bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CovProfile: profiling the viral genome and gene expressions of SARS-COV-2

Yonghan Yu[1,†], Zhengtu Li[2,†], Yinhu Li[1,†], Le Yu[3,†], Wenlong Jia[1], Yiqi Jiang[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.

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

Abstract

The SARS-CoV-2 virus has infected more than one million people worldwide to date. Knowing its genome and gene expressions is essential to understand the virus' mechanism. Here, we propose a computational tool CovProfile to detect the viral genomic variations as well as viral gene expressions from the sequences obtained from Nanopore devices. We applied CovProfile to 11 samples, each from a terminally ill patient, and discovered that all the patients are infected by multiple viral strains, which might affect the reliability of phylogenetic analysis. Moreover, the expression of viral genes ORF1ab gene, S gene, M gene, and N gene are high among most of the samples. While performing the tests, we noticed a consistent abundance of transcript segments of MUC5B, presumably from the host, across all the samples.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has infected more than 1.2 million persons, and has led to more than 68,000 deaths, at the time of preparation of this manuscript [1–4]. The virus is an enveloped and single-stranded RNA betacoronavirus of ∼30k base-pairs (bps), which belongs to the family Coronaviridae [5, 6]. Since the year 2000, we have witnessed and experienced three highly widespread pathogenic coronaviruses in human populations — the other two are SARS-COV in 2002-2003, and MERS-CoV in 2012 [7]. 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 [8–10]. The mechanisms of the severe cases due to SARS-CoV-2 remain unclear.

No drug is currently confirmed to be effective against SARS-CoV-2. Some vaccine candidates are under intensive studies [11]. A nucleotide analog RNA-dependent RNA Polymerase (RdRP) inhibitor called remdesivir has demonstrated effects on SARS-CoV-2 [12]. Hoffmann *et al.* suggested that TMPRSS2 inhibitor approved for clinical use can be adopted as a treatment option [13]. *In silico* methods have as well produced lists of potential drugs [14] and epitopes [15] worthy of further exploration. However, short of details on how the viral molecule compromises the host's cells, 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 reveal therapeutic targets.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

However, there are many pitfalls in sequencing the SARS-CoV-2 virus. Since the samples typically carry very low viral load, they are necessarily amplified through multiplex reverse transcription polymerase chain reaction (RT-PCR) before sequencing. Nanopore sequencing is typically used, in order to generate long reads that allow the detection of large genomic structural changes, viral recombination, and viral integration into the host genome [16]. They also support alternative splicing analyses of microorganisms. The approach has been used to sequence the dengue [17] and Zika viruses [18], as well as the SARS-CoV-2 [19].

This approach, however, has several problems. First, the long reads from Nanopore technology suffers from high error rates. Second, the sub-sequences can be amplified unevenly. Third, reads in the gene region amplified from the viral genome are indistinguishable from the gene transcripts. Finally, chimeric reads are frequent. These inaccuracies severely impact the study of viral expressions. Furthermore, while not inherent in the technology, we noticed that the problem could be further complicated by the existence of multiple viral strains.

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 MinION. To detect the genomic variations, we mapped the reads to SARS-CoV-2 reference genome (MN908947.3). To profile the viral gene expressions, we adopt a regularization method to find out the primer efficiencies. The tool was tested on the data obtained from the sputum samples in eleven severe COVID-19 patients. No structural variation was found, but all eleven patients appear to be infected by two strains. 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 the tests, we noticed that MUC5B transcript segments to be consistently more abundant than segments of other genes in the host. The function of this gene may be associated with the severe symptom of abundant mucus in COVID-19 patients.

The discovery of multiple strains in all our samples has severe implications on the study of phylogeny, which rely heavily on SNPs (Single Nucleotide Polymorphisms). We identify at least two problems. First, coexistence of multiple strains might lead to inference of the wrong allele combinations, resulting in incorrect phylogeny and interpretation. Second, SNV calling might, in fact, be incorrect when mutations are located near the primer region. This problems are severe and requires more careful analysis, since phylogeny is essential in studying the spreading of the pandemic and evolution of the virus [20–23].

Method and Materials

RNA reverse transcription and Nanopore sequencing

Sputum samples were collected from the 11 patients and inactivated under 56°C with 30 minutes in accordance with WHO and Chinese guidelines [24, 25]. Then, the total RNA was extracted from the samples according to the protocol of RNA isolation kit (RNAqueous® Total RNA isolation Kit, Invitrogen, China), and RNA concentration was determined by Qubit (ThermoFisher Scientific, China). Based on 2 pools of primers (98 pairs of primers in total) (Supplementary table1), the entire genomic sequence of SARS-CoV-2 was amplified segmentally by reverse transcription. Then, the library was built by Nanopore library construction kit (EXP-FLP002-XL, Flow Cell Priming Kit XL, YILIMART, China), while the adapter and barcode sequences were also added to the samples. The samples were sequenced on the Oxford Nanopore MinIon.

Data conversion and filtration

Fast5 format data was generated by the sequencer, and converted into fastq format with guppy basecaller (version 3.0.3). By applying NanoFilt (version 1.7.0) [26], we performed data filtration on the raw fastq data with the following criteria: read lengths should be longer than 100 bp after the removal of the adapter sequences, with an overall quality higher than 10. The clean data was kept for subsequent analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data composition

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 (Fig 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.

