34 (2022) 201116

Contents lists available at ScienceDirect

Human Gene

journal homepage: www.journals.elsevier.com/human-gene

Differential gene expression of SARS-CoV-2 transcriptome provides insight into the design of more sensitive diagnostic tests

Mohadeseh Ahmadi [a], Reza Alizadeh-Navaei [a], Mohammadreza Haghshenas [b], Tahoora Mousavi [c], Majid Saeedi [d], Akbar Hedayatizadeh-Omran [a], Reza Valadan [e,*]

[a] Gastrointestinal Cancer Research Center, Non-communicable Diseases Institute, Mazandaran University of Medical Sciences, Sari, Iran
[b] Department of Microbiology, Molecular and Cell-Biology Research Center, Faculty of Medicine, Mazandaran University of Medical Sciences, Sari, Iran
[c] Molecular and Cell Biology Research Center, Hemoglobinopathy Institute, Mazandaran University of Medical Sciences, Iran
[d] Department of Pharmaceutics, Faculty of Pharmacy, Mazandaran University of Medical Sciences, Sari, Iran
[e] Department of Immunology, Molecular and Cell Biology Research Centre, Faculty of Medicine, Mazandaran University of Medical Sciences, Sari, Iran

ARTICLE INFO

Keywords:
Covid-19
Gene expression
SARS-CoV-2
RT-PCR

ABSTRACT

The Coronavirus disease 2019 (COVID-19) pandemic is being addressed through RT-PCR, a frontline diagnostic technique. We evaluated gene expression patterns to improve the accuracy and sensitivity of current diagnostic tests. We downloaded relevant next-generation sequencing (NGS) data from the Sequence Read Archive (SRA) database, checked for quality, and mapped them onto the target reference sequence. It was determined that ORF1ab, N, S, and ORF8 genes are mainly expressed based on the results of the quantitative evaluation after normalization by HPRT and elimination of insufficient expression data. ORF8, ORF3a, and M genes were found to have higher expression values than the E gene as a routine RT-PCR detector gene (p*0.05). M gene expression values are also close to ORF8 values. Taking into account the importance of differential expression of genes in the design of diagnostic kits as well as the findings of from this study, it is likely that the M gene is worth further investigation due to its high expression and low mutation rate.

1. Introduction

COVID-19 is associated with high mortality and morbidity caused by (SARS-CoV-2). The disease is now considered a pandemic and an important global health issue. The disease mainly affects the upper and lower respiratory system, causing various signs and symptoms such as fever, cough, headache, malaise, fatigue, and dyspnea (Zimmermann and Curtis, 2020). In a subset of patients, the disease progresses to the fatal stage leading to bilateral pulmonary consolidation and alveolar damage that eventually ends with acute respiratory distress syndrome (ARDS) (Batah and Fabro, 2021).

A newly identified coronavirus, SARS-CoV-2, is also known as the severe acute respiratory syndrome coronavirus. Compared to some closely related coronaviruses such as SARA-CoV and bat SARS-like coronaviruses (bat SL-CoVZC45), the genomic sequence is very similar. Meta-transcriptomic analysis revealed that SARS-CoV-2 contains a single-stranded RNA genome of about 29,811 bases, encoding a variety of structural and non-structural proteins. In the SARS-CoV-2 genome, about two-thirds of the open reading frames (ORF) encode for two large polyproteins, which are then processed into at least 16 nonstructural proteins by the viral protease. As well as four structural proteins, the genome encodes for spikes (S), envelopes (E), membranes (M), and nucleocapsid (N) (Kirtipal et al., 2020). A study of the Coronavirus replication cycle and assembly inside the cells has revealed that positive-sense viral RNA is available for translation into viral proteins and serves as a template for replicating many subgenomic mRNAs. Subgenomic mRNAs are a 3-nested set of genomic RNAs containing a 70–100 bases leader RNA that is identical to the 5 end of the genomic RNA, joined to subsets of nested transcripts from 3 end of the genomic RNA (Thiel et al., 2003). The production of subgenomic RNA is regulated by a conserved sequence (AAACGAAC for SARS-CoV) called transcription-regulating sequences (TRS) (Thiel et al., 2003). The presence of subgenomic RNA is a defining feature of the order Nidovirales (Van Vliet et al., 2002).

