J Biosci (2021)46:11
DOI: 10.1007/s12038-020-00131-5

Pattern of genomic variation in SARS-CoV-2 (COVID-19) suggests restricted nonrandom changes: Analysis using Shewhart control charts

SAURAV MANDAL[1]*, TANMOY ROYCHOWDHURY[2] and ALOK BHATTACHARYA[3]*

[1]*School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India*

[2]*Department of Internal Medicine, Cardiology, University of Michigan, Ann Arbor, MI, USA*

[3]*Department of Biology, Ashoka University, Sonepat, Haryana, India*

**Corresponding authors (Emails, sauravl3_sit@jnu.ac.in; alok.bhattacharya@gmail.com)*

MS received 4 June 2020; accepted 7 December 2020

SARS-CoV-2 is a member of the Coronavirus family which recently originated from the Wuhan province of China and spread very rapidly through the world infecting more than 4 million people. In the past, other Coronaviruses have also been found to cause human infection, but not as widespread as COVID-19. Since Coronavirus sequences constantly change due to mutation and recombination, it is important to understand the pattern of changes and likely path the virus can take in the future. In this study, we have used the Shewhart control chart to identify and analyze hypervariable (hotspots) and hypovariable (coldspots) regions of the virus. Our analysis shows that SARS-CoV-2 has changed in a few regions of the genome. Analysis of SARS-CoV-1 and MERS sequences suggests that over time, mutations start accumulating in different regions and most likely SARS-CoV-2 may also follow a similar path. The results suggest a possible emergence of modified viruses over some time.

Keywords. SARS-CoV-2; SARS-Cov-1; MERS; Shewhart control chart; hotspots

1. Introduction

Coronaviruses are single-stranded RNA viruses that vary in size from 26 to 32 kilobases (Schoeman and Fielding 2019). Based on the serology and genome, the Coronavirinae subfamily is divided into four major genera: Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and Deltacoronavirus (Wertheim *et al.* 2013). Generally, these display host specificity, the former two primarily infect mammals, whereas the latter two predominantly infect birds (Chan *et al.* 2013). Coronaviruses mainly cause respiratory and gastrointestinal tract infections, and several different instances of human infections have been seen. Some of these are the severe acute respiratory syndrome coronavirus (SARS-CoV-1, SARS1), and the Middle East respiratory syndrome coronavirus (MERS-CoV, MERS) where substantial numbers of humans were infected (Chu *et al.* 2014). These viruses are highly prevalent in animals. Novel viruses emerge occasionally due to genomic alteration as a result of cross-species infections and high rate of sequence diversity and frequent recombination of their genomes (Rehman *et al.* 2020).

The viral genome is the positive strand, and therefore the genomes need to be copied using a virally coded RNA dependent RNA polymerase (RdRp) for replication. RdRps display a high level of error during copying RNA strands, thereby introducing a number of

This article is part of the Topical Collection: COVID-19: Disease Biology & Intervention.

Electronic supplementary material: The online version of this article (https://doi.org/10.1007/s12038-020-00131-5) contains supplementary material, which is available to authorized users.

http://www.ias.ac.in/jbiosci
Published online: 03 February 2021

mutations in the genome (Castro *et al.* 2005). Coronaviruses also undergo a high degree of recombination among each other, probably using a template-switching mechanism (Lai Lai 1992). The genome plasticity allows viral evolution and emergence of novel viruses with altered host range and pathology. The SARS-CoV-2 or COVID-19 virus has emerged from a zoonotic source as late as December 2019 and infected a couple of million people worldwide with high mortality ranging from around 3-10%. This virus displays altered disease transmission and pathology as compared to SARS1 and MERS (Lai *et al.* 2020).