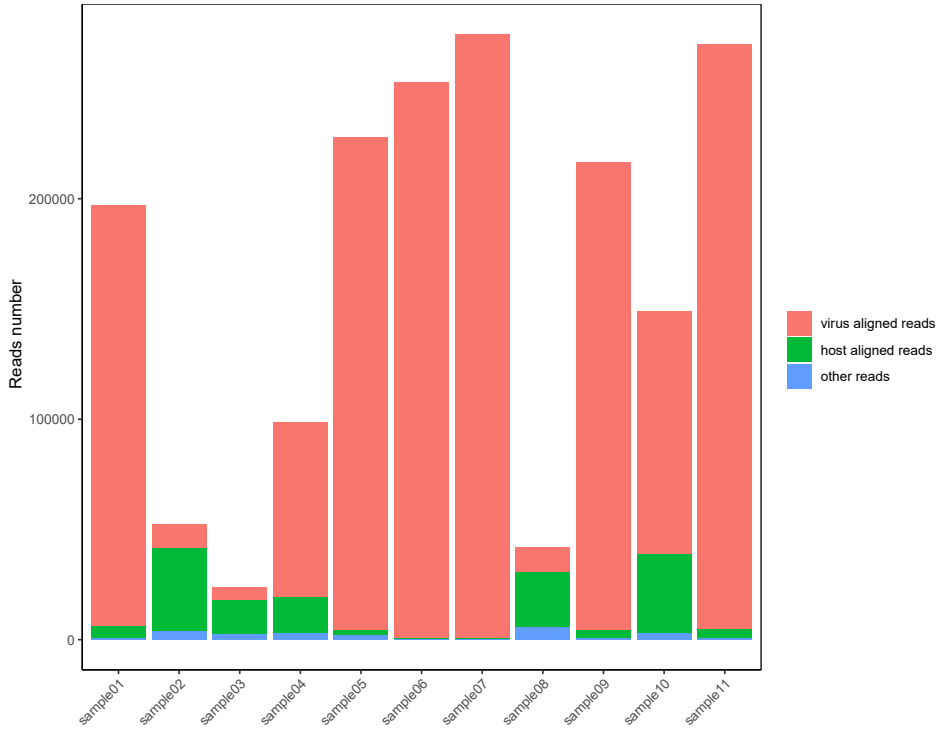

Figure 1. Compositions of the data sets. The heights of the columns represent the data composition. The data from the host, SARS-CoV-2 and other genomes are represented respectively with the colors pink, green and blue.

Remove false-positive chimeric reads

Chimeric Nanopore reads formed by random connection of multiplex RT-PCR amplicons were split into multiple segments if the locations of their bilateral ends match corresponding locations of PCR primers. These chimeric reads were able to 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 the primers on both viral and host genome, and split the identified chimeric reads into segments corresponding to PCR amplicons. This method allowed us to salvage large amount of sequencing data, leading to 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 [26] with default parameters for Oxford Nanopore reads. The aligned PCR amplicons were separated according to

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

corresponding primer pool. With the separated alignment results, the genomic variations with average quality larger than 10 were called with bcftools (version 1.8). Mutations with less than 10 supported reads were filtered. 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 within corresponding primer pool. The filtered mutations for different primer pools were then merged as the final mutations. Based on the gene information in SARS-CoV-2 reference genome, the final mutations were annotated by self-programming software.

Estimating viral load and primer efficiency

Assume that there are n pairs of primer (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

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

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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

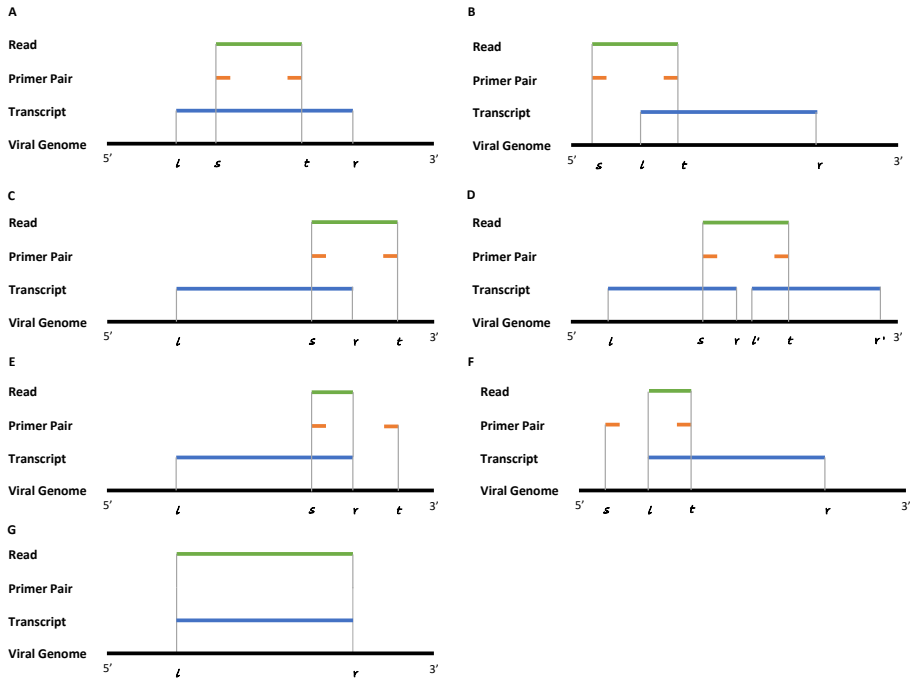

Figure 2. Relations between viral gene, primer pair and sequence. **A**: Primer pair spans fully on gene with sequence aligned with primers. **B**: Primer pair spans across gene boundaries at 5' with sequence aligned with primers. **C**: Primer pair spans across gene boundaries at 3' with sequence aligned 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 [27]. 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 samples (Figure 3). Based on these mutations, we discovered a total of 48 SNPs

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(Supplementary table2); 44 of them located on the genetic regions, including gene ORF1ab, S, N, M, ORF6 and ORF8. Gene ORF1ab contained 28 SNPs in 11 samples, and 20 of them were non-synonymous mutations. Non-synonymous mutations also occupied a large part of SNPs in the gene S, N, M and ORF8: 7 of 10 SNPs in gene S, 1 of 3 SNPs in gene N, 1 of 1 SNPs in gene M, and 1 of 1 SNPs in gene ORF8 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 two SNPs (loci 8,782 and 28,144), which were important SNP loci identified in recent phlogenetic analysis [20], that frequently demonstrates heterogeneous signals, suggesting the existence of at least two different virus strains in each host. However, we did not find creditable InDels (Insertions and Deletions), or structural variations, viral-host integration during the mutation analysis.

