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

Abhinav Jain[1,2,&], Mercy Rophina[1,2,&], Saurabh Mahajan[3], Bhavya Balaji Krishnan[4], Manasa Sharma[5], Sreya Mandal[3], Teresa Fernandez[3], Sumayra Sultanji[3], Samatha Mathew[1,2], Sridhar Sivasubbu[1,2], Vinod Scaria[1,2,$]

[1] *CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Mathura Road, Delhi 110025, INDIA*
[2] *Academy of Scientific and Innovative Research (AcSIR), CSIR-HRDC Campus, Sector 19, Kamla Nehru Nagar, Ghaziabad, Uttar Pradesh 201002, INDIA*
[3] *St. Joseph's College, Langford Gardens, Bengaluru, Karnataka 560027 INDIA*
[4] *Imperial College London, South Kensington, London SW7 2BU, United Kingdom*
[5] *Ramaiah University of Applied Sciences, Bengaluru, Karnataka 560054 INDIA*

[&] *Contributed equally and would like to be known as joint first authors.*
[$] *Address for correspondence: CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Mathura Road, Delhi 110025, INDIA . Email: vinods@igib.in (VS)*

ABSTRACT

An isolated epidemic of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) causing Coronavirus Diseases (COVID-19) originating in Wuhan, China has now 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 disease. The virus has been evolving through base substitutions. The recent availability of genomes of SARS-CoV-2 isolates from different countries including India motivated us to assess the presence and potential impact of variations in target sites for the oligonucleotide primers and probes used in molecular diagnosis. We catalogued a total of 132 primers or probes sequences from the literature and the public domain. Our analysis revealed a total of 125 unique genetic variants in 80 either primers or probes binding sites. A total of 13 unique variants had allele frequency of ≥ 1% in Indian SARS-CoV-2 genomes mapped to the primers or probes binding sites. A total of 15 primers or probes binding sites had cumulative variant frequency of ≥ 1% in the SARS-CoV-2 genomes. These included primers or probes sites which are widely used in India and across the world for molecular diagnosis as well as approved by national and international agencies. This highlights the need for sequencing genomes of emerging pathogens to make evidence based policies for development and approval of diagnostics. To the best of our knowledge, ours is the most comprehensive analysis of genomic variants in genomes of SARS-CoV-2 isolates from India and their potential impact on efficacy of molecular diagnostics.

Keywords: COVID-19, genomes, SARS-CoV-2, variations, reverse transcription polymerase chain reaction, Gibbs free energy

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 the 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 to 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 its sensitivity (Yang et al. 2014). Motivated by the availability of a large number of genomes of SARS-CoV-2 isolates from India, we attempted to understand the genomic variants and their potential impact on molecular assays.

We analysed genomic sequences of SARS-CoV-2 isolates from India in GISAID (Shu and McCauley 2017) as on 23rd of June 2020. Sequences with <99% alignment and ≥1% gaps were not considered for analysis. The genomes were re-aligned to the reference SARS-CoV-2 genome Wuhan-Hu-1 (Wu et al. 2020) using EMBOSS needle (Rice et al. 2000) and parsed for variants using bespoke scripts. The primer/probe sequences were compiled using extensive literature searches as well as databases (COVID-19 Primer, 2020) and were mapped to the reference genome using BLAST (Rice et al. 2000). The SARS-CoV-2 genomic variant coordinates were overlapped with the primer/probe binding sites. Tm and Gibbs free energy (ΔG) was calculated**.** We also evaluated the internal single mismatch as well as terminal mismatch which could have an impact on the thermodynamics stability of the nucleic acid secondary structure as well as on Tm **Supplementary Methods 1**.

