverage in MUC5B gene is inconsistent among the samples, perhaps due to data scarcity.

Conclusion and Discussion

In this study, we implemented a software tool, CovProfile, to analyze multiplex RT-PCRs on Nanopore sequencing data, and studied the sputum samples of COVID-19 patients. Our results show enrichment of PCR amplicons which totally cover 99.7% of SARS-CoV-2 virus genome, revealing that the multiplex RT-PCRs are a reliable method in identification of COVID-19 infection. Interestingly, the variant (26709:G>A) is identified in half of samples, and shows consistent heterozygosity (alteration frequency: 19.36%-26.15%) in all relevant samples. This suggests at least two virus strains in the patients.

The data contained large amount of chimeric reads formed by unintended random connection of multiplex PCR amplicons, making up 1.69% of total sequencing reads. Similar observation has been reported that these chimeric reads can be formed during library preparation and sequencing, and lead to the cross barcode assignment errors in multiplex samples [24, 26]. We applied CovProfile on the data and recovered majority of chimeric reads.

The sequencing reads could be used to evaluate the expression of SARS-CoV-2 genes. However, the coverage unevenness caused by efficiency differences in the multiplex primers impedes our calculation of pathogenic genes expression. To solve this problem, we propose a multi-parameter method which considers the primer efficiency, gene differential expressions, and RNA virus genome copy number simultaneously. Different locations on the viral genome have distinct composition of molecule sources, including RNA genome, expressed viral transcript, and PCR amplicons. According to the gene locations along the viral genome, we enumerate reads types based on aligned positions, and deduce the contributions of different sources to the local region depth. By applying this method on our data, we find that viral gene ORF1ab may maintain higher expression than other genes across most samples. This viral gene occupied a large region on SARS-CoV-2 genome, and it responsible for the amplification and assembly of the virus [15]. For the primer, the multiplex primers efficiency has wide distribution in single sample, and the distributions among samples demonstrate similar trend. This suggests that our estimation of primers efficiency comply to the reality.

We utilize the sequencing data to evaluate host genome expression and find several genes consistently activated in the majority of samples after eliminating the PCR effects. The most abundant gene is the MUC5B, which encodes a member of the mucin protein family. The family proteins are the major components of mucus secretions in normal lung mucus [18]. MUC5B has been reported to be highly expressed in lung disease, such as chronic obstructive pulmonary disease (COPD). Its up-regulated transcription could cause mucociliary dysfunction and enhances lung fibrosis [8, 13], thus might be relevant to the recent report that a lot of mucus are found in the lung of COVID-19 infected patients. Another activated host gene is the TCAF2, which encodes transient receptor potential cation channel-associated factor. It could bind to the TRPM8 channel and promote its trafficking to the cell surface. Previous study reported that TCAF2 significantly increases the migration speed of prostate cancer cells when it co-expresses with TRPM8 [5]. Other enriched genes might also break the cellular homeostasis, including the nucleoredoxin-like gene NXNL2, tyrosine kinases FER, and so on. Further studies of these genes could help researchers gain insights into the dysregulation of host cells after the SARS-CoV-2 infection. We should point out these genes are extracted based on a very limited amount of data, and the conclusions can be different with a large data sets.