Evaluations of the amplification efficiencies 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 NO. 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.

Highly expressed genes from the host

We noticed the existence of human genome in the samples from an earlier study. We similarly analyze the corresponding genes and their expression levels. Aligning these data to human transcript database, we identified the segments with possibly human gene transcripts. The coverage depths of these genes were normalized based on the primer efficiencies (Supplementary table3). 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

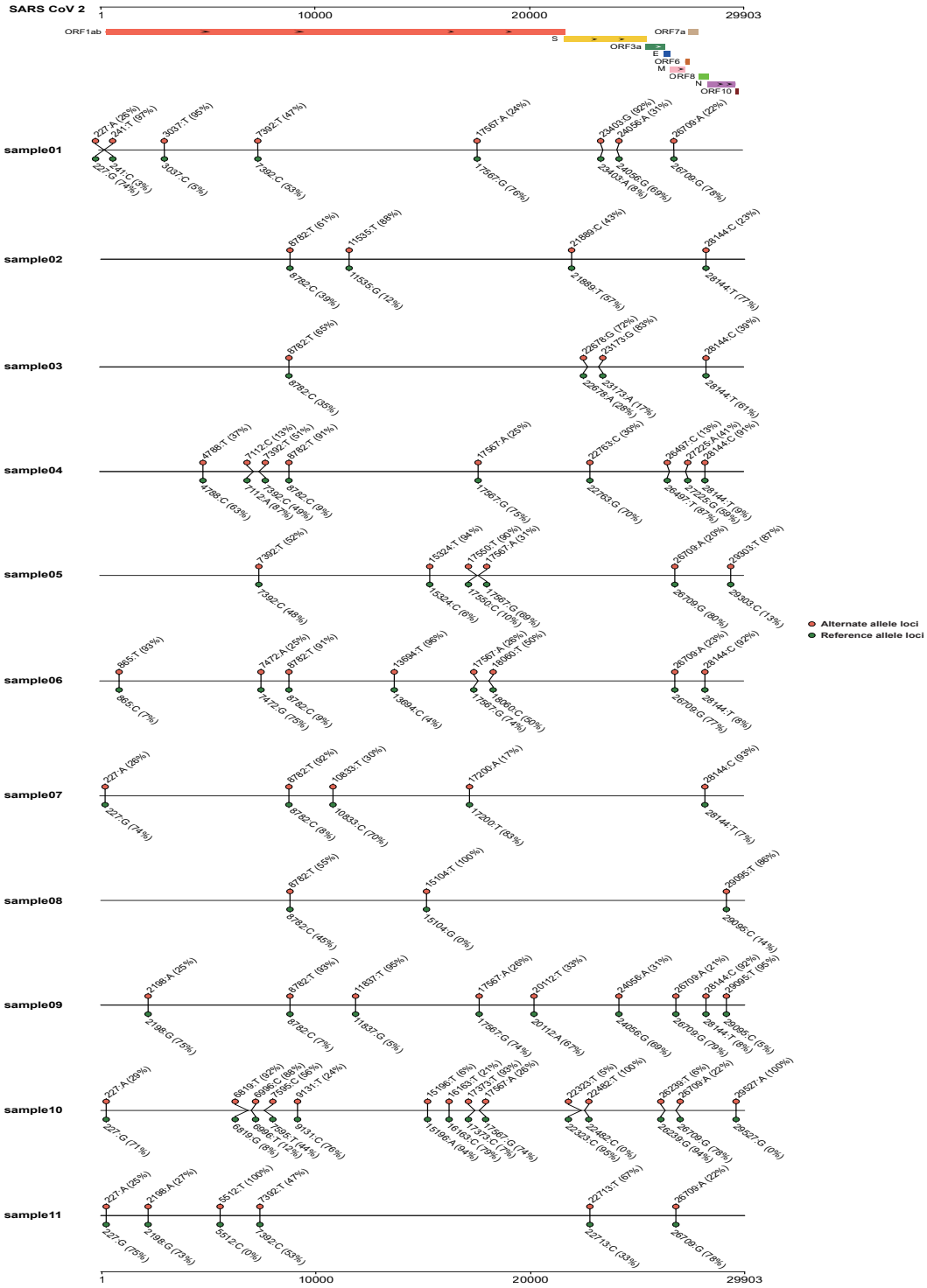

Figure 3. Distributions of SNPs among the 11 COVID-19 samples. The locations of genes in SARS-CoV-2 are displayed on top of the plot, with the distributions of SNPs for the 11 samples. The genotype of SNP and reference were list above and below the line respectively, with the percentages of their supported reads.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