A large number of COVID-19 genomes from different parts of the world have been sequenced. The sequenced information has been used to classify viruses, in order to understand the origin of these viruses. The results from these analyses clearly showed that COVID-19 is a lineage B Betacoronavirus and closely resembles two bat SARS-like coronaviruses with an extremely high degree of sequence identity with each other and bat viruses (Velavan *et al.* 2020). Less than 1% sequence divergence observed suggests that these viruses have evolved in humans recently. Most of the computational approaches towards the analysis of COVID-19 sequences are for better classification of the viral sequences and eventually understanding features and patterns that correlate with disease severity. For achieving these goals both alignment-based and free methods have been used with variable success (Robson 2020).

One of the major concerns of COVID-19 is viral evolution and its potential impact on our future disease surveillance and prevention strategies. For example, if the targeted domains for vaccines change so that antibodies do not recognize the altered virus, vaccinated individuals will not be protected (Wang *et al.* 2019). In order to study the evolution of viruses, there is a need to understand patterns of genomic changes in the viruses that have been isolated after the pandemic has started. In this study, we present the results of an analysis of COVID-19 sequences using Shewhart control charts (SCC) which were developed for the identification of genomic hot and cold spots.

2. Results and discussion

2.1 *Nonrandom distribution of single nucleotide variations (SNV)*

We have carried out a distribution of SNVs between two isolates one being the reference genome NC_045512 of COVID-19 and another a randomly generated SNVs following Poisson distribution. For this analysis, we have used a non-overlapping window of 250 nucleotides. SNV counts were plotted within the 250 nucleotide segments (figure 1). For comparison, we have also generated the same number of SNVs with random map positions (Supplementary figure 1). It is clear that the SNVs are distributed nonrandomly across the genome. The bin size of 250 nucleotides was chosen considering our past experience and the size of the genome. The optimal bin size of 250 has been chosen through prior experience and observations by considering all the SNVs spread over the entire genome. Previously, Das et. al. in 2012 had found that optimal bin size is 2000 for *Mycobacterium tuberculosis H37Rv* genome (4 million genome size). The bin size chosen here was based on the size of the genome being analysed in comparison to that of *M. tuberculosis*.

2.2 *Distribution of SNVs in COVID-19*

Genomic sequences of a total of 552 isolates of COVID-19 were used to generate SCC. SCC a useful statistical test is used for quality control in the manufacturing process (Nelson 1984). This chart can be used to monitor one or more variables that are directly or indirectly associated with a production process and can detect instantly any significant deviation during the manufacturing process. SCC was used for finding hotspot and coldspot regions in bacterial genomes (Das *et al.* 2012). In the application, the genomes were divided into segments of equal length and a number of SNVs were noted that map to these regions. Basic

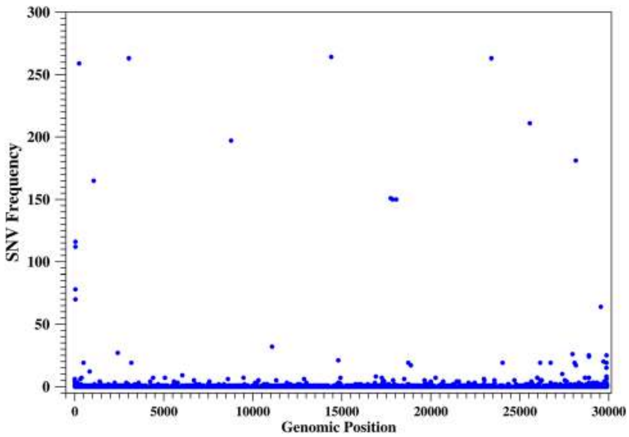

Figure 1. SNVs frequency identified at each genomic location by comparing *NC_045512* genome along with all the SARS-CoV-2 isolates.

characteristics of these charts are Center Line (CL), the Upper Control Limit (UCL) and Lower Control Limit (LCL). This chart graphically displayed the regions that showed the number of SNVs beyond a certain statistical limit (out of control) that is, above UCL in a given location of the genome. Hotspots are defined as the regions that are above UCL and correspondingly coldspots map close to zero lines (Koutras *et al.* 2007). In effect, SCC finds regions that display statistically higher or lower levels of variations compared to all other regions.

