

Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active. Contents lists available at ScienceDirect



International Journal of Infectious Diseases



journal homepage: www.elsevier.com/locate/ijid

Short Communication

Analysis of the potential impact of genomic variants in global SARS-CoV-2 genomes on molecular diagnostic assays



Abhinav Jain^{a,b,1}, Mercy Rophina^{a,b,1}, Saurabh Mahajan^c, Bhavya Balaji Krishnan^d, Manasa Sharma^e, Sreya Mandal^c, Teresa Fernandez^c, Sumayra Sultanji^c, Bani Jolly^{a,b}, Samatha Mathew^{a,b}, Sridhar Sivasubbu^{a,b}, Vinod Scaria^{a,b,*}

^a CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Mathura Road, Delhi 110025, India

^b Academy of Scientific and Innovative Research (AcSIR), CSIR-HRDC Campus, Sector 19, Kamla Nehru Nagar, Ghaziabad, Uttar Pradesh 201002, India

^c St. Joseph's College, Langford Gardens, Bengaluru, Karnataka 560027 India

^d Imperial College London, South Kensington, London SW7 2BU, United Kingdom

^e Ramaiah University of Applied Sciences, Bengaluru, Karnataka 560054, India

ARTICLE INFO

ABSTRACT

Article history: Received 5 August 2020 Received in revised form 25 October 2020 Accepted 26 October 2020

Keywords: COVID-19 Genomes SARS-CoV-2 Variations Reverse transcription polymerase chain reaction Gibbs free energy An epidemic of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causing coronavirus diseases (COVID-19) initially reported in Wuhan, China has rapidly emerged into a global pandemic affecting millions of people worldwide. Molecular detection of SARS-CoV-2 using reverse transcription polymerase chain reaction (RT-PCR) forms the mainstay in screening, diagnosis and epidemiology of the disease. Since the virus evolves by accumulating base substitutions, mutations in the viral genome could possibly affect the accuracy of RT-PCR-based detection assays. The recent availability of genomes of SARS-CoV-2 isolates motivated us to assess the presence and potential impact of variations in target sites of the oligonucleotide primers and probes used in molecular diagnosis. We catalogued a total of 132 primer or probe sequences from literature and data available in the public domain. Our analysis revealed that a total of 5862 unique genetic variants mapped to at least one of the 132 primer or probe binding sites in the genome. A total of 29 unique variants were present in \geq 1% of genomes from at least one of the continents (Asia, Africa, Australia, Europe, North America, and South America) that mapped to 36 unique primers or probes binding sites. Similarly, a total of 27 primer or probe binding sites had cumulative variants frequency of > 1% in the global SARS-CoV-2 genomes. These included primers or probes sites which are used worldwide for molecular diagnosis as well as approved by national and international agencies. We also found 286 SARS-CoV-2 genomic regions with low variability at a continuous stretch of > 20bps that could be potentially used for primer designing. This highlights the need for sequencing genomes of emerging pathogens to enable evidence-based policies for development and approval of diagnostics. © 2020 The Authors. Published by Elsevier Ltd on behalf of International Society for Infectious Diseases. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-ncnd/4.0/).

- SARS-CoV-2 variants impact RT-PCR efficiency in detection.
- A total of 29 global SARS-CoV-2 genetic variants had a frequency ≥ 1%.
- The thermodynamic stability of the virus-primers complex gets perturbed.
- A number of recommended primer or probe sequences had high variant frequency.

Initially reported from a city in China, the coronavirus disease 2019 (COVID-19) has now rapidly emerged as a global pandemic. Reverse transcription polymerase chain reaction (RT-PCR) based assays have been the mainstay for the diagnosis and screening of COVID-19 due to their high sensitivity and specificity (Shen et al. 2020). These assays utilize oligonucleotide primers and probes specific to the viral nucleic acid. The SARS-CoV-2 has been continuously evolving and has an estimated substitution rate of $1.19-1.31 \times 10^{-3}$ per site per year (Li et al. 2020). Recent reports that suggest genetic variation in viruses at the primers or probes binding site could decrease the sensitivity of RT-PCR based assays (Yang et al. 2014). Motivated by the availability of a large number of genomes of SARS-CoV-2 isolates globally, we attempted to

https://doi.org/10.1016/j.ijid.2020.10.086

^{*} Corresponding author at: CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Mathura Road, Delhi 110025, India. *E-mail address*: vinods@igib.in (V. Scaria).

¹ Contributed equally and would like to be known as joint first authors.

