Comparative genomics of global SARS-CoV-2 quasispecies offers insights into its microevolution and holds implications for pathogenesis and control

Santi M. Mandal[1], Suresh K. Mondal[1], Shriparna Mukherjee[2], Wriddhiman Ghosh[3] and Ranadhir Chakraborty[4,*]

[1] Central Research Facility, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India.
[2] Department of Botany, Prasannadeb Women's College, Jalpaiguri, West Bengal, India.
[3] Department of Microbiology, Bose Institute, P-1/12 CIT Scheme VII M, Kolkata 700054, West Bengal, India.
[4] Department of Biotechnology, University of North Bengal, Raja Rammohanpur, Darjeeling 734013, West Bengal, India.

* **Correspondence:** Prof. Ranadhir Chakraborty,
Department of Biotechnology,
University of North Bengal,
Raja Rammohanpur,
P.O. - NBU, District - Darjeeling,
West Bengal, Pin - 734013, India.
Telephone: +91-9434872273; Fax: +913532699001
E-mail: rcnbusiliguri@gmail.com

Running Title: *SARS-CoV-2 microevolution*
Key words: comparative genomics, SARS-CoV-2, microevolution, quasi-species, point mutation, disinfectants as mutagens

Abstract

In the wake of the current SARS-CoV-2 pandemic devastating the world, it is imperative to elucidate the comparative genomics of geographically-diverse strains of this novel coronavirus to gain insights into its microevolution, pathogenesis and control. Here we explore the molecular nature, genome-wide frequency, and gene-wise distribution of mutations in three distinct datasets encompassing 68 SARS-CoV-2 RNA-genomes altogether. While phylogenomic analysis revealed parallelism between the evolutionary paths charted by distinct quasispecies clusters of the virus, occurrence of mutations across genomes was found to be non-random. Whereas deletion mutations are extremely scarce and insertions totally absent, of all the instances of single nucleotide substitution detected, the overwhelming majority were transition mutations with cytidine to uridine being the most prevalent type. Propensity of this transition could be attributed to hydrolytic deamination mediated by ultra-violet irradiation or bisulfite reagent, both of which find wide usage as sterilizer/disinfectant.

Transversions, albeit few and predominated by the guanosine to uridine form, were found concentrated in loci encoding the structural proteins of the virus, so might confer versatile tissue-colonization potentials. Mutation frequency of the three distinct genome-sets ranged narrowly between 0.07-1.08 × 10^{-4} nucleotides positions mutated per nucleotide aligned. Gene-wise mapping of the global mutations illuminated the highly conserved nature of the genes encoding the non-structural proteins Nsp7, Nsp8 (two essential cofactors of the viral RNA-dependent RNA-polymerase) and Nsp9 (Nsp8-interacting single-strand RNA-binding protein), plus the envelope protein E (involved in SARS-CoV-2 assembly, budding and pathogenesis). These mutation-free genomic loci and/or their protein products could be potent targets for future drug designing/targeting.

1. Introduction

On 30 Dec 2019, ophthalmologist Li Wenliang in Wuhan, Hubei province, China, first recognized and communicated about the outbreak of a contagious illness resembling severe acute respiratory syndrome (SARS), which, subsequently, went on to be identified as 2019 novel coronavirus disease (COVID-19; causative agent: SARS coronavirus 2, abbreviated as SARS-CoV-2; see Green, 2020) that has spread to 215 countries, infecting 5,307,298 people, and killing 342,070 as of 26 May 2020 (https://covid19.who.int). The first whole genome sequence of SARS-CoV-2 was deposited to GenBank (MN908947.3) on January 5 by researchers of Shanghai Public Health Clinical Center and School of Public Health, Fudan University, Shanghai, China (Wu et al., 2020). SARS-CoV-2 is an enveloped, positive-sense, single-stranded RNA virus containing a 29,903 nucleotide genome having an untranslated segment of 254 and 229 nucleotides at the 5' and 3' ends respectively. Its putative genes encode a surface spike glycoprotein, an envelope layer glycoprotein, a replicase intricate, a nucleocapsid phosphoprotein, and five other non-basic proteins (Wu et al., 2020). More recent investigations revealed high gene-arrangement similitudes of SARS-CoV-2 with coronaviruses found in bats (*Rhinolophus sinicus*) (Zhou et al., 2020; Paraskevis et al., 2020) and Sunda Pangolin (*Manis javanica*) (Liu et al., 2019), which could be indicative of SARS-CoV-2 being a zoonotic disease (Lu et al., 2020). This said, human to human transmission of SARS-CoV-2 is now well-established, and its infection has spread across geographical and political barriers, courtesy of unbridled human travel across the globe. The virus spread rapidly in Japan, Australia, Southeast Asia, Western Europe and the Middle East. As for the USA, Canada and India, Covid-19 specifically arrived on the shores of these countries on 13, 22 and 30 January respectively, in the year 2020. Whereas no vaccine or therapeutic intervention has yet been developed to cure or mitigate Covid-19, a rapid accumulation of mutations in the genome of SARS-CoV-2 has been identified as a major challenge in taming the pandemic (Benvenuto et al., 2020) that, according to World Health Organization (WHO), has been raging the world since 11 March 2020 (https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19). In this context of SARS-CoV-2 population genetics, the present study analyzes hundreds of whole genome sequences of this novel coronavirus, focusing on 19 potential quasi-species isolated from 15 different countries (or distinct geographies) plus 46 Indian isolates, to delineate SARS-CoV-2 phylogeny, and then identify the distribution pattern and chemical nature of a substantial number of point mutations in the global SARS-CoV-2 genome. Consideration of the findings in the light of the existing knowledge on molecular biology and chemistry of point mutations (Hurlbert et al., 1960; Matallana-Surget et al., 2008; Phe et al., 2009) revealed potential mechanisms that could be instrumental in accelerating mutation recruitment in the RNA genome of SARS-CoV-2.