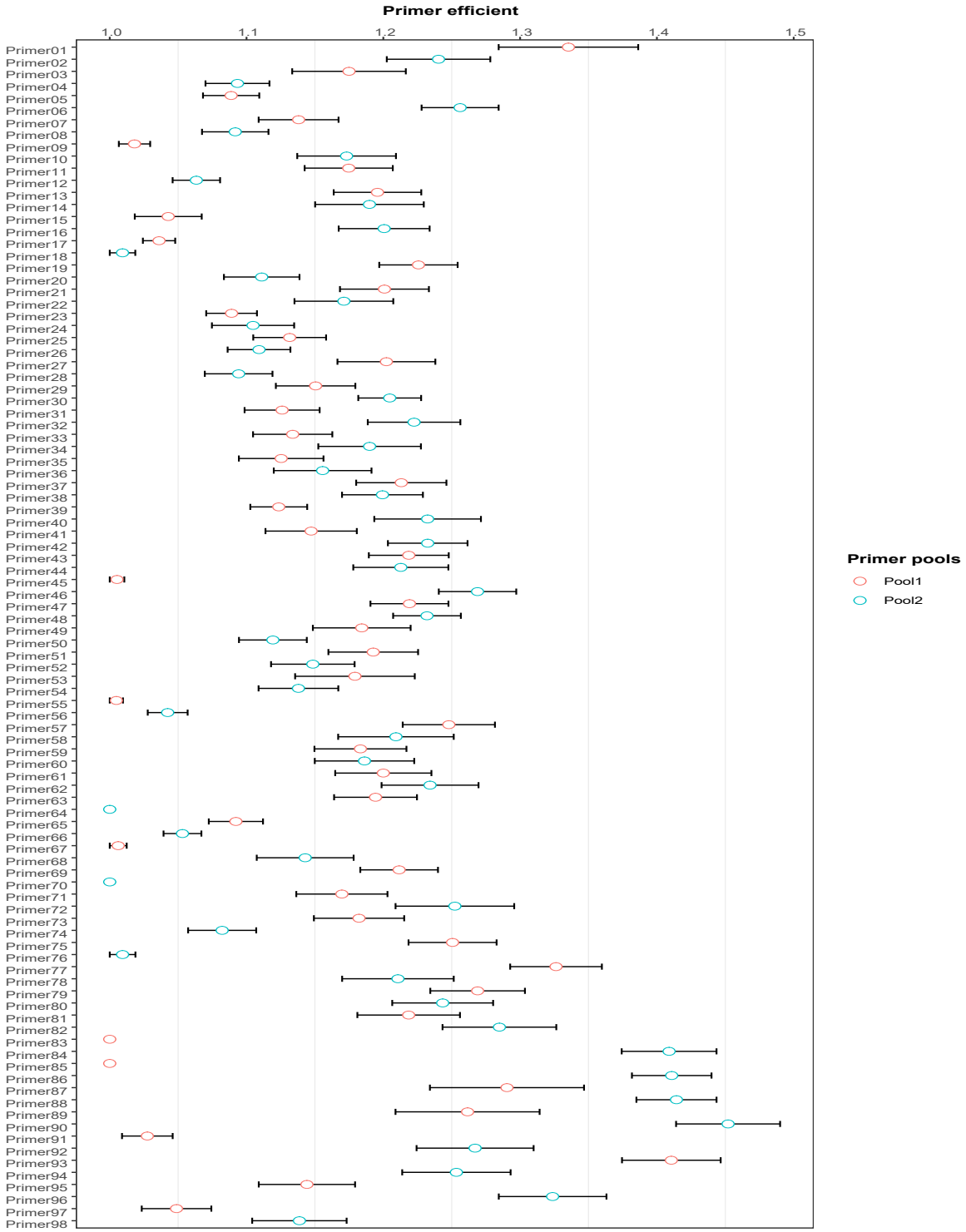

Figure 4. Amplification efficiencies for the 98 pairs of primers. The amplification efficiencies calculated for the primers, and their distributions in 11 samples. The primers in pool1 and pool2 are represented using red and blue markers respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