Amplification of fragments of the viral genome is the primary method for detecting SARS-CoV-2 at the molecular level. So far, the most developed molecular detection method for SARS-CoV-2 is real-time reverse transcriptase-polymerase chain reaction (rRT-PCR); however,

* Corresponding author.
E-mail address: rvaladan@mazums.ac.ir (R. Valadan).

https://doi.org/10.1016/j.humgen.2022.201116
Received 26 December 2021; Received in revised form 20 August 2022; Accepted 3 October 2022
Available online 10 October 2022

other approaches such as nested PCR and colorimetric loop-mediated isothermal amplification have also been described. Clinical specimens can be tested with rRT-PCR, including bronchoalveolar lavage fluid (BALF), sputum, fibrobronchoscope brushes, nasopharyngeal and oropharyngeal swabs, feces, and blood, but with varying results and sensitivity, which makes it necessary to test multiple specimens to in order to improve sensitivity and reduce false-negative results (Wang et al., 2020). Different clinical samples also display different viral loads, ranging from a barely detectable one in urine and blood to an abundant one in sputum, nasal swab, and BALF (Pan et al., 2020; Zhou et al., 2020). As SARS-CoV-2 naturally binds to cells that express ACE-2 as a receptor on their surface, viral load is correlated with that of tissues that express ACE-2. An RNA-sequencing data from the Human Cell Atlas database shows that nasal epithelium, specifically goblet cells and ciliated cells, express the highest expression level of ACE-2 (https://arxiv.org/abs/2003.06122) (Sungnak et al., 2020). Numerous factors can affect the sensitivity of a diagnostic test, including specimen types, sample generation method, and laboratory technical factors. However, caution needs to be taken when designing the test itself to ensure that it is as sensitive as possible (Wiersinga et al., 2020). The commercial and in-house diagnostic rRT-PCR assays are currently based on primers and probed designed for N, ORF1ab, and E genes even though the differential gene expression pattern of SARS Cov-2 in the infected cell may affect the sensitivity of the test (Islam and Iqbal, 2020). Therefore, this study was aimed to evaluate differential gene expression pattern SARS Cov-2 in the upper respiratory epithelial cells based on RNA-seq data deposited on the public database to improve the accuracy and sensitivity of current diagnostic tests.

2. Methods

2.1. Selection of sequence files

Relevant data from many human sequencing studies were filtered and downloaded in July 2020 from the Sequence Read Archive database SRA database (https://www.ncbi.nlm.nih.gov/sra/) based on the following criteria: 1- The next-generation sequencing (NGS) data contained SARS CoV-2 genomic data and was derived only from RNA-seq strategy. 2- Bio samples also contained shedding host cells, such as samples obtained from oropharyngeal(NP), nasopharyngeal(OP), and endotracheal(ETA) swabs. 3- The sequence data were also filtered for the presence of host cell transcriptomic data. 4- Samples with Illumina sequencing platform and PAIRED layout were considered.

2.2. Data processing and quality control of reads

Downloaded bio projects were imported into the CLC genomics workbench (CLC, V20.0.2; Qiagen, USA). The quality of the data was verified. To achieve a reliable and high-quality result, trimming was employed and poor-quality bases were eliminated.

2.3. Target reference building

In this study, the target sequence was generated by assembling SARS CoV-2 (GenBank: NC_045512.2) and HPRT sequences to serve as a reference for mapping experiment reads. HPRT was chosen as a housekeeping gene because of its stable expression during viral infection in the upper respiratory tract (Resa et al., 2014).

2.4. Mapping the reads to the target reference built

After trimming, the sequences were mapped to on the target reference sequence. As noted earlier, this sequence was created by assembling SARS CoV-2 and HPRT sequences in CLC software by the "create tracklist" option.

2.5. Analysis of SARS CoV-2 transcripts expression using RNA-Seq data

Using CLC Genomics Workbench (CLC, V20.0.2; Qiagen, USA), transcripts per million (TPM) for Clean reads mapped to a target were calculated for each gene (Zhao et al., 2020). On 99 patients with COVID-19, the expression of SARS-CoV-2 genes was determined by using CLC Genomics Workbench (V20.0.2; Qiagen, USA). In all samples downloaded their RNA sequences, a comparison between expression values was performed after HPRT normalization. Mapped expression values were used to determine differential gene expression.

2.6. Statistical analysis

Statistical analysis was conducted to compare the SARS-CoV-2 gene by SPSS 23 statistical software. Using a *t*-Test, differential expression values were calculated among COVID-19 patients. Diagrams were drawn using the R Studio software. A *P*-value of 0.05 was considered significant.