^{1201-9712/© 2020} The Authors. Published by Elsevier Ltd on behalf of International Society for Infectious Diseases. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

	imary of Primer and Probe sequences and genomic v	ariants analysis.
	088075- 4/1 56 28601) 7P	107362-000 ATT0506202 TTTG 19 /
Suppose E.G. Al Al model Science (politic) 0 <td< td=""><td>0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0</td><td>Q8720- 39350/# 44 8 6 9 9 0</td></td<>	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0	Q8720- 39350/# 44 8 6 9 9 0
	28 02040201 4 53 T0AMTACIC - 0 CAAAA 28 52 7 1	
Trostocor a 55 o 0T100C 24 3 0 cmeasured of the senses, and details of releases and details of the senses, conclusions and another senses, and details of the senses, conclusions and another senses, and details of the sense, and details of the sense, and details of the senses	201100 /PP TC 4 0 N 2020 0001 0000000000000000000000000	UIC 0 7 287 19 7 28756 47 CT 2 7 N 45 0000 28756 47 CT 2 7 N 45 0000 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	- 1 0 1 3 8 0 2 2 8 1 9 0 1 1 1 0 1 1 3 8 0 2 2 8 1 1 9 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	2 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
C014/C0170 2 33 0. C014/C0170 10 8 0. 0. C014/C0170 26 8 10 10 C014/C0170 26 8 11 17 C014/C0170 26 8 11 17 C014/C0170 15 8 8 10 C014/C0170 15 8 8 10 C014/C0170 15 8 8 10 C014/C0170 16 10 10 10	AC 227 19 / Nut_internation 207030 PP CT 1 7 N 40 00009 apd 3 (Memorital WT) 207030 PP CT 1 7 N 40 0.0009 apd 3 (Memorital WT)	GTTGTAGCA 22 (1 / 0 / 1 GTTGTAGCA 22 (1 / 0 / 1 GTTGTAGCA 24 (1 / 0 /
	2 1 1 3 6 3 1 2 0	
1000TTAC 1/2 3/8 0 1000TTAC 1000TTAC <td>TCCORC air assessment of the second by a delined. 2883 (A 2 3 51 2</td> <td>0 AGGUACG 4 5 5 8 CTTG 28 5 11 / 2003018 CT 5 8 8 11 / 2003018 CT 6 1</td>	TCCORC air assessment of the second by a delined. 2883 (A 2 3 51 2	0 AGGUACG 4 5 5 8 CTTG 28 5 11 / 2003018 CT 5 8 8 11 / 2003018 CT 6 1
AMCTOCADA 27 53 0 DTBACA 10 1/ 1 P 19500 - 10 1/ 1 P 19501 - 01 1/	CALGOCITIT 41 51 TOTOLOGICOL 2003	CTTOCOLACT 4 50 ACCIDATIGA 51.0 1 CREAD 32.0 9 51.1 CREAD 32.0 9 51.1 CREAD 9 8.14 4.1 COUND 2.0000 accidationac
TODOTION 24 0 0 0 ATRODUCOL 15 0 1 T1000 0 1 0 T1000 0 1 0 NUMBER 0 0 No.2000	GAGGAACC 2 53 AGAAGAGG 1 9 3 (7170 281 9 1 9 1 29533) PP C 3 0 8 1 N 425 0.000 B0000000000000000000000000000000	OTTOGOACT (
A COTOCTOC 1071 A 60 1071 A 70 1071 A 70		28456 / / 49 2855) / P 2 7 2 0 9 0 0 0 0 0 0 0 0 0
	2864918P C.T. 3 8 N 194 0002 with 138ergef /#TT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	- 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
01000-0001 2 . 3 8 The 22 1 / / 5 The 22 1 / 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 224479 89 07 20 7 5 506 0.011 additional discrete the 2 22459 89 07 20 7 5 506 0.011 additional discrete the 2 2 2 2 2 2 2 2 2 2 2 2 2 2		00004ACTT 1 9.4 CTOURT, AS 24.2 4 CTOURT, AS 24.2 4 CHARLES, B16, 3 5 0.004 SHR01PP GA, 2 5
	299951 RP C T 3 7 N 596 00122 and 5118er007.0011	4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
