

Mutations on COVID-19 diagnostic targets

Rui Wang¹, Yuta Hozumi¹, Changchuan Yin²*, and Guo-Wei Wei^{1,3,4} †

¹ Department of Mathematics, Michigan State University, MI 48824, USA

² Department of Mathematics, Statistics, and Computer Science,
University of Illinois at Chicago, Chicago, IL 60607, USA

³ Department of Biochemistry and Molecular Biology
Michigan State University, MI 48824, USA

⁴ Department of Electrical and Computer Engineering
Michigan State University, MI 48824, USA

Abstract

Effective, sensitive, and reliable diagnostic reagents are of paramount importance for combating the ongoing coronavirus disease 2019 (COVID-19) pandemic at the time there is no preventive vaccine nor specific drug available for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It would be an absolute tragedy if currently used diagnostic reagents are undermined in any manner. Based on the genotyping of 7818 SARS-CoV-2 genome samples collected up to May 1, 2020, we reveal that essentially all of the current COVID-19 diagnostic targets have had mutations. We further show that SARS-CoV-2 has the most devastating mutations on the targets of various nucleocapsid (N) gene primers and probes, which have been unfortunately used by countries around the world to diagnose COVID-19. Our findings explain what has seriously gone wrong with a specific diagnostic reagent made in China. To understand whether SARS-CoV-2 genes have mutated unevenly, we have computed the mutation ratio and mutation h -index of all SARS-CoV genes, indicating that the N gene is the most non-conservative gene in the SARS-CoV-2 genome. Our findings enable researchers to target the most conservative SARS-CoV-2 genes and proteins for the design and development of COVID-19 diagnostic reagents, preventive vaccines, and therapeutic medicines.

The coronavirus disease 2019 (COVID-19) pandemic outbreak caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), first reported in Wuhan in December 2019, has spread to 187 countries and territories with more than 3.481 million infection cases and 244,633 fatalities worldwide by May 1, 2020. Additionally, travel restrictions, quarantines, and social distancing measures have essentially put the global economy on hold. Unfortunately, there is no specific medication nor vaccine for COVID-19 at this moment. Therefore, reopening economies depends vitally on effective COVID-19 diagnostic testing, patient isolation, contact tracing, and quarantine. It cannot be overemphasized the importance of diagnostic testing for combating COVID-19.

We reveal that there are many mutations on the COVID-19 diagnostic targets commonly used around the world, including those designated by the United States (US) Centers for Disease Control and Prevention (CDC). These mutations seriously undermine the current global effort in COVID-19 testing, prevention, and control. Approved by the US Food and Drug Administration (FDA), the CDC has detailed guidelines for COVID-19 diagnostic testing, called “CDC 2019–Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Diagnostic Panel” (<https://www.fda.gov/media/134922/download>). The CDC has designated two oligonu-

*Address correspondences to Changchuan Yin. E-mail:cyin1@uic.edu

†Address correspondences to Guo-Wei Wei. E-mail:wei@math.msu.edu

cleotide primers from regions of the virus nucleocapsid (N) gene, i.e., N1 and N2, as probes for the specific detection of SARS-CoV-2. The panel has also selected an additional primer/probe set, the human RNase P gene (RP), as control samples. Many other diagnostic primers and probes based on RNA-dependent RNA polymerase (RdRP, also named IP2, IP4, ORF1ab, or ORF1b), envelope (E), and N genes have been designed [3] and/or designated by the World Health Organization (WHO) as shown in Table S1 of the Supporting Material, which provides the details of 41 commonly used diagnostic primers and probes [10].

Diagnostic test reagents were designed based on early clinical specimens containing a full spectrum of SARS-CoV-2, particularly the reference genome collected on January 5, 2020, in Wuhan (SARS-CoV, NC004718) [12]. It has been reported that different primers and probes show nonuniform performance [1,2,4,6,7,11].

Our findings are based on the genotyping of 7818 SARS-CoV-2 genome samples collected up to May 1, 2020, which have 5117 single mutations over about 29.8 kilobases (kb). These mutations occur on all of SARS-CoV-2 genes and proteins, indicating alarming impacts on the current efforts in the development of COVID-19 diagnostic tests, prevention vaccines, and therapeutic medicines. We employ K -means methods to cluster these mutations, resulting in globally at least five distinct subtypes of SARS-CoV-2 genomes, from early Cluster I to late Cluster V. Table 1 shows cluster distributions of samples (N_{NS}) and total mutation counts (N_{TF}) for 11 countries.

Table 1: The cluster distributions of samples (N_{NS}) and total mutation counts (N_{TF}) for 11 countries.