3. Results

3.1. Raw data selection

A total of 81 bio projects containing 42,600 biosamples were initially selected by searching the appropriate sequences in the SRA database. Up to 688 biosamples were obtained from the remaining two bio projects after removing bio projects that did not meet our inclusion criteria (Table 1). In the remaining bioprojects, there were biosamples without data to download or non-mentioned types of biosamples (e.g., Blood, Stool). Some biosamples contained low or no identified reads associated with SARS-Cov2 or the host cells (to perform normalization by HPRT). In total, 118 biosamples were analyzed for differential expression.

3.2. Processing and quality assessment of raw data

Sequence base quality scores are presented in Fig. 1. The median quality score in plot A is started from less than five bases and then rises. In plot B, the distribution of the average read quality is fairly in the upper range of the plot. Plot C shows very high-quality RNA-Seq data because of having narrower distribution than the theoretical one. The distribution of the average read quality is relatively tight in the upper range of plot D.

3.3. Genes expression of SARS CoV-2 in COVID-19 patients

In this study, the differential expression value of eleven genes in 118 COVID-19 patients was analyzed. After normalization by HPRT and elimination of insufficient expression data, 99 samples were retained for final analysis. The result of the quantitative evaluation showed that the ORF1ab has the highest expressed transcript value, followed by S, N, ORF8, ORF3a, M, ORF7a, E, ORF6, ORF7b, and ORF10 (Fig. 2). Genes that expression widely assessed in a diagnostic test for COVID-19 are ORF1ab, N, and E genes (Corman et al., 2020). According to Fig. 3, our analysis showed that ORF1ab, N, S, and ORF8 genes are mainly expressed.

Table 1
Summary of raw data of SARS-CoV-2 infected human samples extracted from the SRA database.

Data properties	Bioprojects		Total
	PRJNA6367748	PRJNA656695	
Experiments	516	171	687
Biosamples	517	171	688
Data volume, Gbases	40	22	62
Data volume, Mbytes	23,430	11,107	24,537