2. Experimental procedure

2.1 Comparative genomics

Of the 4561 SARS-CoV-2 whole genome sequences deposited thus far in GenBank, >84% are from USA, followed by China (>5%), India (>1.3%) and Taiwan (>1%). Genome sequences of this novel coronavirus have also been deposited from other countries (or distinct geographies) such as Australia, Belgium, Brazil, Colombia, Czech Republic, Finland, France, Germany, Greece, Hongkong, Iran, Iraq, Israel, Italy, Japan, Kazakhstan, Malaysia, Nepal, Netherlands, Nigeria, Pakistan, Peru, Philippines, Puerto Rico, South Africa, South Korea, Spain, Serbia, Sri Lanka, Sweden, Thailand, Tunisia, Turkey and Vietnam. From these, a total of 13 genome sequences of SARS-CoV-2 strains (Table 1) isolated from 11 highly-affected countries, namely Brazil (accession number MT126808.1), China (MN908947.3 and MT259230.1), India (MT012098.1), Iran (MT320891.1), Israel (MT276598.1), Italy (MT077125.1), Japan (LC528232.1), South Korea (MT304476.1), Spain (MT292575.1), Sweden (MT093571.1) and the USA (MT295464.1 and MT295465.1) were retrieved from GenBank and each of them was subjected to BlastN analysis with e-value 0.0 (Altshul et al., 1990). From the 100 significant alignments (query coverage and identity cut-offs both >99%) obtained for each of the 13 representative SARS-CoV-2 genome sequences (a total of 12 x 100 = 1200 alignments), a table was constructed where the individual matched-genomes were ascribed to columns representing different levels of query coverage (100% and 99%) and identity (100% - 99.93% at a resolution of 0.01%) (Table 2).

Multiple alignment of the 13 representative SARS-CoV-2 genome sequences was carried out using ClustalW (Thompson et al., 1994); in the process, the comprehensive set of point mutations present in the different protein-coding genes of the individual genomes were identified. The gene-wise distribution and the frequency of occurrence for the various molecular types of nucleotide substitutions/deletions across the 13 representative SARS-CoV-2 genomes were determined and then compared with the corresponding patterns generated for another set of 10 SARS-CoV-2 genomes, having accession numbers (country or geographical territory of origin) MT126808.1 (Brazil), MT996531.1 (China), MT365029.1 (Hong Kong), MT415321.1 (India), LC542809.1 (Japan), MT371050.1 (Sri Lanka), MT385458.1 (USA), MT192772.1 (Vietnam), MT066176.1 (Taiwan) and MT012098.1 (India), which were selected based on their highest sequence similarities with last member of this list (notably, MT012098.1 was also there in the first list of 13 reference genomes). Finally, to understand the microevolution of SARS-CoV-2 in the biogeographical context of its Indian strains, the above analyses were replicated in a dataset of 46 different genomes (Table S1) sequenced from three distinct states of India: Gujarat (26), Karnataka (14) and Andhra Pradesh (06).

Phylogenetic relationships between the 13 represntative SARS-CoV-2 genomes were delineated by constructing a Neighbor Joining tree (Saitou and Nei, 1987) using the MEGA7 software package (Kumar et al., 2016). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) were shown next to the branches (Felsenstein, 1985). The tree was drawn to scale, with branch lengths in the same units as those of the evolutionary distances between the individual genome sequences. Evolutionary distances, in turn, were computed according to Jukes and Cantor (1969) as the number of nucleotides substituted per site, after eliminating all such positions from the final alignment of sequences that contained gap or missing data.

3. Results

3.1 Small but definite divergences in SARS-CoV-2 genome sequences

Independent BlastN analyses carried out with the 13 representative SARS-CoV-2 genomes produced 100 significant alignments with similar sequences from other SARS-CoV-2 isolates. In this way, a total of 13 × 100 = 1300 alignments was obtained. Classification of these alignments based on their identity or query coverage levels showed that an overwhelming number (89%, i.e., 1156 out of the total 1300) of alignments involved identities between 99.97% and 100.00% alongside 99-100% query coverages (Table 1). Maximum number of maximally diverged (identities between 99.93% and 99.96%) SARS-CoV-2 genome alignments were obtained against the US strain having accession number MT295465A, the Swedish strain having accession number MT093571 and the

Indian strain having accession number MT012098. These diverged alignments, in turn, were mostly with regard to strains from USA, Hong Kong and China: for instances, (i) 26 other US strains showed 99.95% identities with the US representative MT295465A; (ii) nine US strains, seven Chinese strains and five Hong Kong strains showed 99.96% identities with the Swedish representative MT093571; while (iii) 19 US strains, seven Hong Kong strains and three Chinese strains showed 99.96% identities with the Indian representative MT012098.

The Neighbor Joining tree constructed based on the 13 representative SARS-CoV-2 genomes revealed a complex phylogenetic relationship between the different strains. A clear dichotomy was evident between the two strains from USA (having genome accession umbers MT295464.1 and MT295465.1) and the rest of the 11 entities included in the analysis. Subsequent branching pattern within the clade containing the 11 non-US strains indicated a polyphyletic relationship between the members. For instance, the prototypical Chinese strain (MN908947.3) formed a single-member branch, parallel to which three other clusters encompassing the strains from (i) Israel (MT276598.1) and Spain (MT292575.1); (ii) China (MT259230.1) and India (MT012098.1), and (iii) Brazil (MT126808.1), Iran (MT320891.1), Italy (MT077125.1), Japan (LC528232.1), South Korea (MT304476.1) and Sweden (MT093571.1) branched from the same node having considerable bootstrap support (88%).