Of the 938 genomes of SARS-CoV-2 isolates from India, a total of 717 were of high quality and were further considered for the variant calling **Supplementary Data 1**. This analysis revealed a total of 1,523 single nucleotide variants (SNVs) as well as 27 indels. We could compile a total of 132 primers or probe sequences **Supplementary Data 2**. A total of 123 SNVs and 2 indels mapped to at least one of the 80 odd primer/probe sites in the genome **Supplementary Data 3** of which, a total of 13 unique variants had allele frequency ≥ 1% **Table 1 and Figure 1**. Of significant note were three primers/probes for the *N* gene, which has variants mapping to the target sites in over 10% of Indian isolates. Variants with >1% frequency were also found in primer / probes encompassing *S, E, RdRP, ORF1a*, and *ORF3a* genes **Table 1**. One of the variant, 15451:G:A had a high frequency of 0.6% in Indian isolates and mapping to the WHO and ICMR-India recommended RdRP_SARSr-F2 “GTGARATGGTCATGTGTGGCGG” primer. Additionally, two indels with 3’ end terminal mismatch were in (N=6) 0.8% in Indian SARS-CoV-2 genomes . One of the primers is a part of WHO protocol E_Sarbeco_R2 (Corman et al. 2020) and widely used (Eurofins Genomics, 2020). Out of these indels, one involves a 3 nt deletion from 3’ end of the primers while the other indel deletes 41 nt that encompasses the whole E_Sarbeco_R2 primer.

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. It has not escaped our attention that a number of genomic loci had significantly low variability in Indian isolates of SARS-CoV-2 isolates suggesting an opportunity to develop better molecular assays. To the best of our knowledge, this report is the most comprehensive report of the assessment of genomic variants and their impact on molecular assays for Indian isolates of SARS-CoV-2. The study highlights the need to widely share genome sequences of isolates as well as molecular probe information during epidemics.

ACKNOWLEDGEMENTS

We acknowledge the researchers who have made the SARS-CoV-2 genomes available in the public domain. A comprehensive list of genomes, contributing laboratories, and acknowledgement is available in **Supplementary Data 1.** Authors acknowledge Paras Sehgal for constructive comments which enriched the manuscript.

Authors acknowledge funding from CSIR India. AJ and SM acknowledge a research fellowship from CSIR India. The funders had no role in the preparation of the manuscript or decision to publish.

AUTHOR CONTRIBUTIONS

MR performed the genome analysis and variant calls. 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 mapping the primers to the genomic loci. AJ performed the analysis of variants mapping to the probe-target sites and was assisted by MR. VS and SS provided the conceptual overview to the analysis. VS, MR and AJ wrote the manuscript, the content and analysis which was read and agreed upon by all authors.

REFERENCES

Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill [Internet]. 2020 Jan;25(3). doi: 10.2807/1560-7917.ES.2020.25.3.2000045

Khan KA, Cheung P. Presence of mismatches between diagnostic PCR assays and coronavirus SARS-CoV-2 genome. R Soc open sci [Internet]. 2020 Jun 10;7(6):200636. Available from: https://royalsocietypublishing.org/doi/10.1098/rsos.200636

Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT, et al. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. J Med Virol [Internet]. 2020 Jun;92(6):602–11. doi: 10.1002/jmv.25731

Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet [Internet]. 2000 Jun;16(6):276–7. doi: 10.1016/s0168-9525(00)02024-2

Shen M, Zhou Y, Ye J, Abdullah Al-Maskri AA, Kang Y, Zeng S, et al. Recent advances and perspectives of nucleic acid detection for coronavirus. J Pharm Anal [Internet]. 2020 Mar 1; doi: 10.1016/j.jpha.2020.02.010

Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill [Internet]. 2017 Mar 30;22(13). doi: 10.2807/1560-7917.ES.2017.22.13.30494

Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature [Internet]. 2020 Mar;579(7798):265–9. doi: 10.1038/s41586-020-2008-3

Yang J-R, Kuo C-Y, Huang H-Y, Wu F-T, Huang Y-L, Cheng C-Y, et al. Newly emerging mutations in the matrix genes of the human influenza A(H1N1)pdm09 and A(H3N2) viruses reduce the detection sensitivity of real-time reverse transcription-PCR. J Clin Microbiol [Internet]. 2014 Jan;52(1):76–82. doi: 10.1128/JCM.02467-13

Website References:

COVID-19 Primer, 2020 Summaries and Discover trends in the latest research papers and the conversations around them https://covid19primer.com/ last accessed on 23 May 2020

Eurofins 2020, https://www.eurofinsgenomics.com/en/products/dnarna-synthesis/coronavirus-portfolio/ last accessed on 3 July 2020