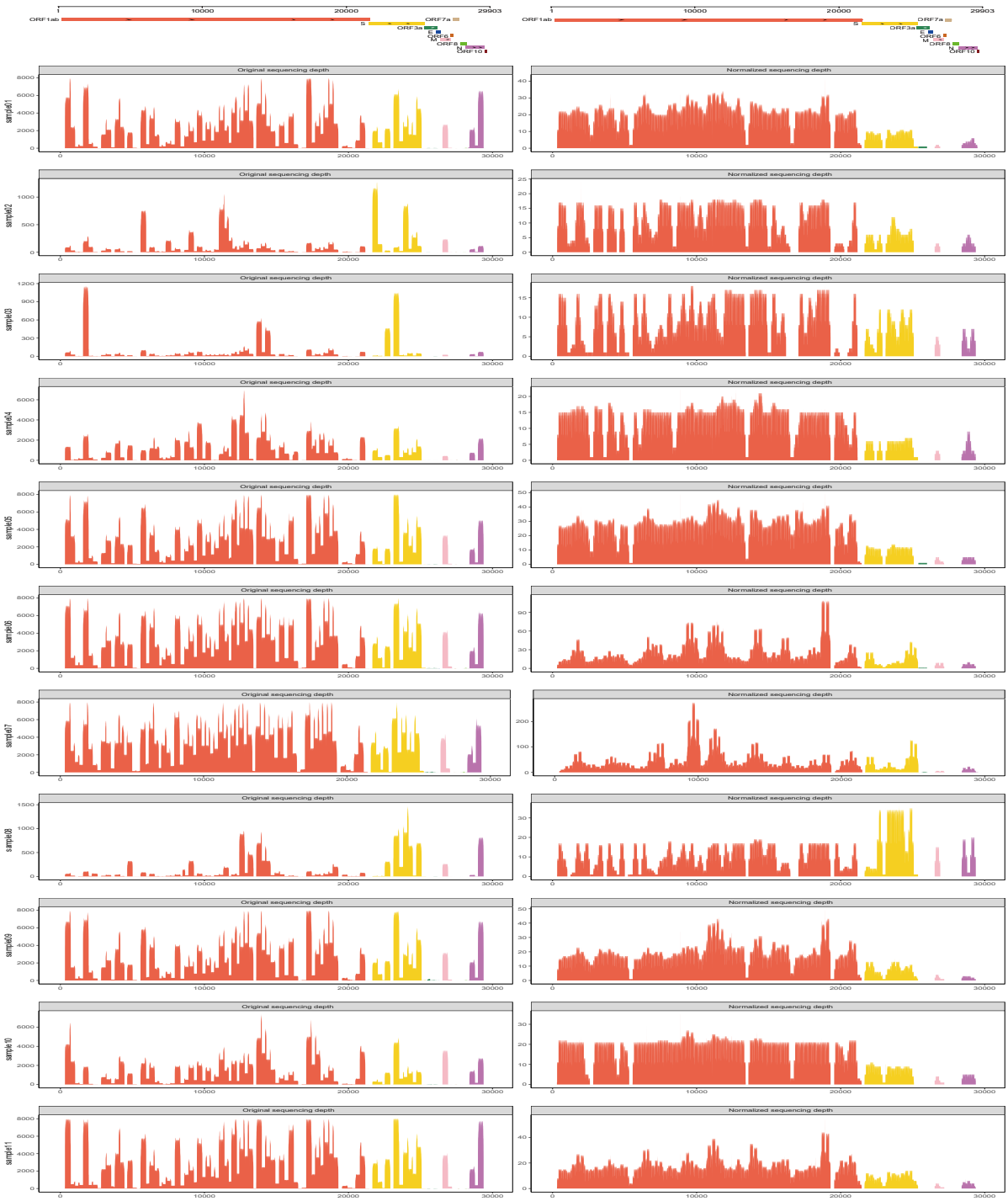

Figure 5. Original and normalized coverage depth for the 11 samples. The locations of the genes in SARS-CoV-2 are displayed on top of the plot, and marked with different colors. Each sample contains two plots, respectively representing the original (left) and normalized coverage (right) depths for the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

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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-PCR on Nanopore sequencing data, and studied the sputum samples of COVID-19 patients. Our results show enrichment of PCR amplicons which in total cover 99.7% of SARS-CoV-2 virus genome, confirming the reliability of the multiplex RT-PCR method in identifying COVID-19 infection.

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 [28, 29]. Our analysis of the host genome expression reveals several genes that are 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 of proteins makes up the major component of mucus secretions in normal lung mucus [30]. MUC5B has been reported to be highly expressed in lung disease, such as chronic obstructive pulmonary disease (COPD). A recent research demonstrated potential abundance of gene MUC1 and gene MUC5AC in the sputum of COVID-19 infected patients [31]. Its up-regulated transcription could cause mucociliary dysfunction and enhances lung fibrosis [32–34], thus might be relevant to the recent report of large quantity of mucus 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 [35]. 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 note that these genes are extracted based on a very limited amount of data, and the conclusions can be different with a large data set.

The viral expressions we uncovered show recurrent heterogeneous SNVs, not necessarily from the same loci, on every one of our 11 samples. This has severe implications on phylogenetic analysis on SARS-CoV-2, as the heterogeneous loci might be mis-linked during assembly.

This observation led us to believe that the coexistence of multiple virus strains may be a common phenomena. To verify this assumption, we downloaded all 4,075 SARS-CoV-2 strains available from GISAID at the time of preparation of this manuscript to analyze the complexity of co-occurrences of SNVs within the dataset. We consider the occurrence of the genotypes on two different loci. If every possible combination of the genotypes on the loci pair exists in at least one strain in the dataset, we say that the loci is *complex*. A complex loci is difficult to explain in phylogeny, but may occur by chance. To remove this possibility, instead we perform an analysis based on *SNP cliques*. We model a single SNP locus as a vertex and create an edge between a loci pair if and only if each of the four combinations of the genotypes of the pair is detected in at least one strain within the dataset. In this case, the existence of a large clique will be intractable to explain using phylogeny, as the chances become diminishingly small with the size of the clique.

From the dataset, we were able to uncover a partial clique of 24 vertices and 138 edges. Two maximal 8-cliques and 14 maximal 7-cliques from the partial cliques are shown in Fig 7. These cliques suggest fast mutations on their loci. On the other hand, genomic recombination between the strains is not detected with the software package RDP4 [36]. Hence, we believe the cliques would be best explained by the coexistence of multiple SARS-CoV-2 strains in a host. In this case, different loci from different viral strains might have been mixed and reported as a single SARS-CoV-2 genome, thus resulting in the prevalence of complex loci. To verify this hypothesis, we examined SARS-CoV-2 Nanopore sequencing data under accession SRR11347377, SRR11313280, SRR13313419, SRR13313494, SRR11140745, SRR11140747, SRR11140749 and SRR11140751, which were collected from San Diego, Wuhan and