Cumulative SCC of 552 number of COVID-19 genomes is shown in figure 2. For this analysis, we have used a window size of 250. The results showed that there are only 9 hotspot regions. The majority of the regions lie in the cold spot areas. It indicates a high degree of similarity and a low level of variations among different isolates. The hotspot regions were further studied by mapping to the protein sequences and 3D structure (UniProtKB). The summary of the results is shown in table 1.

Some of the hotspots map to regions that encode important viral proteins such as helicase, proofreading exonuclease, S protein and transmembrane domain. The longest region was found spanning helicase and proofreading exonuclease enzymes. The changes in these could in principle alter the properties of the enzymes making them more resistant to nucleoside analogues, potential drug candidates (Graepel *et al.* 2017). Variations in the RBD region of S protein can likely alter the nature of interaction with the host receptor changing binding characteristics including host range. Similarly, hotspot regions in important viral proteins ns8 and ORF3a can also affect virus-host interaction.

2.3 *Distribution of SNVs in SARS-CoV-1*

SCC was applied to understand the pattern of genomic variations in SARS1. Since SARS1 and COVID-19 are highly related viruses it was important to see if the hotspot and coldspot regions are the same in these two viruses. The result is shown in figure 3. Unlike COVID-1, there were no clear coldspot regions in this virus. Statistically, significant hotspot regions were