3.2 Gene-wise mapping of mutations in global SARS-CoV-2 genomes

Multiple alignment of the 13 SARS-CoV-2 reference genomes revealed that over the 29,906 completely aligned nucleotide positions, 39 instances of single nucleotide substitution and one instance of deletion of three consecutive nucleotides AUU (Table 2) have occurred in one or more genomes at a frequency of [42 ÷ (29,906 × 13)] = 1.08 × 10^{-4} mutated nucleotides positions per nucleotide participating in the alignment (Table 3). Table 2 shows the nucleotide positions of all these 42 mutations with reference to the 5’ to 3’ sequence of the 29,903-nucleotide-long genome (MN908947.3) of the reference strain from Wuhan, China. Gene-wise mapping of the 40 global mutations encountered across the dataset of 13 reference genomes (Table 2) revealed that 17, out of the total 25 SARS-CoV-2 genes, contained mutation(s) in one or more reference genomes (Table 3). Besides, of the 40 global mutations, the first 24, with reference to the 1-29,903 nucleotide positions of the reference genome MN908947.3, were found to be located between nucleotide positions 313 and 20,268, within the foremost locus of SARS-CoV-2 genome which encodes the 16 non-stuctural proteins Nsp1 through Nsp16 and spans nucleotide positions 266-21,555. All of the 16 Nsp-encoding genes, except *nsp5*, *nsp7 nsp8* and *nsp9*, *nsp11* and *nsp16* have at least one mutation in one or more of the 13 reference genomes. The next seven global mutations, located between nucleotide positions 21,575 and 24,694, mapped in *geneS* which encodes the spike protein and spans nucleotide positions 21,563-25,384. Along the last ~5,000 nucleotides of SARS-CoV-2 genome, nine more mutations were encountered globally: two of these mapped in *orf3a*, one each in *orf6* and *orf8*, four in *geneN* that encodes a membrane protein, and one in the terminal untranscribed region.

Multiple alignment of the comparator set of 10 SARS-CoV-2 genomes showed that over the 29,854 completely aligned nucleotide positions, 18 instances of single nucleotide substitutions and one instance of deletion of three consecutive nucleotides AUU (Table 4) have occurred in one or more genomes at a frequency of [21 ÷ (29,854 × 10)] = 0.70 × 10^{-4} mutated nucleotides positions per nucleotide participating in the alignment (Table 3). Table 4 shows the nucleotide positions of all these 21 mutations with reference to the 5' to 3' sequence of the 29,854 nucleotide-long genome (MN012098.1) of the Indian strain used as the query reference for this dataset. Likewise, multiple alignment of 46 Indian SARS-CoV-2 genomes (listed in Table S1) showed that over the 29,741 nucleotide positions aligned, 138 instances of single nucleotide substitution and one instance of deletion of three consecutive nucleotides GUU (Table S2) have occurred in one or more genomes at a frequency of [141 ÷ (29,741 × 46)] = 1.03 × 10^{-4} mutated nucleotides positions per nucleotide participating in the alignment (Table 3). Table S2 shows the nucleotide positions of all these 141 mutations with reference to the 5' to 3' sequence of the 29,801 nucleotide-long genome (MT451882.1) of one of the 46 Indian strains that was isolated from Ahmedabad (Gujarat) and used as the reference sequence for this dataset.

Gene-wise mapping of the 19 global mutations detected across the 10 comparator genomes (Table 4) showed that nine, out of the total 25 SARS-CoV-2 genes, contained mutation(s) in one or more genomes of this dataset (Table 3). On the other hand, the 139 global mutations were encountered in the dataset of 46 Indian genomes (Table S2) mapped onto 16 out of 25 SARS-CoV-2 genes (Table 3). Comparison of the gene-wise mutation mapping patterns encountered in the three distinct SARS-CoV-2 genome sequence datasets revealed a global propensity of mutation accumulation in *nsp2*, *nsp3*, *nsp4*, *nsp12* and *nsp13*; *geneS* and *orf3a* in tandem with a global stringent sequence conservation in *nsp7*, *nsp8*, *nsp9* and *geneE*. Cross-dataset comparison also revealed a largely comparable mutability in SARS-CoV-2 strains across geographical territories (in Table 3 see data for number of nucleotide positions mutated per nucleotide aligned).

3.3 High frequency of C→ U mutations across global SARS-CoV-2 genomes

Of the 39 instances of single nucleotide substitution detected across the 13 SARS-CoV-2 reference genomes, a substantial 14 (i.e., 35%) involved the conversion C→U. Other substitutions detected, in their descending order of prevalence, were: G→A and G→U (at six different positions each), A→G (at five different positions) U→C (at four different positions), A→U, C→G, G→C, and U→G (at one position each). Of the 18 instances of single nucleotide substitution detected across the comparator set of 10 SARS-CoV-2 genomes, more than 50% (10 in number) involved the conversion C→U, while three and two each involved the conversions G→U and U→C respectively; the conversions A→C, A→G and U→Y (Y = an undefined pyrimidine nucleotide) were detected at one instance each. Furthermore, of the 138 instances of single nucleotide substitution detected across the Indian set of 46 SARS-CoV-2 genomes, slightly more than 40% (56 in number) involved the conversion C→U. Other substitutions detected, in their descending order of prevalence, were: G→U (26), A→G and G→A (at 13 different positions each), U→C (at 11 different positions), C→A (at 7 different positions), G→C (at 5 different positions), A→C (at 3 different positions) U→G (at 2 different positions) A→N and U→A (at 1 position each).