TABLES

Primer/Probe Sequence / Description	Genomic Variant	ΔG (Ref / Alt)	Tm (Ref / Alt)	Gene	No of genomes with variant	Allele frequency	Cumulative No of variants and Freq of variants
GGGGAACTTCTCCTGCTAGAAT (28881-28902) FP	28881:G:A*	-26.1 / -24.82	54.8 / 53	N	93	0.13	283 (0.395)
	28881:G:R	-26.1 / -26.1 to -24.82	54.8 / 53 to 54.8		1	0.001	
	28882:G:A*	-26.1 / -22.66	54.8 / 53		93	0.13	
	28882:G:R	-26.1 / -26.1 to -22.66	54.8 / 53 to 54.8		1	0.001	
	28883:G:C*	-26.1 / -23.64	54.8 / 54.8		93	0.13	
	28883:G:R (A or G)*	-26.1 / -26.1 to -22.66	54.8 / 53 to 54.8		1	0.13	
	28890:C:G	-26.1 / -21.14	54.8 / 54.8		1	0.001	
ACCCCGCATTACGTTTGGTGGACC (28309-28332) Probe	28311:C:T*	-32.95 / -27.67	58.8 / 57.1	N	191	0.266	214 (0.298)
	28311:C:Y*	-32.95 / -32.95 to -27.67	58.8 / 57.1 to 58.8		9	0.013	
	28312:C:T	-32.95 / -27.67	58.8 / 57.1		5	0.007	
	28326:G:T*	-32.95 / -30.05	58.8 / 57.1		9	0.013	
TGAACTGTTGCGACTACGTG (28836-28855) RP	28845:G:T	-25.5 / -20.3	51.8 / 49.7	N	2	0.003	96 (0.134)
	28851:G:T	-25.5 / -21.12	51.8 / 49.7		1	0.001	
	28854:C:T *	-25.5 / -22.09	51.8 / 49.7		93	0.13	

AGGGTCAAGTGCACAGTCTA (22428-22447) RP	22444:C:T*	-23 / -20.25	51.8 / 49.7	S	88	0.123	88 (0.123)
TGGTTTAGCCAGCGTGGTGGT (8774-8794) Probe	8782:C:TG:A*	-28.72 / -23.36	56.3 / 54.4	ORF1a	35	0.049	35 (0.049)
TGCAACTGAGGGAGCCTTGA (28672-28691) FP	28674:C:TG:A	-26.27 / -21.19	53.8 / 51.8	N	1	0.001	28 (0.039)
	28676:A:M (A or C)	-26.27 / -26.27 to -24.21	53.8 / 53.8 to 55.9		2	0.003	
	28688:T:CA:G*	-26.27 / -23.6	53.8 / 55.9		20	0.028	
	28688:T:Y (C or T) A:G or A	-26.27 / -26.27 or -23.6	53.8 / 53.8 to 55.9		2	0.003	
	28690:G:T	-26.27 / -22.37	53.8 / 51.8		3	0.004	
GGGAGCCTTGAATACACCAAAA (28681-28702) FP	28688:T:C*	-26.48 / -23.81	53 / 54.8	N	20	0.028	26 (0.036)
	28688:T:Y (C or T)	-26.48 / -26.48 to -23.81	53 / 53 to 54.8		2	0.003	
	28690:G:T	-26.48 / -22.58	53 / 51.1		3	0.004	
	28699:A:M (A or C)	-26.48 / -26.48 to -23.26	53 / 53 to 54.8		1	0.001	
GTGARATGGTCATGTGTGGCGG (15431-15452) FP	15435:A:R (A or G) T:(T or C)*	-28.79 / -28.79 to -24.58	56.7 to 58.6 / 56.7 to 60.4	RdRP	19	0.026	24 (0.033)
	15444:G:TC:A	-28.79 / -25.02	56.3 / 54.4		1	0.001	
	15451:G:AC:T	-28.79 / -25.57	56.3 / 54.4		4	0.006	
CTACATGCACCAGCAACTGT (23114-23133) FP	23116:A:TT:A	-25.1 / -22.22	51.8 / 51.8	S	1	0.001	15 (0.021)
	23116:A:W (A or T) T:(T or A)	-25.1 / -25.1 to -22.22	51.8 / 51.8		4	0.006	
	23118:A:M (A or C) T(G or T)	-25.1 / -25.1 to -22.72	51.8 / 51.8 to 53.8		3	0.004	