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

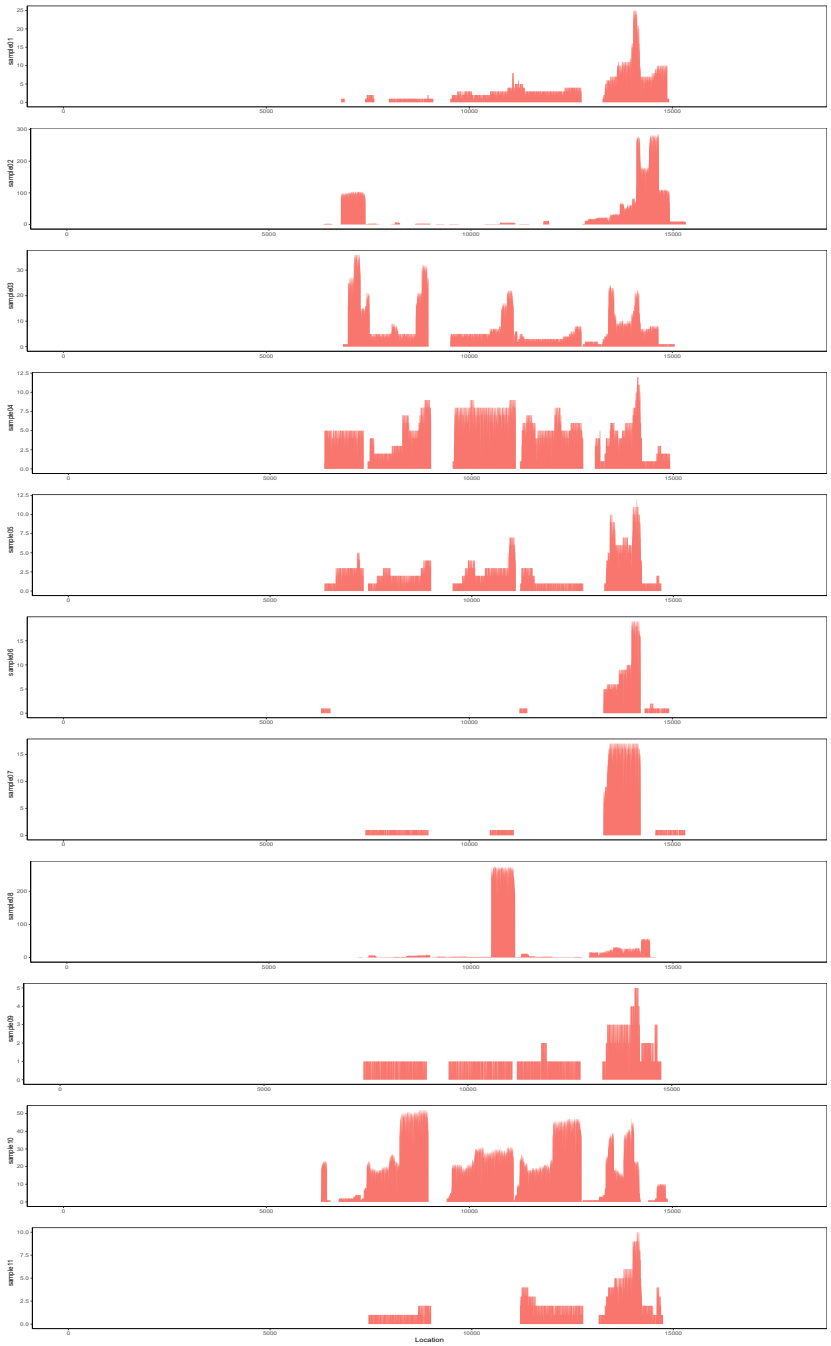

Figure 6. Normalized coverage depth on MUC5B gene in the 11 samples. In each plot, the X and Y coordinates respectively represent the length of MUC5B and the normalized coverage depth along the MUC5B gene.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Wisconsin Madison, respectively. Then, credible evidence of multiple viral strains were identified (See Supplementary Table 4). We also discovered that loci 8782, 26144 and 28144, which were important loci in recent SARS-CoV-2 phylogenetic analysis [20], were involved in some cliques of over 5 loci (Supplementary Figure 1).

In closing, we expect CovProfile to continue to help in detecting genomic variation on 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 data usage. The method also provides robust estimation of primer efficiency through a multi-parameter model. We hope the viral and host gene expressions we reported could help in the development of diagnosis and therapy strategies of SARS-CoV-2 infection candidates.

Data Availability

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive [37] in BIG Data Center [38], Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under BioProject PRJCA002503 with accession ID CRA002522 that are publicly accessible at https://bigd.big.ac.cn/gsa. All other data are available from the authors upon reasonable request.

Code Availability

CovProfile used in the main analysis pipeline are available at https://gitlab.deepomics.org/yyh/covprofile. All other code are available from the authors upon reasonable request.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