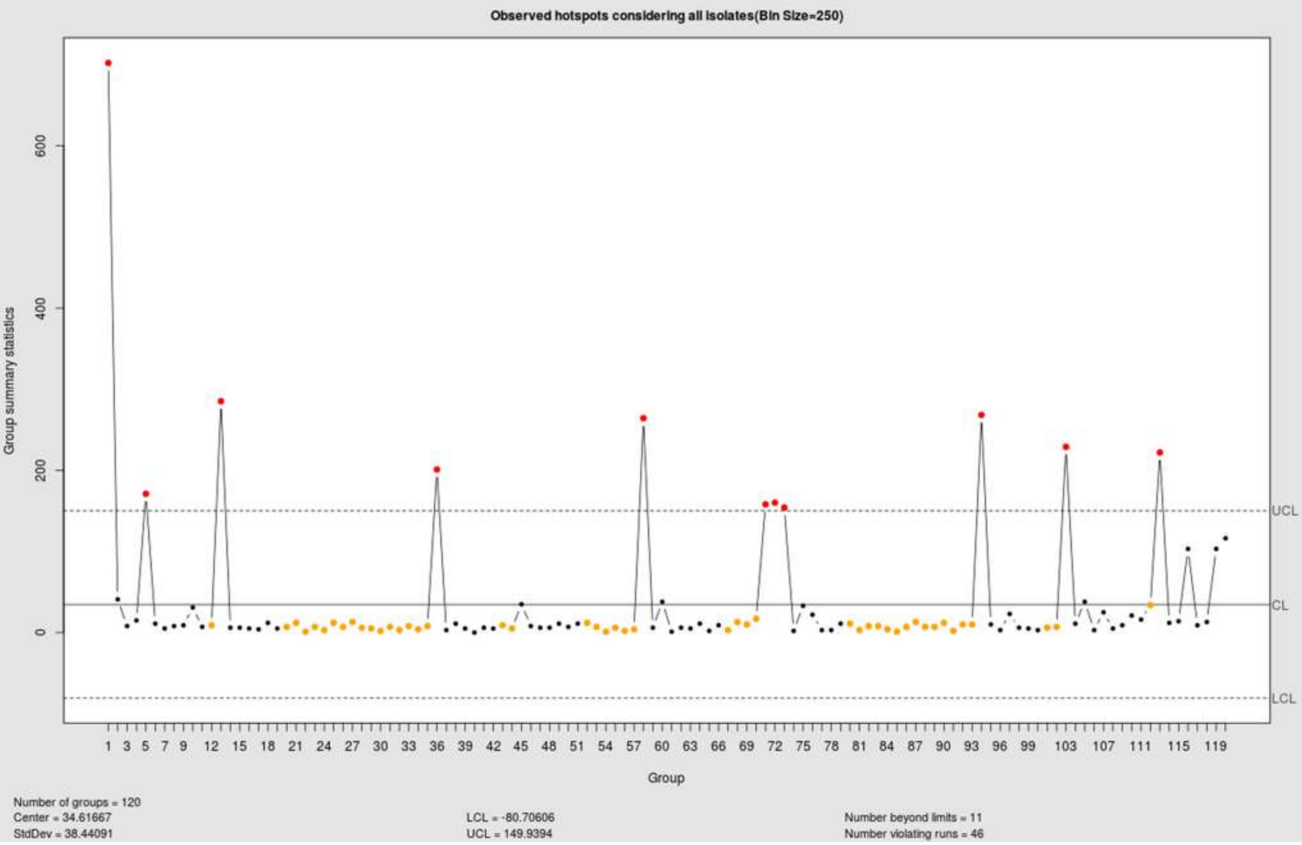

Figure 2. Red dots indicate the Hotspot regions of SARS-Cov-2. Using Figure 1 binning of size 250 was performed to find the SNV frequency in each bin by applying Shewhart control chart. Refer to table 1 for hotspot location details.

Table 1. SARS-CoV-2 hotspot location details.

	Genomic position					
No.	From	To	Protein/peptide	Number of SNVs present	Domain	Comments
1	0	250	5′-UTR ORF1	702	Consists of stem-loop 5	ATG present from location 265
2	1000	1250	nsp2	171	Topological domain	Plays a role in the modulation of host cell survival signaling pathway
3	3000	3250	nsp3	285	Predicted phosphoesterase transmembrane domain 1	Responsible for the cleavages located at the N-terminus of the replicase polyprotein
4	8500	8750	nsp4	8	Transmembrane domain 2 (Helical)	Cleavage by PL-PRO
5	14250	14500	Nsp12 /RNA-dependent RNA polymerase	264	Consists of Helix, Zinc finger and sheet	Domain 2 of the nsp12 protein
6	17500	18000	Helicase	318	CV ZBD (Zinc Binding Domain), RNA virus helicase ATP-binding	ZBD in N-terminus displaying RNA and DNA duplex-unwinding activities with 5' to 3' polarity
7	18000	18250	Proofreading exoribonuclease	154	Chain	Homologous to Guanine-N7 methyltransferase
8	23500	25500	S protein	63	Receptor-binding domain (RBD)	RBD binds to the human ACE2
9	25500	25750	ORF3a	229	Transmembrane domain(helical)	Forms homotetrameric potassium sensitive ion channels (viroporin)
10	28000	28250	ns8/ORF8	222	Signal peptide and chain region	Plays a role in host-virus interaction

mostly concentrated towards the 3′-end of the genome. A low level of variations was observed all along the genome, the main reason for the lack of cold spot reasons. Some of the genes that showed a high degree of variations are helicase, proofreading exoribonuclease, nsp2, nsp3, etc. (supplementary table 1). It appears that helicase and proofreading exonuclease are intrinsically variable in both the SARS viruses.

The difference in patterns observed between SARS-CoV-1 and SARS-CoV-2 may be due to restrained viral evolution due to the recent appearance of the latter (only 4 months). The SARS-CoV-1 sequences that were used to derive hotspots were deposited over several years (from 2003-2019). It is likely that the highly divergent SCC pattern observed by us, maybe due to the long time that it took for the evolution of the virus. Therefore, we carried out an SCC analysis of only those sequences deposited in 2003. The results are shown in figure 3. Viral sequences were found to be much more conserved with several cold spots and only two hotspots. The number of variations found in the hotspots was also much less compared to SARS-CoV-2.

2.4 *Distribution of SNVs in MERS-CoV*

MERS sequences were also analyzed by SCC. The observed pattern was substantially different from that of both SARS1 and COVID-19 (figure 4). Variations were spread all across the genome and only three hotspots were observed. Unfortunately, the number of sequences available for analysis were only a few which did not allow definite conclusions to be made. The three hot spots mapped to Spike, ORF3, ORF4a nsp3 and nsp13 proteins (supplementary table 2).