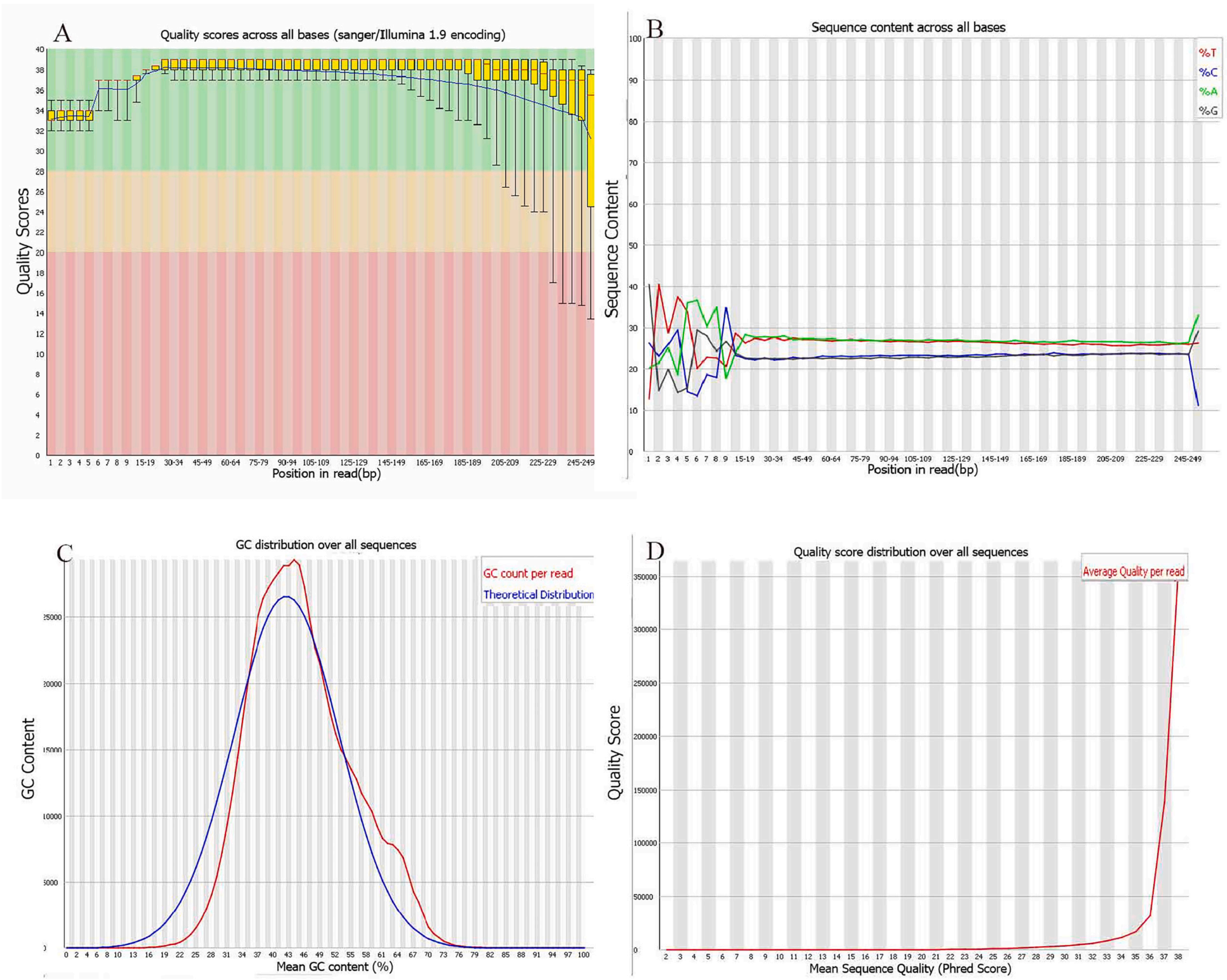

Fig. 1. (A) This plot shows the quality of bases across a read with a green box indicating very good quality for all bases, (B) Per base sequence content plot which shows the parallel lines of bases in the length of reads, (C) per sequence GC plot shows very high-quality RNA-Seq data because the distribution of GC content is normally based on modal GC content (theoretical distribution), and (D) the right skewed of sequence base quality score plot shows the high quality of reads (the minimum quality of score is about 27).