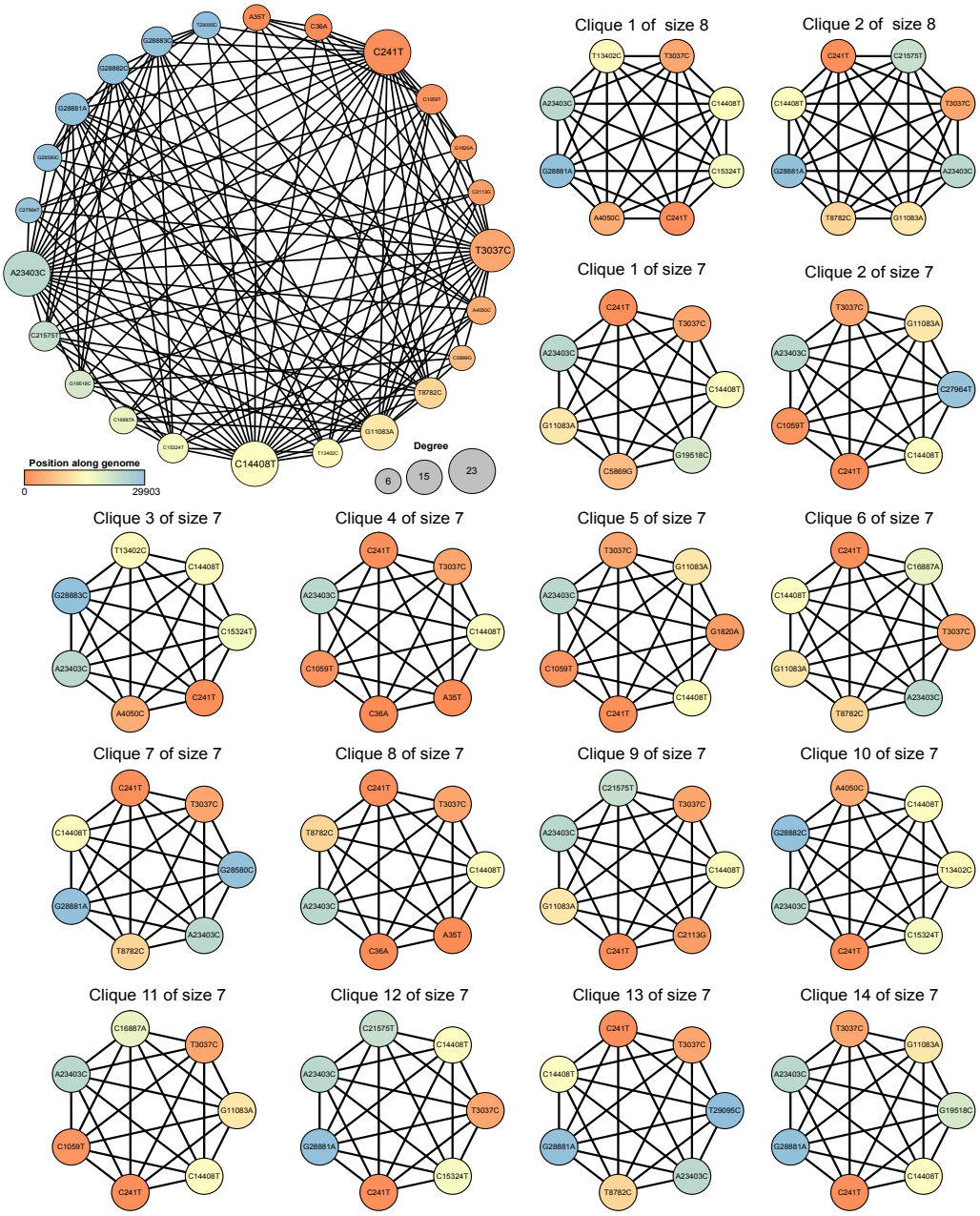

Figure 7. **Maximal SNP k-clique with $k > 6$.**

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian 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.

2. Chen Wang, Peter W Horby, Frederick G Hayden, and George F Gao. A novel coronavirus outbreak of global health concern. *The Lancet*, 395(10223):470–473, 2020.

3. Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, et al. A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798):265–269, 2020.

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

5. Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, pages 1–4, 2020.

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

7. Emmie de Wit, Neeltje van Doremalen, Darryl Falzarano, and Vincent J Munster. Sars and mers: recent insights into emerging coronaviruses. *Nature Reviews Microbiology*, 14(8):523, 2016.

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

9. Abdullah Assiri, Glen R. Abedi, Malak Al Masri, Abdulaziz Bin Saeed, Susan I. Gerber, and John 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.

10. Heshui Shi, Xiaoyu Han, Nanchuan Jiang, Yukun Cao, Osamah Alwalid, Jin Gu, Yanqing Fan, and Chuansheng Zheng. Radiological findings from 81 patients with COVID-19 pneumonia in wuhan, china: a descriptive study. *The Lancet Infectious Diseases*, 20(4):425–434, April 2020.

11. Cynthia Liu, Qiongqiong Zhou, Yingzhu Li, Linda V Garner, Steve P Watkins, Linda J Carter, Jeffrey Smoot, Anne C Gregg, Angela D Daniels, Susan Jervey, et al. Research and development on therapeutic agents and vaccines for covid-19 and related human coronavirus diseases, 2020.

12. Timothy P Sheahan, Amy C Sims, Sarah R Leist, Alexandra Schäfer, John Won, Ariane J Brown, Stephanie A Montgomery, Alison Hogg, Darius Babusis, Michael 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.

13. Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, et al. Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell*, 2020.

14. Pietro Hiram Guzzi, Daniele Mercatelli, Carmine Ceraolo, and Federico M Giorgi. Master regulator analysis of the sars-cov-2/human interactome. *bioRxiv*, 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

15. Marek Prachar, Sune Justesen, Daniel B Steen-Jensen, Stephan P Thorgrimsen, Erik Jurgons, Ole Winther, and Frederik Otzen Bagger. Covid-19 vaccine candidates: Prediction and validation of 174 sars-cov-2 epitopes. *bioRxiv*, 2020.

16. Mohan T Bolisetty, Gopinath Rajadinakaran, and Brenton R Graveley. Determining exon connectivity in complex mrnas by nanopore sequencing. *Genome Biology*, 16(1):204, 2015.

17. Samuel CB Stubbs, Barbara A Blacklaws, Benediktus Yohan, Frilasita A Yudhaputri, Rahma F Hayati, Brian Schwem, Edsel M Salvaña, Raul V Destura, James S Lester, Khin 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.

18. Joshua Quick, Nathan D Grubaugh, Steven T Pullan, Ingra M Claro, Andrew D Smith, Karthik Gangavarapu, Glenn Oliveira, Refugio Robles-Sikisaka, Thomas F Rogers, Nathan 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.

19. Shona C Moore, Rebekah Penrice-Randal, Muhannad Alruwaili, Xiaofeng Dong, Steven T Pullan, Daniel Carter, Kevin Bewley, Qin Zhao, Yani Sun, Catherine Hartley, et al. Amplicon based minion sequencing of sars-cov-2 and metagenomic characterisation of nasopharyngeal swabs from patients with covid-19. *medRxiv*, 2020.

20. Peter Forster, Lucy Forster, Colin Renfrew, and Michael Forster. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences*, page 202004999, April 2020.

21. Xiaolu Tang, Changcheng Wu, Xiang Li, Yuhe Song, Xinmin Yao, Xinkai Wu, Yuange Duan, Hong Zhang, Yirong Wang, Zhaohui Qian, Jie Cui, and Jian Lu. On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, March 2020.

22. Liangsheng Zhang, Jian-Rong Yang, Zhenguo Zhang, and Zhenguo Lin. Genomic variations of SARS-CoV-2 suggest multiple outbreak sources of transmission. February 2020.