COTIGNAM D0 TMANTOAT 271 TOTOCOTT 277 COTOC 3 COTO 3	TotAction 1 0	00000ACTT 1 54 CTCCUCCT 1 58 CTCCUCCT 1 1000.76444. Views 2 5 JUBRIT 0 6 JERDIT 0 0.0002 JERDIT 0 0.0002
	0 0	1 0 1 0 2 1 4 0 0 4 0 1 1 4 0
COLUMPOC 12 0 1 1 0 1 1 0 1 1 0 1 1 0 1	Control C.1 A N 1923 0.942 additionage and comparison matrix C00001 C 1 N 1923 0.942 additionage and comparison matrix C00001 C 1 1 N 1923 0.942 additionage and comparison matrix C00001 C 2 1<	CHAPT 0 (1,1,7) N 13891 0.3031 Bases ad/Christophalbias 2 28996 > 0 5 <td< td=""></td<>
COGATOOC - 13 4 1 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1 1 0	
Th 10010 (2000) 201 0 0 assessment of the specificity of address (2000) assessment of the specificity of address (2000) assessment of the specificity of the specificity of the specificity of the specificity of (2000) assessment of	000GAACTT 1 54 CTCDAT 1 58 CTCDAT 1 38 20 / B CTCDAT 288 20 / B CTCDAT 1 58 20 / B CTCDAT	24 CAACCTTC 1 55 06CAACG 1 2 3 To 28 0 5/ / the listoring plates
	2000/1/2 CU 0 0 0 10 10 100 0 0 0 0 0 0 0 0 0 0 0	
25 1011001LC 1/ 51 X1040A4C 20 00 / 0050 4 1 49 11110100000440000	24 3 CADACATTT 6 3 TOCTCTCAA 28 1 0 MERLINERS MOLESCOMMUNICATION MERLINERS MOLESCOMMUNICATION 10 CERNO 7 5 5	
279491979 CT 3 7 M 502 0010 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	20079107 CT 2 1 N 714 0.0000 banka 2 20179107 CT 2 1 N 714 0.0000 banka 2 2 1 0 8 4 0 3 2 0 3 4 0	OpenA Official (C) OpenA
GALAGOSTT 22 COTTACTOR 1.9 COTACTOR	TCTODITAAL 389 24 60 BBLE Towns Bronk costs:come GGCCAACA 77 97 7 8 5 GGCCAACA 77 97 7 N 214 0.0046 BLE Towns Bronk costs:come	
	TIALGAACGA 67 110000000 11 1 1 110000000 201 19 7 10100000 201 19 7 10100 10 10 10 10 10 10 10 10 10 10 10 1	CRUID: 27 16 48 A
ADCOUNT 10 10 0 1 TODACC C 10 10 0 1 TODACC C 10 10 1 0 0 TODACC C 10 1 0 0 0 TODACC C 10 0 0 0 TODACC C 10 0 TODAC	OF SMPP O I I N O </td <td>- 2 2 1 9 1 0 3 4 7 9 1 0 0 77 1 ATOPO2000 6 5 3 ATOPO2000 6 5 3 ATOPO2000 6 5 3</td>	- 2 2 1 9 1 0 3 4 7 9 1 0 0 77 1 ATOPO2000 6 5 3 ATOPO2000 6 5 3 ATOPO2000 6 5 3
	OCAMATTOT 21 49 GCANTTOC 1 7 00 201 / 7 00 201 / 7 00 201 / 7 01 201 / 7 02 201 / 7 03 201 / 7 04 0005 2005 2005	GA 292 12 1 Macroscopensionshifts 2022 7.9 65 0.0007 202022.2 20241) PP 0.7 1 8 N 305 0.0007 20241) PP 0.7 1 8 N 305 0.0007 200722.2 2041) PP 0.7 1 8 N 305 0.0007 200724.0 2 0 <t< td=""></t<>
25 5 COTTODTD 91 53 Recotacionententicionententico		2 2 1 9 1 0 2 4 9 1 0 2 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
ensempter AQ 3 B 19 0.0013 Lad?bhome422106_4 - <	COOFTCTTC 7 80 GOARTATICS 202 20 / CONTACTCS 202 20 / CONTACTCS 202 20 /	20223 27 12 54 and control of the state of t
30 31 100007100 1 3 1000007100 1 3 AVTC10 207 7 AVTC10 207 7 AVTC10 4 107 AVTC10 4 107	Observer 0.1 9 0.1 0.0007 0.200 2 2 1 0	2 2 2 1 9 1 0 3 4 7 9 1 0 0
Open Open <th< td=""><td></td><td>201417/RP 0.1 0.1 0.1 0.0 0.0 0.0 0.0 0.0 0.0 0.0</td></th<>		201417/RP 0.1 0.1 0.1 0.0 0.0 0.0 0.0 0.0 0.0 0.0
GGGAGCCT 88: 25 8 TGA T.C. 8 / N 282 0.0061 /		