4. Discussion

4.1 Microevolutionary dynamics of SARS-CoV-2

Pace of mutation accumulation due to replication errors is higher in the RNA genomes of viruses than the corresponding spontaneous mutation rates in the DNA genomes of most other living entities. This leads to an unrelenting creation of mutant genomes for any RNA virus, alongside a rivalry among the extant variants, including the more advanced ones that are added to the viro-diversity over time (Eigen et al., 1988). The active genomic variants that are in circulation within global/local RNA virus populations are regarded as quasispecies (Eigen et al., 1988), and it is obvious that the mutations eventually maintained within the natural populations of a virus prequalify as they cause no impediment to the ability of the virus to replicate and complete the infection cycle. In the present study, the mutation frequencies calculated for the three distinct sets of SARS-CoV-2 genomes fell within the narrow range of 0.07 – 1.08 × 10^{-4} nucleotides positions mutated per nucleotide aligned. This indicated that a majority of the individual genomes explored differed in one or more nucleotides from the consensus sequence of the pan-genome. The small but definite genomic divergences revealed further corroborated that during the present pandemic the reservoir of SARS-CoV-2 quasispecies is rapidly expanding across geographies. This rich stock of genotypic, and therefore potentially phenotypic, variants is likely to hold major implications for potential multifaceted adaptations of this novel coronavirus within human hosts, and in doing so have serious consequences on the resultant pathogenesis, disease complications and control (Domingo and Holland, 1992). Phylogenetically, SARS-CoV-2 seems to be a complex group. Possibilities remain that subsequent revelation/incorporation of key intermediary links in the phylogenomic framework of SARS-CoV-2 can resolve the polyphylies apparent in the current tree topology. Subject to whether the retrogressively convergent past of SARS-CoV-2 evolution is at all unearthed, it may well be that distinct quasispecies clusters of the novel coronavirus are actually evolving along multiple parallel and discrete paths of genome innovation.

4.2 Conserved versus variable regions of the global SARS-CoV-2 genome

Gene-wise mapping of mutations in global SARS-CoV-2 genomes revealed the highly conserved nature of *nsp7*, *nsp8*, *nsp9* and *geneE*. The longest polycistronic RNA genome known for any virus is that of SARS-CoV-2. Two lengthy overlapping open-reading frames, *orf1a* and *orf1b*, constituting 71.2% of the 5$^{'}$-proximal polycistronic region, are translated to generate two replicase polyproteins, Pp1a and Pp1ab (Subissi et al., 2014). The two polyproteins are subsequently cleaved by the protease encoded by *orf1*, into 16 non-structural proteins Nsp1 through Nsp16, which in turn give rise to a replication and transcription complex (RTC). RTC renders the complex mechanism of interactome-driven viral RNA synthesis involving multiple copies of the full length genome and also sub-genome length negative-stranded RNAs that produce subgenomic mRNAs to be translated into structural and accessory proteins (Subissi et al., 2014). One of the most important aspect of genome replication in SARS-CoVs is the processivity of their RNA synthesis machinery. In SARS-CoV-2, a hexadecamer formed by Nsp7 and Nsp8 (Kirchdoerfer and Ward, 2019) has been established as the processivity cofactor for RNA-dependent RNA-polymerase (Nsp12) activity (Gao et al., 2020). The *in vitro* polymerase activity of Nsp12 is reportedly weak and non-processive, as desirable for efficient replication of large RNA genomes; but its processivity in primer extension mode is apparently enhanced by orders of magnitude upon the formation of a complex with Nsp7 and Nsp8 (teVelthius et al., 2010). The centrality of this tripartite holoenzyme (Nsp7-Nsp8-Nsp12), which is unique among all RNA viruses, in rendering RNA synthesis in the absence of exogenous RNA primers is corroborated by previous reports of point mutations in *nsp7* and *nsp8* of other coronaviruses jeopardizing genome replication and growth cycle (Subissi et al., 2014). Nsp9, which is also a unique protein of the coronaviruses, possesses a novel fold constituted of an open β-barrel having six strands along with an α-helix containing GXXXG protein-protein interaction motif engaged in the formation of a dimer that is critical for virus replication (Milkins et al., 2009). Furthermore, Nsp9 binds non-specifically with single-stranded RNA and can also interact with Nsp8 (Egloff et al., 2004; Sutton et al., 2004), thereby potentially protecting the nascent RNA genomes from degradation during replication.

When we carried out a BlastP search using the Nsp7 sequence (YP_009725303.1) translated from the SARS-CoV-2 genome MN908947.3 as the query, all homologs from diverse SARS-CoV-2 sources exhibited 100% identities along the entire 83 amino acids-long alignments. Notably, Nsp7 homologs from other SARS-related coronaviruses exhibited the single mismatch K→R at amino acid position 70, whereas no homolog of this protein from any source was found to exhibited any sequence variation that could affect the RNA-binding properties to the polymerase complex at large (for mechanistic details see Subissi et al., 2014). Nsp7 sequence from Bat coronavirus too showed 100% identity with YP_009725303.1. Similarly, BlastP analysis with the Nsp8 sequence (YP_009725304.1) from the SARS-CoV-2 genome MN908947.3 revealed 100% identities with all Nsp8 homologs from diverse SARS-CoV-2 isolates;only two amino acid mismatches along the query length of 198 residues were identified at positions 15 (F→Y) and 132 (I→V) of the Nsp8 sequence NP828866.1 and 2AHM_E from two other SARS-related CoVs respectively. The three Nsp8 residues, K58 (responsible for interaction of the polymerase complex with RNA), P183 and R190 (essential for Nsp8-Nsp12 interaction), were found to be conserved across all SARS-CoVs lineages. BlastP analysis using the Nsp9 sequence (YP_009725305.1) from the SARS-CoV-2 genome MN908947.3 as the query, revealed 100% identities with all Nsp9 homologs retrieved from diverse SARS-CoV-2 genomes, while along the query length of 113 residues, only two amino acid mismatches were noted at positions 34 (T→N) and 35 (T→S) of Nnsp9 sequences (1UW7_A and NP828867.1) of other SARS-related CoVs. In this way, the current gene-and protein-level analyses of mutations across global SARS-CoV-2 genomes revealed important filters in RNA biology of the novel coronaviruses that ensures zero point mutation accumulation in loci encoding *nsp7*. *nsp8* or *nsp9* (Tables 2, 3, 4 and S2), thereby warding off any undesired impairment of its replication machinery. The mutation spectra derived from the quasispecies structure, on one hand explains the limitation and short-life of all anti-viral vaccines innovated (Domingo and Holland, 1992), while on the other it brings to the fore novel potentials of discovering molecules aimed at shattering the processivity of SARS-CoV-2 replicase by targeting Nsp7, Nsp8 and/or Nsp9.