23. Aiping Wu, Peihua Niu, Lulan Wang, Hangyu Zhou, Xiang Zhao, Wenling Wang, Jingfeng Wang, Chengyang Ji, Xiao Ding, Xianyue Wang, Roujian Lu, Sarah Gold, Saba Aliyari, Shilei Zhang, Ellee Vikram, Angela Zou, Emily Lenh, Janet Chen, Fei Ye, Na Han, Yousong Peng, Haitao Guo, Guizhen Wu, Taijiao Jiang, Wenjie Tan, and Genhong Cheng. Mutations, recombination and insertion in the evolution of 2019-nCoV. March 2020.

24. World Health Organization. Clinical management of severe acute respiratory infection (sari) when covid-19 disease is suspected: interim guidance, 13 march 2020. Technical documents, 2020.

25. Chinese clinical guidance for covid-19 pneumonia diagnosis and treatment, http://kjfy.meetingchina.org/msite/news/show/cn/3337.html, 2020.

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

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

28. Ruby White, Christophe Pellefigues, Franca Ronchese, Olivier Lamiable, and David Eccles. Investigation of chimeric reads using the minion. *F1000Research*, 6, 2017.

29. Yifei Xu, Kuiama Lewandowski, Sheila Lumley, Steven Pullan, Richard Vipond, Miles Carroll, Dona Foster, Philippa C Matthews, Timothy Peto, and Derrick Crook. Detection of viral pathogens with multiplex nanopore minion sequencing: be careful with cross-talk. *Frontiers in microbiology*, 9:2225, 2018.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

30. Michelle G Roy, Alessandra Livraghi-Butrico, Ashley A Fletcher, Melissa M McElwee, Scott E Evans, Ryan M Boerner, Samantha N Alexander, Lindsey K Bellinghausen, Alfred S Song, Youlia M Petrova, et al. Muc5b is required for airway defence. *Nature*, 505(7483):412–416, 2014.

31. Nanshan Zhong. Perspectives of the management of covid-19 infection in china, 2020.

32. Laura A Hancock, Corinne E Hennessy, George M Solomon, Evgenia Dobrinskikh, Alani Estrella, Naoko Hara, David B Hill, William J Kissner, Matthew R Markovetz, Diane E Grove Villalon, et al. Muc5b overexpression causes mucociliary dysfunction and enhances lung fibrosis in mice. *Nature communications*, 9(1):1–10, 2018.

33. Susan K Mathai, Stephen Humphries, Jonathan A Kropski, Timothy S Blackwell, Julia Powers, Avram D Walts, Cheryl Markin, Julia Woodward, Jonathan H Chung, Kevin K Brown, et al. Muc5b variant is associated with visually and quantitatively detected preclinical pulmonary fibrosis. *Thorax*, 74(12):1131–1139, 2019.

34. Erik P Lillehoj, Kosuke Kato, Wenju Lu, and Kwang C Kim. Cellular and molecular biology of airway mucins. In *International review of cell and molecular biology*, volume 303, pages 139–202. Elsevier, 2013.

35. Dimitra Gkika, Loic Lemonnier, George Shapovalov, Dmitri Gordienko, Céline Poux, Michela Bernardini, Alexandre Bokhobza, Gabriel Bidaux, Cindy Degerny, Kathye 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.

36. Darren P. Martin, Ben Murrell, Michael Golden, Arjun Khoosal, and Brejnev Muhire. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, 1(1), March 2015.

37. Yanqing Wang, Fuhai Song, Junwei Zhu, Sisi Zhang, Yadong Yang, Tingting Chen, Bixia Tang, Lili Dong, Nan Ding, Qian Zhang, Zhouxian Bai, Xunong Dong, Huanxin Chen, Mingyuan Sun, Shuang Zhai, Yubin Sun, Lei Yu, Li Lan, Jingfa Xiao, Xiangdong Fang, Hongxing Lei, Zhang Zhang, and Wenming Zhao. GSA: Genome sequence archive $*.Genomics, Proteomics\&Bioinformatics, 15(1) : 14 - -18, February 2017.$

38. Zhang Zhang, Wenming Zhao, Jingfa Xiao, Yiming Bao, Shunmin He, Guoqing Zhang, Yixue Li, Guoping Zhao, Runsheng Chen, Yang Gao, Chao Zhang, Liyun Yuan, Guoqing Zhang, Shuhua Xu, Chao Zhang, Yang Gao, Zhilin Ning, Yan Lu, Shuhua Xu, Jingyao Zeng, Na Yuan, Junwei Zhu, and et al. Database resources of the national genomics data center in 2020. *Nucleic Acids Research*, November 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.05.026146; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgments

We thank every participants contributing to this work.

Author Information

These authors contributed equally: Yonghan Yu, Zhengtu Li, Yinghu Li and Le Yu.

Affiliations

City University of Hong Kong Shenzhen Research Institute, Shenzhen 518057, China
Yonghan Yu, Yinhu Li, Wenlong Jia, Yiqi Jiang and Shuai Cheng Li
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
Zhengtu Li and Feng Ye
Beijing YuanShengKangTai (ProtoDNA) Genetech Co. Ltd., Beijing 100190, China
Le Yu

Contributions

SCL conceived the idea.
SCL and FY supervised the project.
SCL and YY designed the algorithms
SCL, WJ, YY and YL discussed and designed the experiments.
YY and YL implemented the code and conducted the analysis.
YJ conducted the clique 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.

Corresponding authors

Correspondence to Feng Ye and Shuai Cheng Li.

Ethics declarations

Competing interests

All authors declare no competing interest.

Ethics oversight

This study was approved by the First Affiliated Hospital of Guangzhou Medical University, and the sample and data collection procedures were conducted in accordance with the principles expressed in the Declaration of Helsinki. All patients provided written informed consent, and volunteered to receive investigation for scientific research.