	Cluster I		Cluster II		Cluster III		Cluster IV		Cluster V	
Country	N_{NS}	N_{TF}	N_{NS}	N_{TF}	N_{NS}	N_{TF}	N_{NS}	N_{TF}	N_{NS}	N_{TF}
US	739	5149	248	1623	514	4968	60	555	677	5035
CA	40	240	13	72	28	193	14	119	19	126
AU	63	434	354	3810	182	1315	99	873	96	691
UA	2	13	554	3785	607	4206	597	5730	60	457
CN	23	54	179	865	7	63	1	13	1	7
DE	0	0	12	42	3	18	8	70	20	131
FR	0	0	14	55	105	755	6	49	66	463
UK	0	0	23	90	10	55	4	30	0	0
IT	0	0	6	134	22	161	12	140	0	0
JP	0	0	67	194	0	0	0	0	0	0
KR	0	0	26	160	0	0	0	0	0	0

The US, Canada (CA), Australia (AU), Ukraine (UA), and China (CN) samples involve all of the five clusters. Among them, China initially had samples only in Clusters I and II and its sample distributions reached to other Clusters after March 2020. Germany (DE) and France (FR) samples are in Cluster II, III, IV, and V. United Kingdom (UK) and Italy (IT) samples are mainly in Clusters II, III, and IV. Japan (JP) and Korea (KR) samples belong to Cluster II only. Cluster II is common to all countries.

Table 2 provides all mutations on various primers and probes and their occurring frequencies in various clusters. More detailed mutation information is given in Tables S2-S42 of the Supporting Material. It is interesting to note that N-China-F [10] is the most inefficient reagent among all primers/probes and its SARS-CoV-2 target has eight mutations involving samples in all five clusters, which may explain many media reports about the inefficiency of certain COVID-19 diagnostic kits made in China. Note that primers and probes typically have a small length of around 20 nucleotides.

Table 2: Summary of mutations on COVID-19 diagnostic primers and probes and their occurrence frequencies in clusters.

Primer/probe	# of mutations	Total frequency	Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V
RX7038-N1 primer (Fw) ^a	4	5	0	1	3	1	0
RX7038-N1 primer (Rv) ^a	6	42	0	28	3	13	0
RX7038-N2 primer (Fw) ^a	2	5	0	1	2	1	1
RX7038-N2 primer (Rv) ^a	3	9	2	5	2	0	0
RX7038-N3 primer (Fw) [6]	5	110	0	98	9	2	1
RX7038-N3 primer (Rv) [6]	6	17	1	3	11	1	1
N1-U.S.-P [10]	3	116	1	108	5	2	0
N2-U.S.-P [10]	3	31	27	3	2	1	0
N3-U.S.-P [10]	8	19	4	5	6	2	2
N-Sarbeco-F ^b [3]	5	15	3	4	6	0	2
N-Sarbeco-P ^b [3]	2	3	0	0	1	2	0
N-Sarbeco-R ^b [3]	7	33	7	6	8	0	12
N-China-F [10]	8	4194	9	76	29	4062	15
N-China-R [10]	7	14	2	1	7	3	1
N-China-P [10]	0	0	0	0	0	0	0
N-HK-F [10]	4	44	0	3	16	25	0
N-HK-R [10]	3	12	2	0	7	2	1
N-JP-F [10]	2	5	3	2	0	0	0
N-JP-R [10]	2	4	0	2	2	0	0
N-TL-F [10]	7	40	1	32	4	3	0
N-TL-R [10]	6	14	0	6	6	1	1
N-TL-P [10]	3	12	0	1	2	9	0
E-Sarbeco-F1 ^c	1	1	0	1	0	0	1
E-Sarbeco-R2 ^c	2	2	1	1	0	0	0
E-Sarbeco-P1 ^c	2	8	0	6	2	0	0
E-DE-F [10]	1	2	0	0	2	0	0
nCoV-IP2-12669Fw ^c	0	0	0	0	0	0	0
nCoV-IP2-12759Rv ^c	7	39	1	13	22	0	3
nCoV-IP2-12696bProbe(+) ^c	1	4	0	0	4	0	0
nCoV-IP4-14059Fw ^c	1	8	0	0	8	0	0
nCoV-IP4-14146Rv ^c	3	13	0	4	5	0	4
nCoV-IP4-14084Probe(+) ^c	3	9	0	4	5	0	0
RdRP-SARSR-F2 ^d	3	13	0	0	11	0	2
RdRP-SARSR-R1 [3] ^d	1	1	0	0	1	0	0
RdRP-SARSR-P2 [3] ^d	2	6	0	5	1	0	0
ORF1ab-China-F [10]	0	0	0	0	0	0	0
ORF1ab-China-R [10]	0	0	0	0	0	0	0
ORF1ab-China-P [10]	0	0	0	0	0	0	0
ORF1b-nsp14-HK-F [10]	2	2	0	0	1	0	1
ORF1b-nsp14-HK-R [10]	4	9	0	6	2	2	0
ORF1b-nsp14-HK-P [10]	2	4	1	0	2	1	0