The variant frequency, primers/probes frequency, Gibbs free energy (Δ G), and melting temperature (Tm) for reference and alternate allele in the Indian SARS-CoV-2 isolates. Only primers/probes with a frequency of more than 1% in any of the six continents are included in this Table.Tm-Melting Temperature Temperature, Δ G-Gibbs Free Energy, Ref- Reference, Alt- Alternate, No.- Number, Afr- Africa, Aus- Australia, Eur- Europe, NA-North America, SA- South America.



Figure 1. Genome Wide distribution (A) and frequency (B) of the 29 genetic variants with allele Frequency $\geq 1\%$ in at least one of 6 continents. Note: Colour required in Print.

understand the genomic variants and their potential impact on molecular diagnostic assays.

We analysed the genome sequences of SARS-CoV-2 isolates deposited in GISAID (Shu and McCauley 2017) as on 26th September 2020. Only complete genome sequences that had ≥99% alignment with the Wuhan-Hu-1 reference genome (NC_045512.2) (Wu et al. 2020) and <5% degenerate bases were considered for the analysis. Genome sequences having clustered mutations and higher than expected divergence were also excluded from the analysis. The individual genomes were realigned to the Wuhan-Hu-1 reference genome using EMBOSS needle (Rice et al. 2000) and the pairwise alignments were parsed to identify variants using bespoke scripts. The primer/probe sequences were compiled using extensive literature searches as well as from public databases and were mapped to the reference genome using BLAST (Altschul et al. 1990). The SARS-CoV-2 genomic variant coordinates were overlapped with the primer/ probe binding sites. Melting temperature (Tm) and Gibbs free energy (ΔG°_{37}) at standard condition were calculated for primer or probe sequences. We also evaluated the internal and terminal mismatches that could have an impact on the thermodynamic stability of the nucleic acid secondary structure as well as on the Tm Supplementary Method 1. We have also identified regions in the SARS-CoV-2 genome with low variability. We have considered variants below 95th percentile of the frequency and identified continuous stretches of \geq 20bps for designing primers. The choice of 20 bps is guided by the standard length of the primer/probe sequences.

A total of 45,830 high quality genome sequences which comprise 4779 sequences form Asia, 25,091 sequences from Europe, 859 from Africa, 12,949 sequences from North America, 791 sequences from South America and 1361 from Australia were used in the analysis. Our analysis revealed a total of 88,880 unique single nucleotide variants (SNVs) across the genome. We compiled a total of 132 primers or probe sequences Supplementary Data 1. A total of 5862 unique genetic variants mapped to 132 primers or probes binding sites in the SARS-CoV-2 genome. Out of these, a total of 29 unique variants had allele frequency $\geq 1\%$ in at least one

of the six continents from where the SARS-CoV-2 genomes were isolated Table 1 and Figure 1. We have also observed potential differences in the ΔG°_{37} and Tm, that affect the thermodynamic stability of secondary structure and the annealing of the primers and probes to the viral cDNA/RNA respectively Table 1. Of significant interest, three variants with over 30% frequency each genomes mapped to the primer 2019-nCoV-NFP in GGGGAACTTCTCCTGCTAGAAT that targets N gene which is a part of the China Centers for Disease Control and Prevention (CDC) protocol (WHO in house assay, 2020). A cumulative variant frequency of 93.5% was found in 2019-nCoV-NFP binding site. Variants with >1% frequency were also found in primer / probes encompassing S, M, ORF1ab, and ORF3a genes Table 1. A total of 27 primers and probes sequences had cumulative variant frequency >1% of which 11 were approved by the national regulatory bodies mainly by the Centers for Disease Control and Prevention (CDC) and World Health Organization (WHO) and has been widely used across the globe Supplementary Data 2. Our analysis also suggests 286 genomic regions/sites with variants frequency below 95th percentile (corresponding to variant frequency of 1.7×10^{-4}) Supplementary Figure 1 and Supplementary Data 3.

Our analysis suggests that genome sequencing of isolates in an epidemic could provide useful insights into assessing the diagnostic efficacies as also suggested by previous authors (Khan and Cheung, 2020). We surmise that this could possibly drive policies on evaluation and approvals of the assays for screening and diagnosis. The study also highlights the need for rapid and wide-spread sharing of genomic data of pathogens as well as molecular probe information through public archives during pandemics.

Author contributions

MR performed the genome analysis and variant calls. BJ performed the quality assessment of the genome dataset. AJ and SM1 co-ordinated the compendium of primers and probes with help of Bhavya Balaji Krishnan, Manasa Sharma, Sreya Mandal, Teresa Fernandez and Sumayra Sultanji. SM2 contributed to