4.3 High frequency of C→ U transition and A→G transversion mutations

Occurrence of mutations across SARS-CoV-2 genomes is not random; instead certain types are more frequent than the others. Of the 39 instances of single nucleotide substitution detected across the 13 SARS-CoV-2 reference genomes, 29 were transition mutations while the remaining were transversions. Among all the four transition mutations possible (A→G, G→A, C→U and U→C), C→U was found to be the most prevalent one. On the other hand, among all transversion mutations detected, G→U was most frequent. Furthermore, the relatively lesser number of transversion mutations present are concentrated in loci encoding the structural proteins. These trends were also found to hold good for the 10-genome comparator dataset and the 46-genome Indian dataset.

In view of the overwhelming preponderance of cytidine to uridine transition in the global mutation spectrum of SARS-CoV-2 (as compared to all other transition/transversion mutations present) it seems likely that in the ecological context of this novel coronavirus some physicochemical and/or biochemical mutagen is more instrumental in bringing about this selective change, over and above the general replication error-induced mechanism of mutagenesis. Cytosine can convert to uracil through hydrolytic deamination, under the action of ultra-violet (UV) irradiation or the potential mediation of the enzyme activation-induced cytidine deaminase, which is abundant in mammalian cells and known to act on single-stranded nucleic acids (Bransteitter *et al*., 2003). C→U conversion is also possible chemically under the mediation of bisulfite reagents (Figure S1A; for details see Hyatsu, 2008) that are frequently used as disinfectants, antioxidants and preservative agents. Incidentally, several control techniques involving heating, sterilization, ultraviolet germicidal irradiation (UVGI) (Tseng and Li, 2007) and/or chemical disinfectants (Matallana-Surget et al., 2008) are being used currently to reduce the risk of viral infection from contaminated surfaces. Of these, intense UV-C irradiation is at the fore front of our fight against Covid-19, so indiscriminate use of the same may well accelerate the incidence of C→U mutations in global SARS-CoV-2 genomes. Furthermore, UV's specificity for targeting two adjacent pyrimidine nucleotides is long known (Miller, 1985), while in the context of DNA, UV-induced signature mutations collated from existing data on cells exposed to UVC, UVB, UVA or solar simulator light, have been confirmed as C→T in ≥ 60% dipyrimidine sites, of which again ≥ 5% is CC→TT (Brash, 2015). In consideration of the above facts it seems likely that UV irradiation is the potential cause of not only the global preponderance of C→U point mutations across SARS-CoV-2 genomes but also the low abundance of two consecutive cytidines in all lineages of this novel coronavirus. For instance, the 29,903 nucleotide RNA genome (MN908947.3) of the SARS-CoV-2 reference strain from Wuhan (China) has 22.28% of its genome in the form of two consecutive pyrimidine nucleotides (YY), with the most predominant being UU (8.15%) followed by CU (6.85%), UC (4.70%), and lastly CC (2.57%).

In the same way as envisaged for C→U mutations, environmental mutagens may also be instrumental in enhancing the rate of other deamination-based transitions such A→G prevalent across SARS-CoV-2 genomes. Of the various disinfectants widely in use for public sanitization during the ongoing Covid-19 pandemic (www.epa.gov/pesticide-registration/list-n-disinfectants-use-against-sars-cov-2), quaternary ammonium compounds such as benzalkonium chloride and dimethyldioctadecyl-ammonia bromide, being cationic surfactants, can cause genotoxic effects (Ferk et al., 2007); chlorine (as a mixture of HClO and ClO^-) can preferentially attack exocyclic-NH_2 groups of adenosine (Phe et al., 2009). Based on this chemistry, a themodynamically and kinetically favorable, three-step transition route can be conceived for A→G via hydrolytic deamination involving water and formic acid (an age old biocide and ensilation agent of forage crops that is abundant in the environment) and ammonia (Figure S1B; for details see Tolosa et al., 2019).

4.4 Conclusion

The present study conducted a comparative genomic investigation of SARS-CoV-2 quasispecies isolated and sequenced from globally distinct geographies. Mutation spectrum revealed through genotypic analysis highlighted the frequent occurrence of transversion mutations in the structural proteins that might have conferred fast transmissibility of SARS-CoV-2. The study further revealed a mutation-free genomic region coding for Nsp7, Nsp8, and Nsp9, which provide optimum fidelity attuned to maintaining the consensus genome sequence of this novel coronavirus. Phylogenomics pointed out discreteness and parallelism among the evolutionary paths followed by the distinct quasispecies clusters of the virus. Receptivity of major segments of the SARS-CoV-2 global genome towards innovation, in conjunction with the widespread exposure of the virus to UV radiation and other disinfecting agents, constitutes a complex driver of emergence of newer variant strains.

Supplementary Material

One MS Word file named Supplementary_Information containing Figure S1, and one MS Excel file named Supplementary_Dataset containing Tables S1 and S2, accompany this paper.

Author Contributions

RC conceived the study, designed the experiments, interpreted the results and wrote the paper with the help of WG. SM, SKM and SM performed the bioinformatic experiments, and checked and vetted the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgement

We thank Ms. Riddhi Chakraborty, Raja Rammohanpur, West Bengal, India for copy editing the manuscript.

References

Altschul, S.F., Gish, W., Miller, W.,Myers, E. W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215(3), 403-410. http:// doi: 10.1016/S0022-2836(05)80360-2.