	23120:G:K (T) C:A	-25.1 / -25.1 to -20.93	51.8 / 49.7 to 51.8		1	0.001	
	23120:G:T C:A	-25.1 / -20.93	51.8 / 49.7		5	0.007	
	23123:C:T G:A	-25.1 / -20.24	51.8 / 49.7		1	0.001	
CGGATGGCTTATTGTTGGCG (25521-25540) FP	25528:C:T G:A*	-26.06 / -21.19	53.8 / 51.8	ORF3a	13	0.018	14 (0.020)
	25528:C:Y (C or T) G:A	-26.06 / -21.19	53.8 / 51.8 to 53.8		1	0.001	
ACACTAGCCATCCTTACTGCGCTTCG (26332-26357) Probe	26332:A:T T:A	-34.48 / -34	61.1 / 61.1	E	1	0.001	10 (0.014)
	26336:T:C A:G	-34.48 / -32.51	61.1 / 62.7		1	0.001	
	26338:G:T C:A	-34.48 / -31.19	61.1 / 59.5		2	0.003	
	26351:CGCTTC:-GCGAAG	-34.48 / -26.49	61.1 / 51.8		5	0.007	
	26356:C:A G:T	-34.48 / -29.64	61.1 / 59.5		1	0.001	
CGTTTGGTGGACCCTCAGAT (28320-28339) FP	28326:G:T C:A*	-25.51 / -21.87	53.8 / 51.8	N	9	0.013	10 (0.014)
	28337:G:T C:A	-25.51 / -22.49	53.8 / 51.8		1	0.001	
AGCAGTACGCACACAATCGAA (26354-26374) RP	26351:CGCTTC:-GCGAAG	-23.4 / -19.6	52.4 / 48	E	5	0.007	10 (0.014)
	26356:C:A G:U	-23.4 / -19.63	52.4 / 50.5		1	0.001	
	26356:C:T G:A	-23.4 / -20.05	52.4 / 50.5		1	0.001	
	26358:ATTGTGTGCGTACTGCTGCAATATTGTTAACGTGAGTCTTG:-	-23.4 / 0	52.4 / 12		1	0.001	
	26361:G:A C:U	-23.4 / -18.86	52.4 / 50.5		1	0.001	
	26370:C:T G:A	-23.4 / -20.59	52.4 / 50.5		1	0.001	

TTCGTCCGTGTTGCAGCCGA (201-220) Probe	203:C:T G:A	-28.54 / -22.99	55.9 / 53.8	ORF1a	1	0.001	8 (0.011)
	204:G:T C:A	-28.54 / -24.68	55.9 / 53.8		1	0.001	
	218:C:T G:A	-28.54 / -22.99	55.9 / 53.8		2	0.003	
	219:G:T C:A	-28.54 / -25.17	55.9 / 53.8		4	0.006	
ATATTGCAGCAGTACGCACACA (26360-26381) RP	26358:ATTGTGTGCGTACTGCTGCAATATTGTTAACGTGAGTCTTG:-	-25 / 0	53 / 0	E	1	0.001	8 (0.011)
	26361:G:A	-25 / -20.46	53 / 51.1		1	0.001	
	26370:C:T	-25 / -22.19	53 / 51.1		1	0.001	
	26375:G:A	-25 / -20.53	53 / 51.1		2	0.003	
	26376:C:A	-25 / -21.31	53 / 51.1		2	0.003	
	26378:A:R	-25 / -22.26	53 / 53 to 54.8		1	0.001	

Table 1: Summary of Primer and Probe sequences and genomic variants analysis.

The variant frequency, primers/probes cumulative variant frequency, Gibbs free energy (ΔG), melting temperature (Tm), and Extinction coefficient for reference and alternate in the Indian SARS-CoV-2 isolates. Only primers/probes with a cumulative variant frequency of more than 1% is included in this Table.