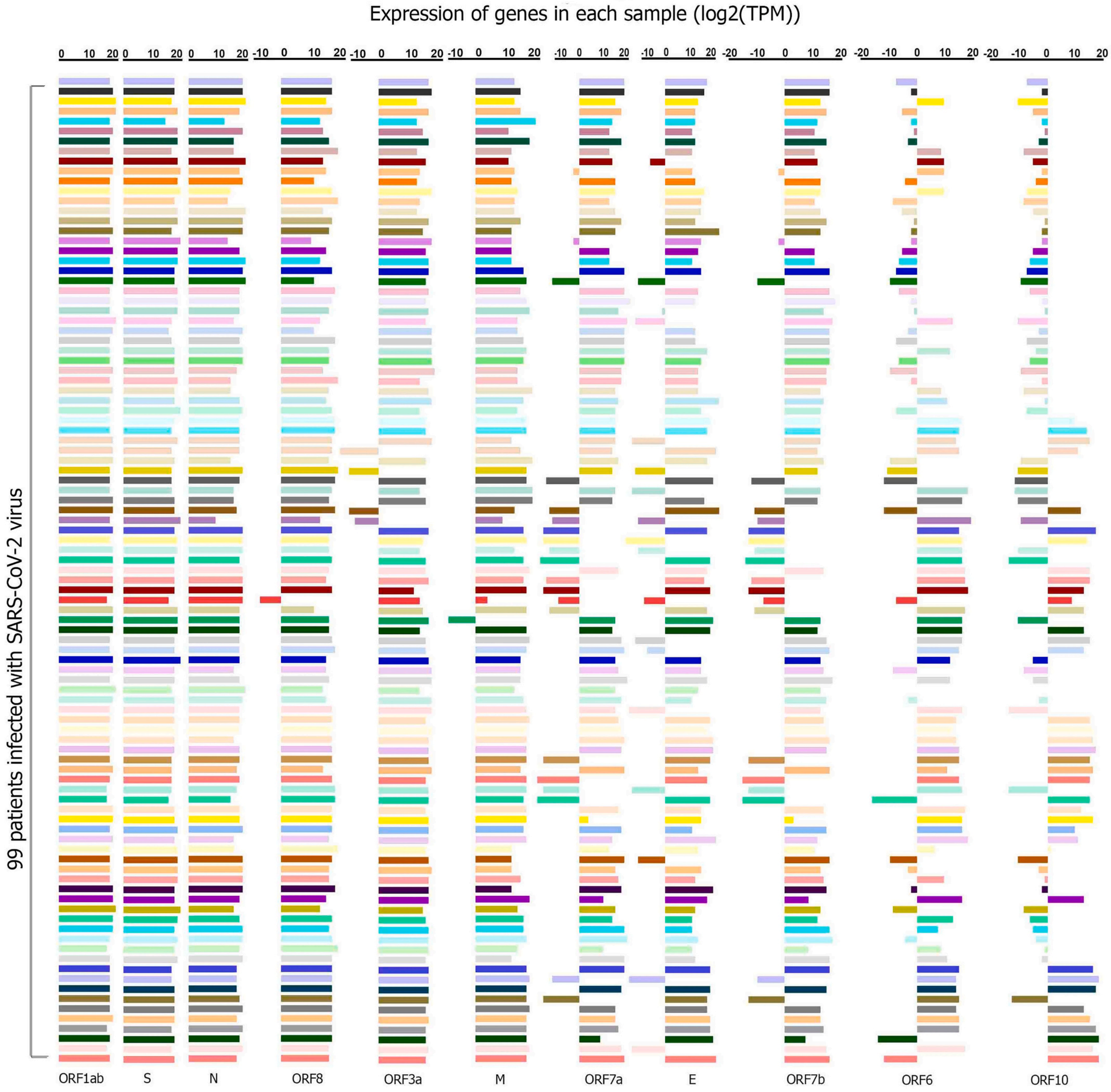

Fig. 2. The values of gene expression of eleven SARS-CoV-2 genes have been presented individually for each sample based on log2(TPM) indices. The highest amount of the expression value belongs to ORF1ab. The expression of ORF1ab, S, and N genes was positive in all samples, and also ORF8 and M genes, except in one case, showed positive expression in all samples.

3.4. Expression comparing within SARS-CoV-2 transcriptomes in samples from the upper respiratory tract

We calculated the expression value of SARS-Cov-2 genes in the 99 COVID-19 patients using transcriptomics data from oral swabs. Since current commercial and in-house diagnostic RT-PCR assay is based on primers and probes designed for N, ORF1ab, and E genes, exploring novel differential gene expression patterns of SARS-Cov2 can improve the specificity of diagnosis. Overall, in this study, the TPM value showed that ORF1ab, S, N, and ORF8 genes have the most expression value. As shown in Fig. 4, the ORF1ab gene has more expression value than N ($p < 0.05$*). Also, seven genes in SARS-Cov-2 (S, N, ORF1ab, M, ORF3a, ORF7a, and ORF8) have more expression significantly compared to the E gene ($p < 0.005$ **). Except for ORF1ab and S genes, no other gene had significantly higher expression rather than ORF8 ($p < 0.0005$ ***). Furthermore, three genes (S, N, and ORF1ab) showed significantly more expression compared to ORF3 and M genes ($p < 0.0000$ ****). The only gene with more expression than the S gene was ORF1ab ($P > 0.05$). The expression comparison between some important SARS-CoV-2 genes was summarized in Table 2. As shown in this table, the increased expression of genes that were placed in the vertical section was compared based on the *P*-value rather than other genes and significant increases were highlighted. As an example, a significant increase in expression has been observed for ORF1ab rather than for other genes except for S.