3. Methods

3.1 *Datasets*

The Wuhan seafood market pneumonia virus (COVID-19 virus/SARS-CoV-2) isolate Wuhan-Hu-1 was downloaded from the National Center for Biotechnology Information (NCBI) database (Brister *et al.* 2015). All of the available 617 sequences of COVID-19 virus, SARS-Cov-1 and MERS sequences were obtained

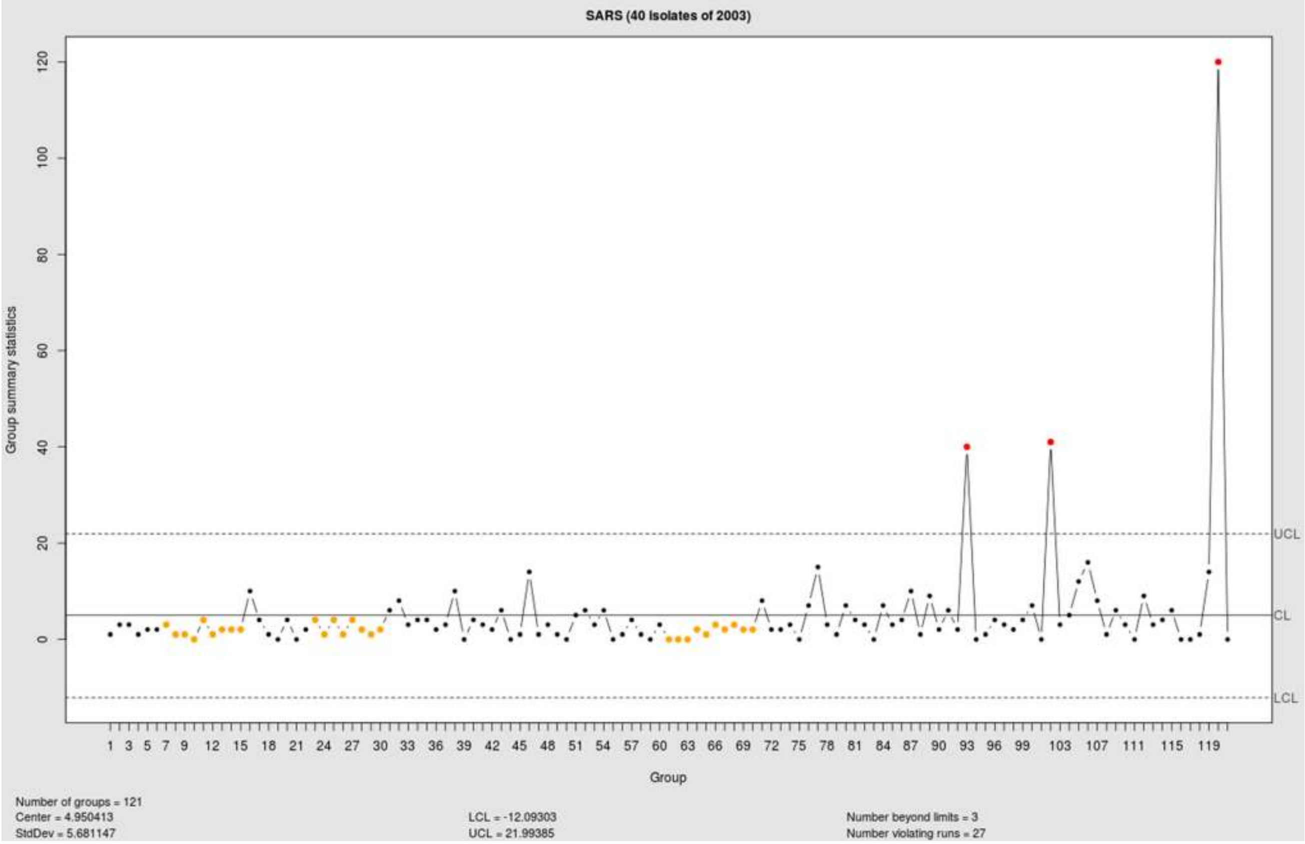

Figure 3. Shewhart Control Chart showing the hotspots in red dots above the UCL of SARS-CoV. The hotspot details are presented in supplementary table 1.