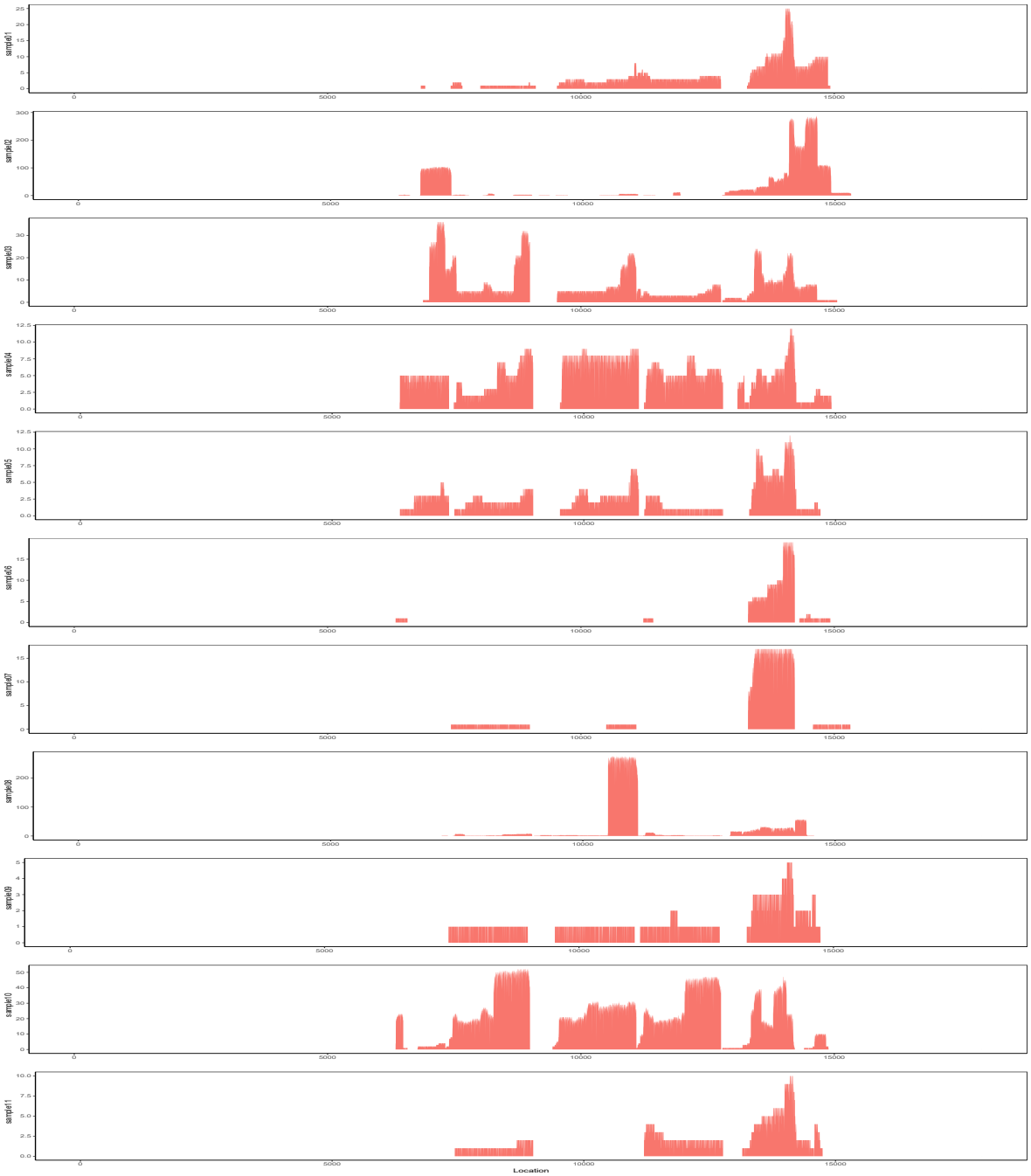

Figure 6. Normalized coverage depth on MUC5B gene in the 11 samples. In each plot, the X and Y coordinates exhibited the length of MUC5B and the normalized coverage depth on along the MUC5B gene respectively.

Overall, we present a tool for genomic variation detection and the transcriptional profile of SARS-CoV-2 infection from multiplex Nanopore sequencing data. The tool decomposes and realign the chimeric reads that commonly exist in the multiplex Nanopore sequencing, and this will greatly promote the data usage. The method also provides a robust estimation of primer efficiency through a multi-parameter model. The viral and host gene expression we report could pave steps to the development of diagnosis and therapy strategies of SARS-CoV-2 infection candidates.

Acknowledgments

We thank just about everybody.

Funding

Not applicable.

Author's contributions

SCL conceived the idea.
SCL and FY supervised the project.
SCL and YY designed the algorithms
SCL, WJ, YY and HL discussed and designed the experiments.
YY and JL implemented the code and conducted the analysis.
ZL and LY prepared the data.
SCL YY WJ and YL drafted the manuscript.
SCL, FY revised the manuscript.
All authors read and approved the final manuscript.

References

1. Novel coronavirus (2019-ncov) situation reports, https://www.who.int/emergencies/diseases/novelcoronavirus-2019/situation-reports, 2020.

2. A. Assiri, G. R. Abedi, M. A. Masri, A. B. Saeed, S. I. Gerber, and J. T. Watson. Middle east respiratory syndrome coronavirus infection during pregnancy: A report of 5 cases from saudi arabia: Table 1. *Clinical Infectious Diseases*, 63(7):951–953, June 2016.

3. M. T. Bolisetty, G. Rajadinakaran, and B. R. Graveley. Determining exon connectivity in complex mrnas by nanopore sequencing. *Genome Biology*, 16(1):204, 2015.

4. E. de Wit, N. van Doremalen, D. Falzarano, and V. J. Munster. Sars and mers: recent insights into emerging coronaviruses. *Nature Reviews Microbiology*, 14(8):523, 2016.

5. D. Gkika, L. Lemonnier, G. Shapovalov, D. Gordienko, C. Poux, M. Bernardini, A. Bokhobza, G. Bidaux, C. Degerny, K. Verreman, et al. Trp channel–associated factors are a novel protein family that regulates trpm8 trafficking and activityidentification of trpm8 partner proteins. *The Journal of Cell Biology*, 208(1):89–107, 2015.

6. A. Gorbalenya, S. Baker, R. Baric, et al. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-ncov and naming it sars-cov-2. *Nature Microbiology*, 2020.