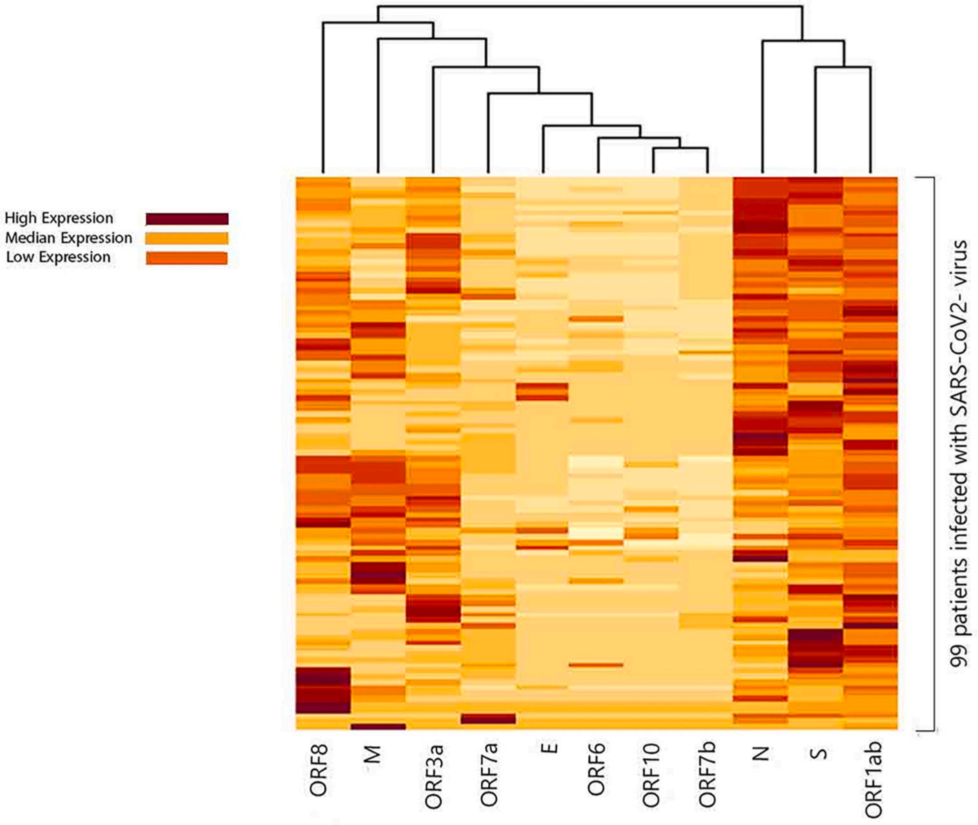

Fig. 3. Heat map of differentially expressed genes (DEGs) based on TPM expression value. The cream color represents a lower expression level. Its most scattered area is related to E, ORF6, ORF10, and ORF7b; the brown color represents a higher expression level, and its most scattered area is associated with ORF1ab, S, N ORF8, and M.

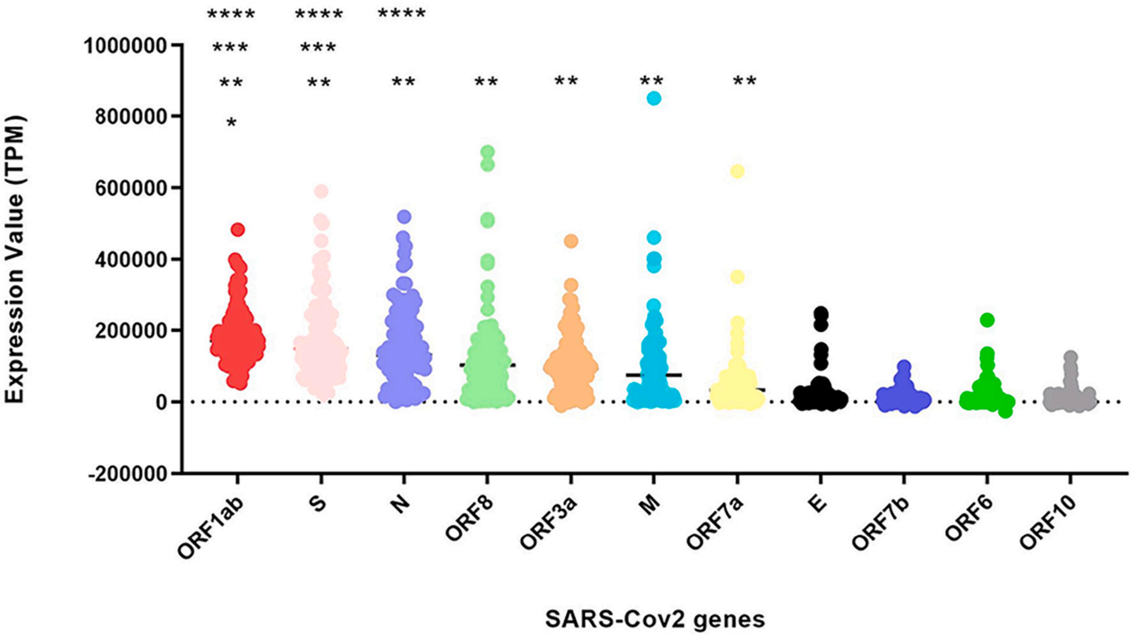

Fig. 4. Visualization of the difference in the average expression values of each gene of the SARS-COV-2. ORF1ab gene has significantly more expression values than N ($p < 0.05$ *). ORF1ab, S, N, ORF8, ORF3a, M, and ORF7a have more expression than E ($p < 0.005$ **). Also, ORF1ab and S have more expression than ORF8 ($p < 0.0005$ ***), and ORF1ab, S, and N are higher than ORF3a and M genes ($p < 0.0000$ ****).

Table 2
The summary of expression comparison between indicated SARS-CoV-2 genes.