Benvenuto, D., Angeletti, S., Giovanetti, M., et al. 2020. Evolutionary analysis of SARS-CoV-2: How mutation of non-structural protein 6 (NSP6) could affect viral autophagy. J. Infect. S0163-4453 (20), 30186-9. http://doi: 10.1016/j.jinf.2020.03.058.

Bransteitter, R., Pham, P., Scharff, M.D., Goodman, M.F. 2003. Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. Proc. Natl. Acad. Sci. USA. 100 (7) 4102-4107; https://doi.org/10.1073/pnas.0730835100.

Brash, D.E., 2015. UV signature mutations. Photochem. Photobiol. 91(1),15-26. http://doi:10.1111/php.12377.

Domingo, E., Holland, J.J., 1992. Complications of RNA heterogeneity for the engineering of virus vaccines and antiviral agents, in: Setlow, K.K., (Ed.), Genetic engineering, principles and methods, Plenum press, New York, NY , vol.14, pp.13-32.

Egloff, M.P., Ferron, F., Campanacci, V., et al. 2004. The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA binding subunit unique in the RNA virus world. Proc. Natl. Acad. Sci. U SA. 101(11), 3792-3796. https://doi.org/10.1073/pnas.0307877101.

Eigen, M., McCaskill, J., Schuster, P., 1988. Molecular quasispecies. J. Phys. Chem. 92, 6881–6891. https://doi.org/10.1021/j100335a010.

Felsenstein, J., 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution. 39:783-791. doi: 10.1111/j.1558-5646.1985.tb00420.x.

Ferk, F., Misík, M., Hoelzl, C., et al. 2007. Benzalkonium chloride (BAC) and dimethyldioctadecyl-ammonium bromide (DDAB), two common quaternary ammonium compounds, cause genotoxic effects in mammalian and plant cells at environmentally relevant concentrations. Mutagenesis. 22(6), 363-70. https://doi.org/10.1093/mutage/gem027.

Gao Y., Yan L., Huang, Y., et al. 2020. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science. 368: 779-782. DOI: 10.1126/science.abb7498.

Green, A., 2020. Li wenliang. The Lancet, 395(10225), 682. http://doi.org/10.1016/S0140-6736(20)30382-2.

Hayatsu, H., 2008. Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis- A personal account. Proc. Jpn. Acad. Ser. B Phys. Biol. Sci. 84: 321-330. DOI: 10.2183/pjab/84.321.

Hurlbert, R.B., Kammen, H.O., 1960. Formation of Cytidine Nucleotides from Uridine Nucleotides by Soluble Mammalian Enzymes: Requirements for Glutamine and Guanosine Nucleotides. J. Biol. Chem. 235, 443-449.

Jukes, T.H. and Cantor, C.R., 1969. Evolution of Protein Molecules. In: Munro, H.N., Ed., Mammalian Protein Metabolism,Academic Press,New York, 21-132.http://dx.doi.org/10.1016/B978-1-4832-3211-9.50009-7

Kirchdoerfer, R.N., Ward, A.B., 2019. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. Nat. Commun. 10(1), 2342. http://doi.org/10.1038/s41467-019-10280-3.

Kumar, S., Stecher. G., Tamura, K., 2016. MEGA 7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33(7), 1870-1874. http://doi: 10.1093/molbev/msw054.

Liu, P., Chen, W., Chen, J.P., 2019. Viral metagenomics revealed Sendai Virus and coronavirus infection of Malayan Pangolins (*Manis javanica*). Viruses. 11(11), 979.http://doi: 10.3390/v11110979.

Lu, R., Zhao, X., Li, J., et al., 2020. Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 395,565-74. https://doi.org/10.1016/ S0140-6736(20)30251-8.

Matallana-Surget, S., Meador, J., Joux, F., Douki, T., 2008. Effect of the GC content of DNA on the distribution of UVB-induced bipyrimidine photoproducts. Photochem. Photobiol. Sci. 7(7), 794-801. http://DOI: 10.1039/b719929e.

Miknis, Z.J., Donaldson, E.F., Umland, T.C., et al. 2009. Severe acute respiratory syndrome coronavirus nsp9 dimerization is essential for efficient viral growth. J. Virol. 83 (7), 3007–3018.http:// doi: 10.1128/JVI.01505-08.

Miller, J.H., 1985. Mutagenic specificity of ultraviolet light. J. Mol. Biol. 182, 45–68.

Paraskevis, D., Kostaki, E.G., Magiorkinis, G., et al. 2020. Full genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infect. Genet. Evol. 79, 104212. https://doi.org/10.1016/j.meegid.2020.104212.

Phe, M.H., Hajj, M.C., Guilloteau, H., et al. 2009. Assessment of damage to nucleic acids and repair machinery in *Salmonella typhimurium* exposed to chlorine. Int. J. Microbiol. 2009, 201868. http:// doi: 10.1155/2009/201868.

Saitou, N., Nei, M., 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4(4), 406-425. https://doi.org/10.1093/oxfordjournals.molbev.a040454.

Subissi, L., Posthuma, C. C., Collet, A., et al., 2014. One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. Proc. Natl. Acad. Sci. USA. 111 (37), E3900-E3909. https://doi.org/10.1073/pnas.1323705111.

Sutton, G., Fry, E., Carter, L., et al. 2004. The nsp9 replicase protein of SARS-coronavirus, structure and functional insights. Struct. Lond. Engl. 12, 341–353. http://doi:10.1016/j.str.2004.01.016.

te Velthuis, A.J,W., Arnold, J.J., Cameron, C.E., et al. 2010. The RNA polymerase activity of SARS-coronavirus nsp12 is primer dependent. Nucleic. Acids. Res. 38(1):203–214.http:// doi: 10.1093/nar/gkp904.

Thompson, J.D., Higgins, D.G., Gibson. T.J., 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(22), 4573-4680. http://doi: 10.1093/nar/22.22.4673.