Tm- Melting Temperature, ΔG- Gibbs Free Energy, Ref- Reference, Alt- Alternate

* - Variants with >1% frequency were also found in primer / probes encompassing S, E, RdRP, ORF1a, and ORF3a genes

FIGURES

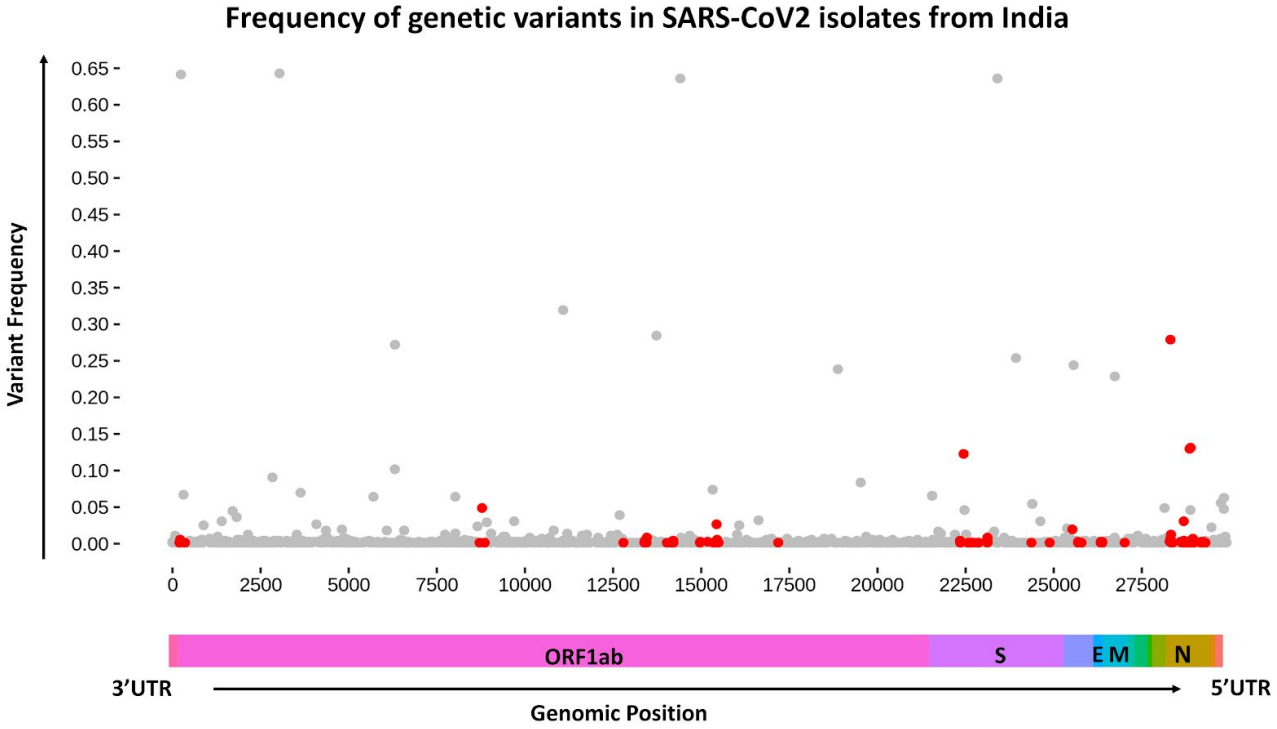

Figure 1. The frequency of genetic variants in SARS-CoV-2 isolates from India.
The variants mapping to oligonucleotide primer/probe sites are marked in red and others in grey. The panel at the bottom depicts the SARS-CoV-2 genome with genomic annotation.

SUPPLEMENTARY DATASETS

Supplementary Methods 1: Methodology for calculating primers/probes melting temperature and Gibbs free energy with variants impact.

Supplementary Data 1. Summary of the genomes of Indian isolates of SARS-CoV-2 available in public domain

Supplementary Data 2. Curated primers and probes sequence and their genomic coordinate used in the molecular assays for detection of SARS-CoV-2.

Supplementary Data 3. Summary of Primer and Probe sequences and genomic variants
Calculated variant frequency, cumulative variant frequency, and melting temperature (Tm) for reference and alternate in the Indian SARS-CoV-2 isolates in the primers and probes binding sites.