	ORF1ab	S	N	ORF8	M	E
ORF1ab		0.287	0.032	0.000	0.000	0.000
S			0.032	0.006	0.000	0.000
N	0.032	0.340		0.610	0.002	0.000
ORF8					0.020	0.000
M						0.000
ORF3a					0.731	0.000

The p-value in green blocks represents the higher statistically significant expression in vertical genes than in other horizontal genes.

4. Discussion

Early detection of SARS-CoV-2 infected persons is crucial to reducing the rate of disease transmission (Younes et al., 2020; Pollard et al., 2020). RT-PCR is a frontline diagnostic technique, according to the WHO declaration, for rapid detection of SARS-CoV-2 in clinical samples (Islam and Iqbal, 2020; Sule and Oluwayelu, 2020). This test has a restriction in that the false-negative results can further spread the SARS-CoV-2 virus within a community (Younes et al., 2020; Arevalo-Rodriguez et al., 2020; Dao et al., 2021).

To improve the accuracy and sensitivity of current diagnostic tests, we hypothesized that evaluation of the expression value in SARS-CoV-2 transcriptomes could provide new insights into high-sensitivity target detection.

We analyzed the expression values of eleven genes within 99 samples from the SRA database that had high-quality sequencing. In the present study, we found that ORF8, ORF3a, and M genes have higher expression values than the E gene used as a routine detector gene. This makes them suitable for testing and use as detector genes for infected cases. There is no statistically significant difference in the expression of ORF8 and N, used as a routine detector gene.

Quantitative comparison of expression value shows that the ORF1ab is the most abundantly expressed transcript, followed by S, N, ORF8, ORF3a, M, ORF7a, E, ORF7b, ORF6, and ORF10.

It has been demonstrated that the most abundant expression value in N transcripts in SARS-CoV-2 infected cells by direct RNA sequencing technique (using nanopore arrays), and also more expression values are followed by S, ORF7a, ORF3a, ORF8, M, E, ORF6 and ORF7b in their study (Kim et al., 2020). Their Sample was extracted RNA from Vero cells infected with SARS-CoV-2, which were isolated from a patient diagnosed with COVID-19. Based on our study result, the order of the most expression value of SARS-CoV-2 transcripts partly was similar to this study (N and S genes) (Kim et al., 2020). Even though this study was conducted by the nanopore arrays technique as a powerful method to characterize the transcripts of virus-infected cells, the sample size investigated was not comparable to our study (one vs. ninety-nine).

To reduce false-negative results due to biosample selection, we ensured swabs samples contained sufficient quantities of SARS-CoV-2, and by using a reference genome containing HPRT, which showed stable expression during viral infection (Resa et al., 2014), we could able to reduce gene expression diversity errors caused by different viral loads in 99 evaluated samples.

Rahimi et al. reported the 17 high-frequency missense and synonymous mutations. The genomic analysis of various SARS-CoV-2 genes in some studies revealed multiple mutations, including ORF1ab, N, S, M, E, ORF8, ORF3a, ORF7, ORF10, and ORF6. Among them, most mutations were related to ORF1ab, *S*, as well, as *ORF8* genes. Such mutations compromise the sensitivity of RT-PCR results because of their impact on matching the RNA sequence of the specimen and the primers, which can lead to false negatives (Rahimi et al., 2021).

Although the use of genes with high expression can improve the sensitivity of the test, but mutations in these genes lead to incorrect results. Since only two or three genes are used to detect infected cases by Real-time PCR, utilization of genes with lower expression rather than ORF1ab or S while having lower mutation could improve the test's specificity. Compared to ORF8 and ORF3a, which have higher expression values followed by N, the M gene expression value is lower, but not significant, and no mutations have been reported on the M gene so far (Rahimi et al., 2021). Hence, the M gene seems appropriate for a more comprehensive evaluation.

5. Conclusion

Considering the effect of differential expression of genes on the improvement of the accuracy and sensitivity of the diagnostic test and the findings of this study, due to the high expression and low mutation rate, the M gene is suitable for further study.

Ethics approval

This study was approved by the Mazandaran University of Medical Sciences with grant number: (7326).