from the NCBI, on 22 April 2020. The data included a total of 552 of COVID-19, 40 of SARS-1 and 25 of MERS genomes by considering NC_045512, NC_004718 and NC_019843 as reference genomes respectively. The datasets within a time frame of one year are chosen for SARS-1(2003 to 2004) as well as MERS (2012 to 2013) in comparison with the COVID-19 (2020) considering the evolutionary rate of mutation. The whole-genome length of these reference genomes was 29903 for NC_045512, 29751 for NC_004718 and 30119 for NC_019843. All the isolates studied were isolated from the human host.

3.2 *Identification of SNVs*

Multiple sequence alignment was performed using *clustalw* (Thompson *et al.* 1994). Reference genomes of COVID-19, SARS-1 and MERS were used to identify SNVs from isolates of respective categories. For each genomic position, the frequency of SNVs was calculated and this information was used for further downstream analysis. SNV frequency is defined here as the number of SNVs present or absent in a particular genomic location among all the genomic isolates being considered. In figure 1 we observe that in the *x-axis* we have genomic positions and the *y-axis* represents the different SNVs present in numbers of isolates (out of a total of 552) of SARS-CoV-2. For example, if an SNV is present in a large number of isolates, then that position is a hotspot.

The whole-genome (29903 bp) of COVID-19 was divided into 120 non-overlapping bins of 250 bp. The number of SNVs in each bin was used as an input to the Shewhart Control Chart (SCC). SCC is a statistical quality control method for the identification of hotspots and coldspots. SSC has previously been successfully implemented to identify the hotspot and the coldspot regions in the *Mycobacterium tuberculosis* genome (Das *et al.* 2012). We have identified a total of 4189, 2375 and 480 SNPs from SARS-CoV-2 (552 isolates), MERS (25 isolates) and SARS-1(40 isolates) isolates respectively. The unique number of SNPs present were 715, 282 and 292 from SARS-CoV-2, MERS and SARS-1 respectively.