Tolosa, S., Sanson, J.A., Hidalgo, A., 2019. Theoretical study of adenine to guanine transition assisted by water and formic acid using steered molecular dynamic simulations. Front. Chem. 7: 414. http://doi.org/10.3389/fchem z019.00414.

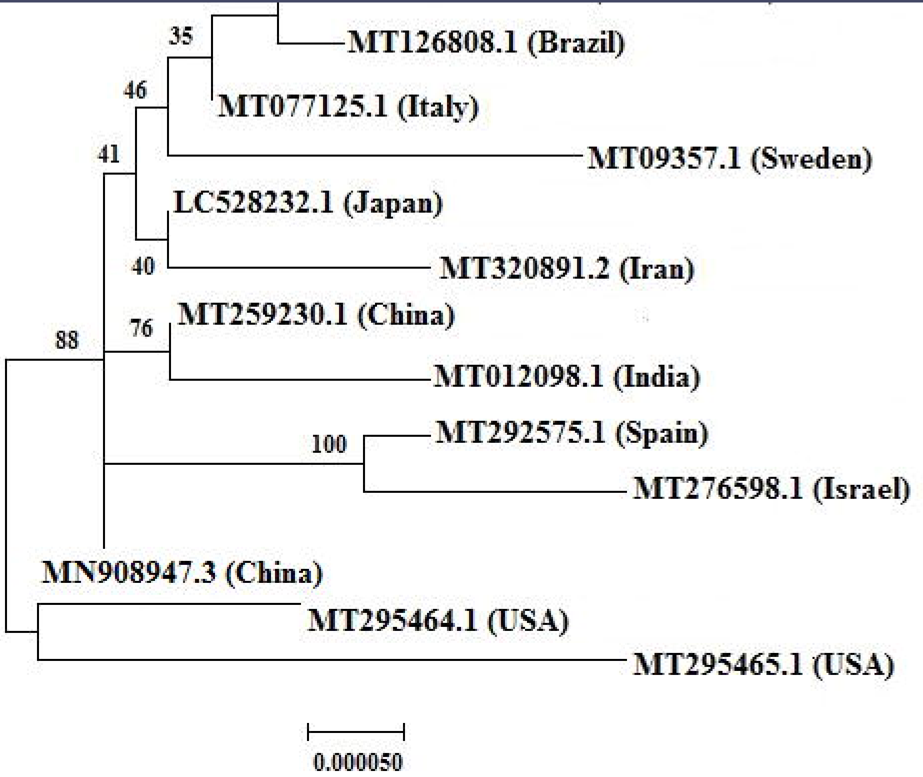
Tseng, C.C., Li, C-S., 2007. Inactivation of viruses on surfaces by ultraviolet germicidal irradiation. J. Occup. Environ. Hyg. 4(6), 400-5. http://doi: 10.1080/15459620701329012.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., et al., 2020. A new coronavirus associated with human respiratory disease in China. Nature. 579(7798), 265-269. http://doi: 10.1038/s41586-020-2008-3.

Zhou, YX-L.P., Wang, X-G., Hu, B., et al. 2020. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. bioRxiv. 2020. Epub Januar 23(2020). https://doi.org/10.1101/2020.01.22.914952.

Figure Legends

Figure 1. The evolutionary history of 13 Pan-SARS-CoV-2 genome sequences inferred using Neighbor Joining algorithm. The optimal tree with the sum of branch length = 0.23622660 has been shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) is shown next to each branch. The scale bar denotes the number of nucleotides substituted per site considered in the final alignment.



35
MT126808.1 (Brazil)
46
MT077125.1 (Italy)
41
MT09357.1 (Sweden)
LC528232.1 (Japan)
40
MT320891.2 (Iran)
MT259230.1 (China)
88
76
MT012098.1 (India)
MT292575.1 (Spain)
100
MT276598.1 (Israel)
MN908947.3 (China)
MT295464.1 (USA)
MT295465.1 (USA)
0.000050

Table 1: Summary of the 13 sets of 100 best matches each ontained from BlastN searches condced independently with the 13 SARS-CoV-2 reference genome sequences (a total of 13 x 100 = 1300 alignments were obtained in this way).

Sl. No.	**Accession no. of the genome (Country / geography of origin)**	**Size of the genome (no. of bases)**	**No. of genomes matched with 100% query coverage**	**No. of genomes matched with 99% query coverage**	**No. of genomes matched at 100% identitity**	**No. of genomes matched at 99.99% identitity**	**No. of genomes matched at 99.98% identitity**	**No. of genomes matched at 99.97% identitity**	**No. of genomes matched at 99.96% identitity**	**No. of genomes matched at 99.95% identitity**	**No. of genomes matched at 99.94% identitity**	**No. of genomes matched at 99.93% identitity**
1.	MN908947.3 (China)	29,903	37	63	03	60	36	-	-	01	-	-
2.	LC528232.1 (Japan)	29,902	01	99	22	71	05	01	01	-	-	-
3.	MT295464.1 (USA)	29,892	44	56	24	42	29	05	-	-	-	-
4.	MT295465.1 (USA)	29,893	56	44	01	15	05	03	40	34	02	-
5.	MT093571.1 (Sweden)	29,886	43	57	01	-	16	57	25	-	-	01
6.	MT304476.1 (Korea)	29,882	92	08	03	07	58	30	01	01	-	-
7.	MT126808.1 (Brazil)	29,876	85	15	08	32	56	02	01	01	-	-
8.	MT259230.1 (China)	29,866	85	15	03	59	37	-	-	01	-	-
9.	MT012098.1 (India)	29,854	93	07	01	-	-	66	33	-	-	-
10.	MT276598.1 (Israel)	29,870	01	99	09	24	66	01	-	-	-	-
11	MT320891.1 (Iran)	29,822	99	01	01	06	84	07	01	01	-	-
12	MT077125.1 (Italy)	29,785	97	03	06	89	05	-	-	-	-	-
13	MT292575.1 (Spain)	29,782	100	00	01	25	74	-	-	-	-	-

Table 2. Gene-wise localization and molecular character of all the point mutations that were identified in one or more of the 13 SARS-CoV-2 reference genomes considered in this study.

