International Journal of Infectious Diseases 102 (2021) 460–462

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

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

ABSTRACT

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 $\geq$ 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 $\geq$ 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.

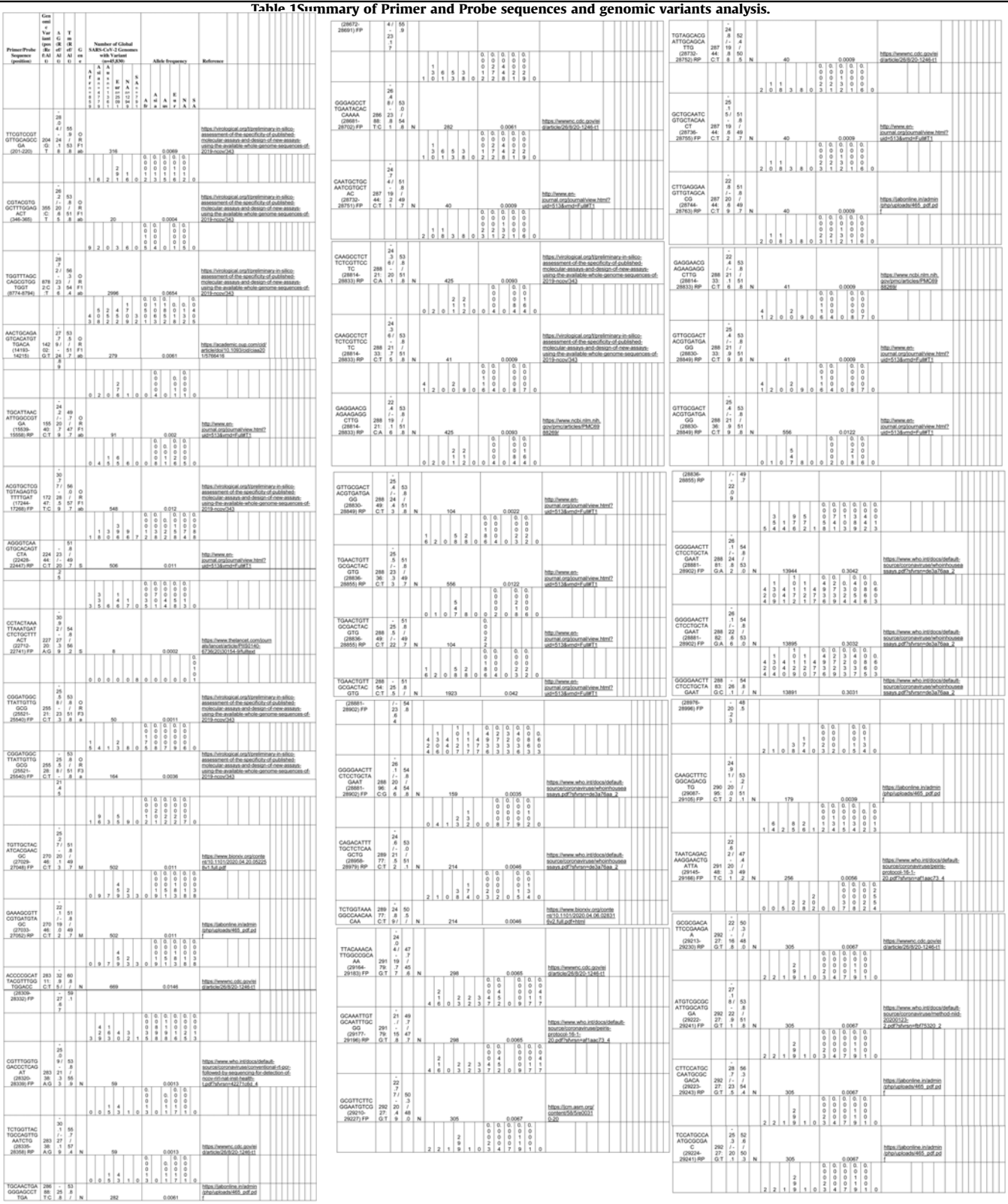
- SARS-CoV-2 variants impact RT-PCR efficiency in detection.
- A total of 29 global SARS-CoV-2 genetic variants had a frequency $\geq$ 1%.
- The thermodynamic stability of the virus-primers complex gets perturbed.
- A number of recommended primer or probe sequences had high variant frequency.

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

* 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.

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

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



The variant frequency, primers/probes frequency, Gibbs free energy (ΔG), 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, ΔG-Gibbs Free Energy, Ref- Reference, Alt- Alternate, No.- Number, Afr- Africa, Aus- Australia, Eur- Europe, NA-North America,SA- South America.

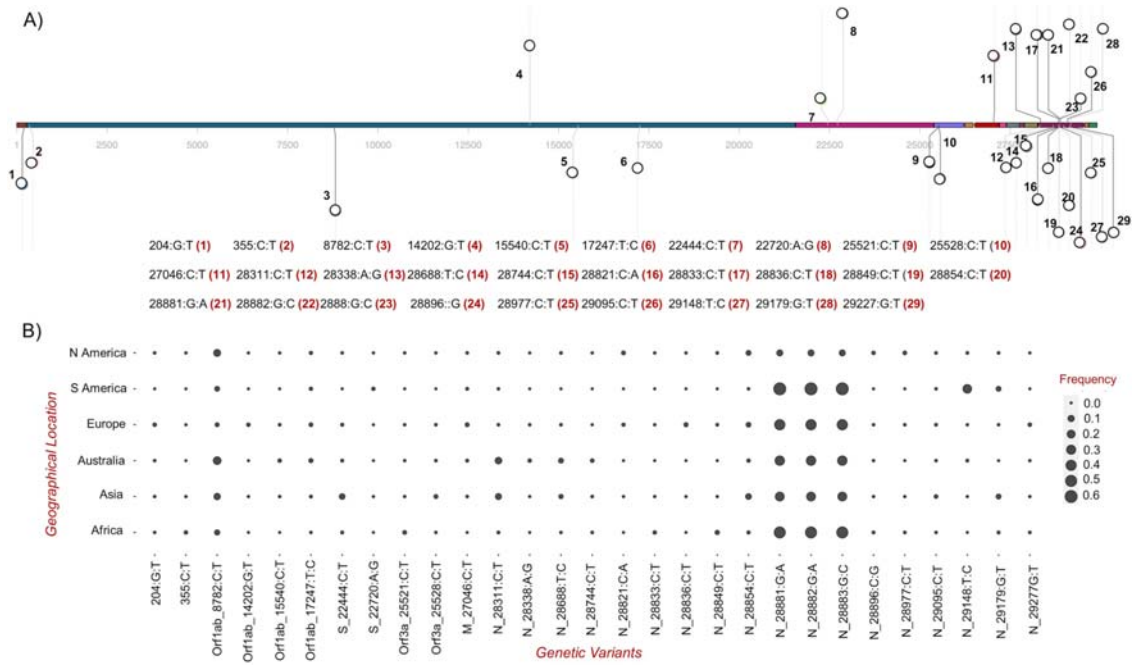


Figure 1. Genome Wide distribution (A) and frequency (B) of the 29 genetic variants with allele Frequency ≥ 1% in at least one of 6 continents.
Note: Colour required in Print.

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 re-aligned 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 ≥ 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 ≥ 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 in genomes mapped to the primer 2019-nCoV-NFP 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