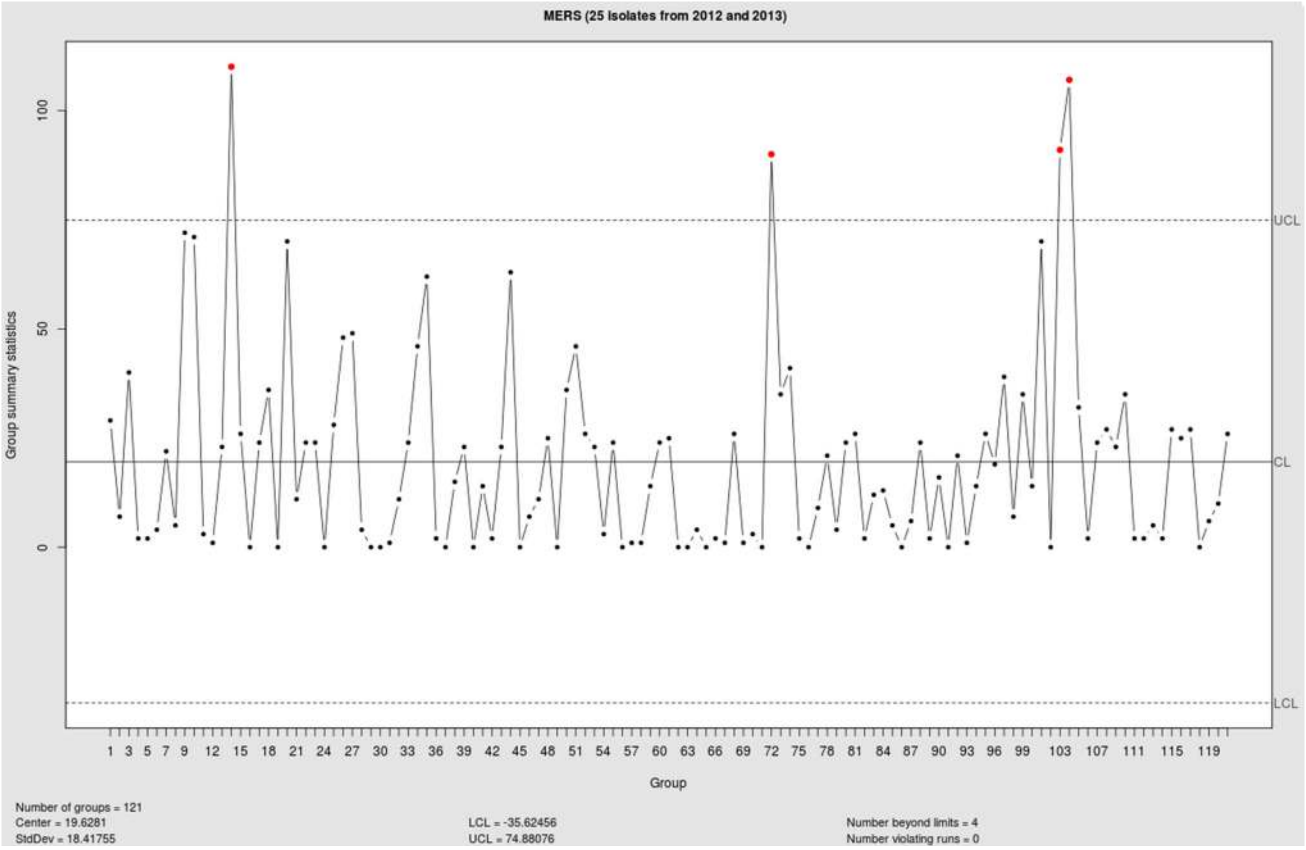

Figure 4. Hotspots identified using SCC indicated as red dots with a bin size of 250. Supplementary table 2 shows the details of the hotspot locations.

3.3 *Shewhart Chart*

SCC is a method to graphically display the quality of the process or the product that has been measured from a sample with respect to the sample number. SCC is characterized by Upper Control Limit (UCL), the Lower Control Limit (LCL) and the CenterLine (CL) (figure 2). These quantities are considered as shown below

$$\begin{aligned}\text{UCL} &= E[Y] + 3 \times \sqrt{VAR(Y)}\\ \text{CL} &= E[Y]\\ \text{LCL} &= E[Y] - 3 \times \sqrt{VAR(Y)}\end{aligned}$$

where Y is g(X). For the process data vector g(X) is a statistical function considered for estimating the mean of the process. E(Y) and VAR(Y) are the mean and variance of Y respectively. For mean μ and standard deviation σ the CL is m μ, UCL is $\mu + 3\sigma$ and LCL is $\mu - 3\sigma$. Essentially the CL, UCL and LCL are the three-sigma limits (Koutras *et al.* 2007; Benneyan *et al.* 2003).

LCL is the cutoff line or limit to figure out the outliers of the data points. The data points that are outside the limits of UCL and LCL or in other words 3σ (3 standard deviations) away from the mean(μ) are the genomic location exhibiting higher or lower mutations compared to other regions.

4. Conclusion

Viral replication introduces a number of sequence variations due to inaccuracy during the copying process. Potentially these alterations can change the behavior of the viruses in the future. The variation of SNVs across the entire genome of SARS-CoV-2 can be observed in figure 1. Some of the SNVs tend to occur more in certain genomic regions. These hotspots are more elucidated in our statistical approach of Shewhart Control Chart which identifies regions that have statistically higher or lower SNVs compared to other regions. We have also performed our analysis of closely related viruses such as MERS and SARS1. The analysis of SARS-CoV-1 and MERS sequences suggests that over time mutations start accumulating in different regions and most likely SARS-CoV-2 may also follow a similar path. Recently some analysis has been done to understand the variation in SARS-C0V2 genomes. Zhou *et al.* performed the identification and characterization of a 2019-nCoV and used nucleotide identity in percentage from SARS-CoV BJ01 and other bat genomes (Zhou *et al.* 2020). In the second study essentially viruses from different geographic origins were included and region-specific variations were enumerated (Laamarti *et al.* 2020). The method used and the results obtained are different from this report.