Serial no.	Nucleotide position* at which mutation has occurred	Identity of the corresponding gene	Nature of the mutation	Serial no.	Nucleotide position* at which mutation has occurred	Identity of the corresponding gene	Nature of the mutation	Serial no.	Nucleotide position* at which mutation has occurred	Identity of the corresponding gene	Nature of the mutation
1.	313	Nsp1	C→U	16.	17247	Nsp13	U→C	31.	24694	S	A→U
2.	1397	Nsp2	G→A	17.	17373	Nsp13	C→U	32.	26084	ORF3a	G→U
3.	2277	Nsp2	U→C	18.	17376	Nsp13	A→G	33.	26144	ORF3a	G→U
4.	2717	Nsp2	G→A	19.	17747	Nsp13	C→U	34.	26729	ORF6	U→C
5.	3037	Nsp3	C→U	20.	17858	Nsp13	A→G	35.	28251	ORF8	C→U
6.	3177	Nsp3	C→U	21.	18060	Nsp14	C→U	36.	28821	N	G→A
7.	5572	Nsp3	G→U	22.	18377	Nsp14	C→U	37.	28822	N	G→A
8.	6695	Nsp3	C→U	23.	18736	Nsp14	U→C	38.	28823	N	G→C
9.	8782	Nsp4	C→U	24.	20268	Nsp15	A→G	39.	29374	N	G→A
10.	9274	Nsp4	A→G	25.	21575	S	C→U	40.	29742	UTR	G→U
11.	11083	Nsp6	G→U	26.	21993 - 21995	S	AUU-deletion				
12.	11557	Nsp6	G→A	27.	22785	S	G→U				
13.	13167	Nsp10	C→G	28.	23403	S	A→G				
14.	14657	Nsp12	C→U	29.	23952	S	U→G				
15.	14805	Nsp12	C→U	30.	24034	S	C→U				

* Nucleotide positions have been designated based on the 5’ to 3’ sequence of the 29,903 nucleotide genome (MN908947) of the reference strain from Wuhan, China. Underlined nucleotide positions denote mutations that were identified in more than one sequence of the dataset.

doi:10.20944/preprints202005.0463.v1

Table 3. Gene-wise presence/absence of the global point mutations identified in the aligned clusters of genomes belonging to the 13 reference§, 10 validation$^{\#}$, and 46 Indian†, strains of SARS-CoV-2.

	No. of point mutations detected globally	**No. of nucleotides positions mutated per nucleotide aligned**	**Genes encoding non-structural proteins (*nsp* genes$^{\Delta}$)**																**Genes encoding structural proteins**										
			1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***gene S***¶	***orf 3a***	***gene E***$^{\$}$	***gene M***$^{\clubsuit}$	***orf 6***	***orf 7a***	***orf 8***	***gene N***$^{@}$	***orf 10***	**UTR**$^{¥}$
13 reference genomes§	42	1.08×10^{-4}	+	+	+	+	−	+	−	−	−	+	−	+	+	+	+	−	+	+	−	−	+	−	+	+	−	+
10 comparator genomes$^{\#}$	21	0.70×10^{-4}	−	+	+	+	+	−	−	−	−	−	−	+	+	−	−	−	+	+	−	−	−	−	−	−	+	−
46 Indian genomes†	141	1.03×10^{-4}	+	+	+	+	+	+	−	−	−	−	−	+	+	+	+	+	+	+	−	+	−	+	+	+	+	−

§ These strains have been isolated from 11 different countries (Brazil, China, India, Iran, Israel, Italy, Japan, Korea, Spain, Sweden, USA).
# These strains have been isolated from eight different countries (Brazil, China, India, Japan, Sri Lanka, Taiwan, USA, Vietnam).
† These strains have been isolated from three Indian states: Gujarat (26), Karnataka (14) and Andhra Pradesh (06).
Δ nsp genes are numbered as 1 to 16.
¶ *geneS* encodes the spike protein of SARS-CoV-2.
$ *geneE* encodes the envelop of SARS-CoV-2.
♣ *geneM* encodes a membrane glycoprotein.
@ *geneN* encodes the nucleocapsid phosphoprotein.
¥ Untranscribed region.

Table 4. Gene-wise localization and molecular character of all the point mutations that were identified in one or more of the 10 SARS-CoV-2 genomes of the comparator dataset.

Serial number	Nucleotide position* at which mutation has occurred	Identity of the corresponding gene	Nature of the mutation
1.	2277	*nsp2*	U→C
2.	3037	*nsp3*	C→U
3.	6695	*nsp3*	C→U
4.	8001	*nsp3*	A→C
5.	9034	*nsp4*	A→G
6.	9499	*nsp4*	C→U
7.	9534	*nsp4*	C→U
8.	10232	*nsp5*	C→U
9.	11083	*nsp6*	G→U
10.	14657	*nsp12*	C→U
11.	14804	*nsp12*	C→U
12.	17247	*nsp13*	U→C
13.	17373	*nsp13*	C→U
14.	21895	SPIKE	U→Y[§]
15.	21993 to 21995	SPIKE	AUU- deletion
16.	22785	SPIKE	G→U
17.	26111	Orf3a	C→U
18.	26144	Orf3a	G→U
19.	29635	Orf10	C→U

[*] Nucleotide positions have been designated based on the 5’ to 3’ sequence of the 29,854 nucleotide genome (MT012098.1) of the Indian strain that is the sole common entry between this comparator dataset and the initial dataset of 13 reference genomes. Underlined nucleotide positions denote mutations that were identified in more than one sequence of this comparator dataset.

[§] Y stands for an undefined pyrimidine base in the sequence, at that position.