7. P. H. Guzzi, D. Mercatelli, C. Ceraolo, and F. M. Giorgi. Master regulator analysis of the sars-cov-2/human interactome. *bioRxiv*, 2020.

8. L. A. Hancock, C. E. Hennessy, G. M. Solomon, E. Dobrinskikh, A. Estrella, N. Hara, D. B. Hill, W. J. Kissner, M. R. Markovetz, D. E. G. Villalon, et al. Muc5b overexpression causes mucociliary dysfunction and enhances lung fibrosis in mice. *Nature communications*, 9(1):1–10, 2018.

9. M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, et al. Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell*, 2020.

10. W. J. Kent. BLAT—the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, Mar. 2002.

11. H. Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 05 2018.

12. C. Liu, Q. Zhou, Y. Li, L. V. Garner, S. P. Watkins, L. J. Carter, J. Smoot, A. C. Gregg, A. D. Daniels, S. Jervey, et al. Research and development on therapeutic agents and vaccines for covid-19 and related human coronavirus diseases, 2020.

13. S. K. Mathai, S. Humphries, J. A. Kropski, T. S. Blackwell, J. Powers, A. D. Walts, C. Markin, J. Woodward, J. H. Chung, K. K. Brown, et al. Muc5b variant is associated with visually and quantitatively detected preclinical pulmonary fibrosis. *Thorax*, 74(12):1131–1139, 2019.

14. S. C. Moore, R. Penrice-Randal, M. Alruwaili, X. Dong, S. T. Pullan, D. Carter, K. Bewley, Q. Zhao, Y. Sun, C. Hartley, et al. Amplicon based minion sequencing of sars-cov-2 and metagenomic characterisation of nasopharyngeal swabs from patients with covid-19. *medRxiv*, 2020.

15. T. Phan. Genetic diversity and evolution of SARS-CoV-2. *Infection, Genetics and Evolution*, 81:104260, July 2020.

16. M. Prachar, S. Justesen, D. B. Steen-Jensen, S. P. Thorgrimsen, E. Jurgons, O. Winther, and F. O. Bagger. Covid-19 vaccine candidates: Prediction and validation of 174 sars-cov-2 epitopes. *bioRxiv*, 2020.

17. J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, et al. Multiplex pcr method for minion and illumina sequencing of zika and other virus genomes directly from clinical samples. *nature protocols*, 12(6):1261, 2017.

18. M. G. Roy, A. Livraghi-Butrico, A. A. Fletcher, M. M. McElwee, S. E. Evans, R. M. Boerner, S. N. Alexander, L. K. Bellinghausen, A. S. Song, Y. M. Petrova, et al. Muc5b is required for airway defence. *Nature*, 505(7483):412–416, 2014.

19. T. P. Sheahan, A. C. Sims, S. R. Leist, A. Schäfer, J. Won, A. J. Brown, S. A. Montgomery, A. Hogg, D. Babusis, M. O. Clarke, et al. Comparative therapeutic efficacy of remdesivir and combination lopinavir, ritonavir, and interferon beta against mers-cov. *Nature Communications*, 11(1):1–14, 2020.

20. H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, Y. Fan, and C. Zheng. Radiological findings from 81 patients with COVID-19 pneumonia in wuhan, china: a descriptive study. *The Lancet Infectious Diseases*, 20(4):425–434, Apr. 2020.

21. S. C. Stubbs, B. A. Blacklaws, B. Yohan, F. A. Yudhaputri, R. F. Hayati, B. Schwem, E. M. Salvaña, R. V. Destura, J. S. Lester, K. S. Myint, et al. Assessment of a multiplex pcr and nanopore-based method for dengue virus sequencing in indonesia. *Virology Journal*, 17(1):1–13, 2020.

22. G. M.-K. Tse. Pulmonary pathological features in coronavirus associated severe acute respiratory syndrome (SARS). *Journal of Clinical Pathology*, 57(3):260–265, Mar. 2004.

23. C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao. A novel coronavirus outbreak of global health concern. *The Lancet*, 395(10223):470–473, 2020.

24. R. White, C. Pellefigues, F. Ronchese, O. Lamiable, and D. Eccles. Investigation of chimeric reads using the minion. *F1000Research*, 6, 2017.

25. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, et al. A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798):265–269, 2020.

26. Y. Xu, K. Lewandowski, S. Lumley, S. Pullan, R. Vipond, M. Carroll, D. Foster, P. C. Matthews, T. Peto, and D. Crook. Detection of viral pathogens with multiplex nanopore minion sequencing: be careful with cross-talk. *Frontiers in microbiology*, 9:2225, 2018.

27. P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, pages 1–4, 2020.

28. N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, et al. China novel coronavirus investigating and research team. a novel coronavirus from patients with pneumonia in china, 2019. *N Engl J Med*, 382(8):727–733, 2020.