Our method of finding hotspots as well as cold spots (something other studies normally do not take into account) is different from others and helps us to understand the future course of virus evolution There is a possibility that the evolved virus may have different properties.

Acknowledgements

SM is supported by the Indian Council of Medical Research (ICMR) through the Senior Research Fellowship. AB and SM thank CDAC-Pune for providing high-performance computing (HPC) facility.

References

Benneyan JC, Lloyd RC and Plsek PE 2003 Statistical process control as a tool for research and healthcare improvement. *Qual. Saf. Health Care* **12** 458–464

Brister JR, Ako-Adjei D, Bao Y and Blinkova O 2015 NCBI viral genomes resource. *Nucleic Acids Res.* **43** D571–D577

Castro C, Arnold JJ and Cameron CE 2005 Incorporation fidelity of the viral RNA-dependent RNA polymerase: a kinetic, thermodynamic and structural perspective. *Virus Res.* **107** 141–149

Chan JF, To KK, Tse H, Jin DY and Yuen KY 2013 Interspecies transmission and emergence of novel viruses: lessons from bats and birds. *Trends Microbiol.* **21** 544–555

Chu H, Zhou J, Wong BH, *et al.* 2014 Productive replication of Middle East respiratory syndrome coronavirus in monocyte-derived dendritic cells modulates innate immune response. *Virology* **454** 197–205

Das S, Duggal P, Roy R, *et al.* 2012 Identification of Hot and Cold spots in genome of *Mycobacterium tuberculosis* using Shewhart Control Charts. *Sci. Rep.* **2** 1–6

Graepel KW, Lu X, Case JB, *et al.* 2017 Proofreading-deficient coronaviruses adapt for increased fitness over long-term passage without reversion of exoribonuclease-inactivating mutations. *MBio* **8** e01503–e01517

Koutras MV, Bersimis S and Maravelakis PE 2007 Statistical process control using Shewhart control charts with supplementary runs rules. *Methodol. Comput. Appl. Probab.* **9** 207–224

Laamarti M, Alouane T, Kartti S, *et al.* 2020 Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geodistribution and a rich genetic variation of hotspots mutations. *PLoS One* **15** e0240345

Lai CC, Shih TP, Ko WC, Tang HJ and Hsueh PR 2020 Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges. *Int. J. Antimicrob. Agents* **55** 105924

Lai MM 1992 RNA recombination in animal and plant viruses. *Microbiol. Mol. Biol. Rev.* **56** 61–79

Nelson LS 1984 The Shewhart control chart—tests for special causes. *J. Quality Technol.***16** 237–239

Robson B 2020 Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus. *Comput. Biol. Med.* 103670

Schoeman D and Fielding BC 2019 Coronavirus envelope protein: current knowledge. *Virol. J.* **16** 69

Rehman SU, Shafique L, Ihsan A and Liu Q 2020 Evolutionary trajectory for the emergence of novel coronavirus SARS-CoV-2. *Pathogens* **9** 240

Thompson JD, Higgins DG and Gibson TJ 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22** 4673–4680

Velavan TP and Meyer CG 2020 The COVID19 epidemic. *Trop. Med. Int. Health* **25** 278

Wang XY, Wang B and Wen YM 2019 From therapeutic antibodies to immune complex vaccines. *NPJ Vaccines* **4** 1–8

Wertheim JO, Chu DK, Peiris JS, Kosakovsky Pond SL and Poon LL 2013 A case for the ancient origin of coronaviruses. *J. Virol.* **87** 7039–7045

Zhou P, Yang XL, Wang XG, *et al.* 2020 A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579** 270–273

Corresponding editor: Sreenivas Chavali