CRediT authorship contribution statement

Mohadeseh Ahmadi: Methodology, Software, Formal analysis, Writing – original draft, Visualization. **Reza Alizadeh-Navaei:** Software, Validation, Formal analysis. **Mohammadreza Haghshenas:** Supervision, Investigation. **Tahoora Mousavi:** Methodology, Software. **Majid Saeedi:** Supervision, Validation. **Akbar Hedayatizadeh-Omran:** Supervision, Investigation. **Reza Valadan:** Conceptualization, Methodology, Software, Validation, Investigation, Resources, Writing – review & editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Arevalo-Rodriguez, I., Buitrago-Garcia, D., Simancas-Racines, D., Zambrano-Achig, P., Del Campo, R., Ciapponi, A., et al., 2020. False-negative results of initial RT-PCR assays for COVID-19: a systematic review. PLoS One 15 (12), e0242958-e.

Batah, S.S., Fabro, A.T., 2021. Pulmonary pathology of ARDS in COVID-19: a pathological review for clinicians. Respir. Med. 176, 106239.

Corman, V.M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D.K., et al., 2020. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill. 25 (3), 2000045.

Dao, T.L., Hoang, V.T., Gautret, P., 2021. Recurrence of SARS-CoV-2 viral RNA in recovered COVID-19 patients: a narrative review. Eur. J. Clin. Microbiol. Infect. Dis. 40 (1), 13–25.

Islam, K.U., Iqbal, J., 2020. An update on molecular diagnostics for COVID-19. Front. Cell. Infect. Microbiol. 10, 560616.

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J.W., Kim, V.N., Chang, H., 2020. The architecture of SARS-CoV-2 transcriptome. Cell. 181 (4), 914–921 e10.

Kirtipal, N., Bharadwaj, S., Kang, S.G., 2020. From SARS to SARS-CoV-2, insights on structure, pathogenicity and immunity aspects of pandemic human coronaviruses. Infect. Genet. Evol. 85, 104502.

Pan, Y., Zhang, D., Yang, P., Poon, L.L., Wang, Q., 2020. Viral load of SARS-CoV-2 in clinical samples. Lancet Infect. Dis. 20 (4), 411–412.

Pollard, C.A., Morran, M.P., Nestor-Kalinoski, A.L., 2020. The COVID-19 pandemic: a global health crisis. Physiol. Genomics 52 (11), 549–557.

Rahimi, A., Mirzazadeh, A., Tavakolpour, S., 2021. Genetics and genomics of SARS-CoV-2: a review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection. Genomics. 113 (1 Pt 2), 1221–1232.

Resa, C., Magro, S., Marechal, P., Barranger, C., Joannes, M., Miszczak, F., et al., 2014. Development of an efficient qRT-PCR assay for quality control and cellular quantification of respiratory samples. J. Clin. Virol. 60 (3), 270–275.

Sule, W.F., Oluwayelu, D.O., 2020. Real-time RT-PCR for COVID-19 diagnosis: challenges and prospects. Pan Afric. Med. J. 35 (Suppl. 2), 121.

Sungnak, W., Huang, N., Bécavin, C., Berg, M., 2020 May. SARS-CoV-2 entry genes are Most highly expressed in nasal goblet and ciliated cells within human airways. Nat Med 26(5), 681–687.

Thiel, V., Ivanov, K.A., Putics, A., Hertzig, T., Schelle, B., Bayer, S., et al., 2003. Mechanisms and enzymes involved in SARS coronavirus genome expression. J. Gen. Virol. 84 (9), 2305–2315.

Van Vliet, A., Smits, S., Rottier, P., De Groot, R., 2002. Discontinuous and non-discontinuous subgenomic RNA transcription in a nidovirus. EMBO J. 21 (23), 6571–6580.

Wang, W., Xu, Y., Gao, R., Lu, R., Han, K., Wu, G., et al., 2020. Detection of SARS-CoV-2 in different types of clinical specimens. Jama. 323 (18), 1843–1844.

Wiersinga, W.J., Rhodes, A., Cheng, A.C., Peacock, S.J., Prescott, H.C., 2020. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. JAMA. 324 (8), 782–793.

Younes, N., Al-Sadeq, D.W., Al-Jighefee, H., Younes, S., Al-Jamal, O., Daas, H.I., et al., 2020. Challenges in laboratory diagnosis of the novel coronavirus SARS-CoV-2. Viruses. 12 (6), 582.

Zhao, S., Ye, Z., Stanton, R., 2020. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA (New York, NY). 26 (8), 903–909.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., et al., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. nature. 579 (7798), 270–273.

Zimmermann, P., Curtis, N., 2020. COVID-19 in children, pregnancy and neonates: a review of epidemiologic and clinical features. Pediatr. Infect. Dis. J. 39 (6), 469–477.