^a <https://www.fda.gov/media/136691/download>

^b <https://www.eurosurveillance.org/content/table/10.2807/1560-7917.ES.2020.25.3.2000045.t1?fmt=ahah&fullscreen=true>

^c <https://www.who.int/docs/default-source/coronavirus/real-time-rt-pcr-assays-for-the-detection-of-sars-cov-2-institu>

^d https://www.who.int/docs/default-source/coronavirus/protocol-v2-1.pdf?sfvrsn=a9ef618c_2

Currently, all the primers and probes used in the US target the N gene [10]. Unfortunately, Table 2 shows

that all of the US CDC designated COVID-19 diagnostic primers have been compromised. The targets of N gene primers and probes used in Japan, Thailand, and China, including Hong Kong, except for that of N-China-P, have undergone multiple mutations involving many clusters as well.

It is interesting to note that the targets of four E gene primers and probes have only six mutations. No mutation has been found on the targets of RNA-dependent RNA polymerase-based primers or probes, nCoV-IP2-12669Fw primer, ORF1ab-China-F, ORF1ab-China-R, and ORF1ab-China-P. However, the target of nCoV-IP2-12759R recommended by Institut Pasteur, Paris has 7 mutations. Overall, targets of the envelope and RNA-dependent RNA polymerase based primers and probes have fewer mutations than those of the N gene. This observation leads us to wonder whether the N gene is particularly prone to mutations.

Table 3: Gene-specific statistics of SARS-CoV-2 single mutations.

Gene type	Gene site	Gene length	Unique SNPs	mutation ratio	h -index
NSP1	266:805	540	121	0.2241	8
NSP2	806:2719	1914	407	0.2126	16
NSP3	2720:8554	5835	912	0.1563	18
NSP4	8555:10054	1500	203	0.1353	11
NSP5(3CL)	10055:10972	918	130	0.1416	10
NSP6	10973:11842	870	133	0.1529	8
NSP7	11843:12091	249	37	0.1486	5
NSP8	12092:12685	594	77	0.1296	4
NSP9	12686:13024	339	48	0.1416	6
NSP10	13025:13441	417	44	0.1055	4
NSP11	13442:13480	39	5	0.1282	2
RNA-dependent-polymerase	13442:16236	2796	363	0.1298	15
Helicase	16237:18039	1803	227	0.1259	12
3'-to-5' exonuclease	18040:19620	1581	241	0.1524	10
endoRNase	19621:20658	1038	143	0.1378	10
2'-O-ribose methyltransferase	20659:21552	894	115	0.1286	8
Spike protein	21563:25384	3819	622	0.1629	17
ORF3a protein	25393:26220	825	231	0.28	13
Envelope protein	26245:26472	225	30	0.1333	6
Membrane glycoprotein	26523:27191	666	105	0.1577	11
ORF6 protein	27202:27387	183	47	0.2568	6
ORF7a protein	27394:27759	363	88	0.2424	6
ORF7b protein	27756:27887	129	10	0.0775	2
ORF8 protein	27894:28259	363	90	0.2479	8
Nucleocapsid protein	28274:29533	1257	340	0.2705	29
ORF10 protein	29558:29674	114	27	0.2368	4

To understand whether there is a differentiation in SARS-CoV-2 gene mutation pattern, we analyze the gene-specific statistics of SARS-CoV-2 single mutations. Table 3 lists the mutation ratio, i.e., number of unique single-nucleotide polymorphisms (SNPs) over the corresponding gene length, for all SARS-CoV-2 genes. A smaller mutation ratio for a given gene indicates its higher degree of conservativeness. Clearly, ORF7b gene has the smallest mutation ratio of 0.0775. The N gene has the second largest mutation ratio of 0.2705, which is very close to the largest ratio of 0.2800 for ORF3a gene. To take into the consideration of mutation frequency, we introduce the mutation h -index, defined as the maximum value of h such that the given gene section has h single mutations that have each occurred at least h times. Normally, larger genes tend to have higher h -index. Table 3 shows that, with a moderate length, the N gene has the largest h -index of 29, which is significantly higher the second largest h -index of 18 for NSP3. Therefore, it was truly

unfortunate for the world to have selected SARS-CoV-2 N gene primers and probes as diagnostic reagents for combating COVID-19.

In summary, the targets of currently used COVID-19 diagnostic reagents have had numerous mutations that have seriously undermined our ability to combat COVID-19. In the Supporting Material, we provide a full list of all 5117 SNP variants, including their positions and mutation types. This information, together with ranking of the degree of the conservativeness of SARS-CoV-2 genes or proteins given in Table 3, enables researchers to avoid non-conservative genes (or their proteins) and mutated nucleotide segments in designing COVID-19 diagnosis, vaccine and drugs.

Methods and materials SARS-CoV-2 genome sequences from infected individuals dated between January 5, 2020, and May 1, 2020, are downloaded from the GISAID database [8] (<https://www.gisaid.org/>). We only consider the records in GISAID with complete genomes and submission dates. The resulting 7818 complete genome sequences are rearranged according to the reference SARS-CoV-2 genome [12] by using the Clustal Omega multiple sequence alignment with default parameters [9]. Gene variants are recorded as single-nucleotide polymorphisms (SNPs). The Jaccard distance [5] is employed to compute the similarities among genome samples. The resulting distance matrix is used in the k -means clustering of all samples.

1 Data Availability

The nucleotide sequences of the SARS-CoV-2 genomes used in this analysis are available, upon free registration, from the GISAID database (<https://www.gisaid.org/>). Supporting Material presents a list of 5117 SNP variants of 7818 SARS-CoV-2 samples across the world, a list of 41 commonly used diagnostic primers and probes, and tables of mutation details on 41 diagnostic primers and probes. The acknowledgments of the SARS-COV-2 genomes are also given in the Supporting Material.

Acknowledgment

This work was supported in part by NIH grant GM126189, NSF Grants DMS-1721024, DMS-1761320, and IIS1900473, Michigan Economic Development Corporation, Bristol-Myers Squibb, and Pfizer. The authors thank The IBM TJ Watson Research Center, The COVID-19 High Performance Computing Consortium, and NVIDIA for computational assistance.

References

- [1] A. M. Casto, M.-L. Huang, A. Nalla, G. A. Perchetti, R. Sampoleo, L. Shrestha, Y. Wei, H. Zhu, A. L. Greninger, and K. R. Jerome. Comparative performance of SARS-CoV-2 detection assays using seven different primer/probe sets and one assay kit. *medRxiv*, 2020.
- [2] J. F.-W. Chan, C. C.-Y. Yip, K. K.-W. To, T. H.-C. Tang, S. C.-Y. Wong, K.-H. Leung, A. Y.-F. Fung, A. C.-K. Ng, Z. Zou, H.-W. Tsoi, et al. Improved molecular diagnosis of COVID-19 by the novel, highly sensitive and specific COVID-19-rdrp/hel real-time reverse transcription-pcr assay validated in vitro and with clinical specimens. *Journal of Clinical Microbiology*, 58(5), 2020.
- [3] V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance*, 25(3):2000045, 2020.

- [4] Y. J. Jung, G.-S. Park, J. H. Moon, K. Ku, S.-H. Beak, S. Kim, E. C. Park, D. Park, J.-H. Lee, C. W. Byeon, et al. Comparative analysis of primer-probe sets for the laboratory confirmation of SARS-CoV-2. *BioRxiv*, 2020.
- [5] M. Levandowsky and D. Winter. Distance between sets. *Nature*, 234(5323):34–35, 1971.
- [6] A. K. Nalla, A. M. Casto, M.-L. W. Huang, G. A. Perchetti, R. Sampoleo, L. Shrestha, Y. Wei, H. Zhu, K. R. Jerome, and A. L. Greninger. Comparative performance of SARS-CoV-2 detection assays using seven different primer/probe sets and one assay kit. *Journal of Clinical Microbiology*, 2020.
- [7] S. Pfefferle, S. Reucher, D. Nörz, and M. Lütgehetmann. Evaluation of a quantitative rt-pcr assay for the detection of the emerging coronavirus SARS-CoV-2 using a high throughput system. *Eurosurveillance*, 25(9):2000152, 2020.
- [8] Y. Shu and J. McCauley. GISAID: Global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance*, 22(13), 2017.
- [9] F. Sievers and D. G. Higgins. Clustal omega. *Current protocols in bioinformatics*, 48(1):3–13, 2014.
- [10] B. Udugama, P. Kadhiresan, H. N. Kozlowski, A. Malekjahani, M. Osborne, V. Y. Li, H. Chen, S. Mubareka, J. Gubbay, and W. C. Chan. Diagnosing COVID-19: The disease and tools for detection. *ACS nano*, 2020.
- [11] C. B. Vogels, A. F. Brito, A. L. Wyllie, J. R. Fauver, I. M. Ott, C. C. Kalinich, M. E. Petrone, M.-L. Landry, E. F. Foxman, and N. D. Grubaugh. Analytical sensitivity and efficiency comparisons of SARS-CoV-2 qrt-pcr assays. *medRxiv*, 2020.
- [12] 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.