 doi:10.20944/preprints202005.0463.v2

Trends of mutation accumulation across global SARS-CoV-2 genomes: Implications for the ecology and evolution of the novel coronavirus

Chayan Roy[1], Santi M. Mandal[2], Suresh K. Mondal[2], Shriparna Mukherjee[3], Wriddhiman Ghosh[4] and Ranadhir Chakraborty[5,*]

[1] College of Veterinary Medicine, Western University of Health Sciences, 309 East Second Street, Pomona, CA, 91766, USA
[2] Central Research Facility, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India.
[3] Department of Botany, Prasannadeb Women's College, Jalpaiguri, West Bengal, India.
[4] Department of Microbiology, Bose Institute, P-1/12 CIT Scheme VII M, Kolkata 700054, West Bengal, India.
[5] 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: Microevolution of SARS-CoV-2
Key words: SARS-CoV-2, COVID-19, genome-wide mutations, transition and transversion, microevolution, disinfectants as mutagens

Abstract

The all-pervasiveness and dynamic nature of the COVID-19 pandemic warrants comprehensive and constant surveillance of the numerous mutations that are accumulating in global SARS-CoV-2 genomes and contributing to the microevoution of the various lineages of the novel coronavirus. This would help

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

us gain insights into the evolving pathogenicity of the virus, and thereby improvise our control and therapeutic strategies. Here we explore the genome-wide frequency, gene-wise distribution, and molecular nature, of the large repertoire of point mutations detected across the global dataset of 3,608 SARS-CoV-2 RNA-genomes short-listed from a total 5,485 whole genome sequences deposited in GenBank till 4 June 2020 using a download filter that eliminated all incomplete/gapped sequences. Phylogenomic analysis involving all existing SARS-CoV-2 lineages, represented by 3,740 whole genome sequences from human-source (out of a total of 63,894 sequences stored in the GISAID repository, as on 15 July, 2020), illustrated that the two major-lineages of the virus contributed almost equivalently to the pandemic. However, entities belonging to the early lineages are still mostly spread over Asian countries, whereas those affiliated to recently-derived lineages have a relatively more global distribution. Mutation frequency in the SARS-CoV-2 pan genome was found to be 2.27×10^{-5} nucleotide positions mutated per nucleotide analyzed. An overwhelming majority (count: 1,797) of the total 2,452 instances of single nucleotide substitution detected (in the SARS-CoV-2 pan genome) were found to be transition mutations with cytidine to uridine (C→U) being the most ubiquitous molecular-type (count: 987). Of the 655 instances of transversion detected, the guanosine to uridine (G→U) variant was most widespread (count: 367). All transcribed and untranscribed loci of the pan genome were found to contain mutation(s). *nsp3*, and *S*, *N* and *orf3a*, were the most point-mutation-ridden non-structural and structural protein-coding genes, respectively, with 435, 300, 171 and 128 total mutations; 349/86, 192/108, 107/64 and 76/52 transitions/transversions; and 189/48, 106/55, 59/42 and 43/31 C→U/G→U substitutions, respectively. Potential mechanistic backgrounds were envisaged for the molecular bias of mutations observed in SARS-CoV-2.

Introduction

On 30 Dec 2019, ophthalmologist Li Wenliang in Wuhan, Hubei province, China, first recognized and communicated about the outbreak of a contagious illness

 doi:10.20944/preprints202005.0463.v2

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 216 countries, infecting 14,562,550 people, and killing 607,781 as of 22 July 2020 (https://covid19.who.int). The first whole genome sequence of SARS-CoV-2 was deposited in 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). 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) indicate SARS-CoV-2 to be a zoonotic disease (Lu et al., 2020). However, human to human transmission of SARS-CoV-2 is also 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 Italy, Spain, France, UK and Iran, and then in other parts of Western Europe, USA, Brazil, Russia, Southeast Asia, South Asia and Middle East.

At the same time as the scientific community is racing to develop vaccines and therapeutics against COVID-19 (Folegatti et al., 2020), the virus on its part is busy accumulating mutations across its pan genome, some of which may well help it evade clinical interventions (Benvenuto et al., 2020; Korber et al., 2020; Saha et al., 2020; Wang et al., 2020). In this context of SARS-CoV-2 evolution, the present study analyzes 3,608 whole genome sequences of this novel coronavirus, focusing on representative quasi-species isolated from 47 different countries (or distinct geographies), to delineate SARS-CoV-2 phylogeny, and then identify the distribution pattern of global point mutations in the SARS-CoV-2

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

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; Kimsey et al., 2018) revealed potential mechanisms that could be instrumental in accelerating mutation recruitment in the RNA genome of SARS-CoV-2.

Experimental procedure

Comparative genomics

Of the 5,485 SARS-CoV-2 whole genome sequences deposited in GenBank till 4 June 2020, an overwhelming percentage (62.9%) were from USA, followed by Australia (16.97%), Thailand (4.1%), India (>3.5%), China (>2.8%), Greece (1.78%) and France (1.47%). Genome sequences of this novel coronavirus have also been deposited from other countries (or distinct geographies) such as Bangladesh, Belgium, Brazil, Colombia, Czech Republic, Finland, Germany, Guam, Hong Kong, Iraq, Israel, Italy, Jamaica, Japan, Kazakhstan, Kenya, Malaysia, Morocco, Nepal, Netherlands, Nigeria, Pakistan, Peru, Philippines, Poland, Puerto Rico, Russia, Serbia, South Africa, South Korea, Spain, Sri Lanka, Sweden, Taiwan, Tunisia, Turkey, Uruguay, Uzbekistan and Vietnam. From these, a total of 3,608 whole genome sequences of SARS-CoV-2 strains were downloaded from the NCB-SARS-CoV-2 database (https://www.ncbi.nlm.nih.gov/sars-cov-2/) on 4 June 2020, using a download filter that eliminated all incomplete or gapped genome sequences from the curated dataset. Sequence similarity between the 3,608 genomes was computed in a pairwise manner involving all the 6,507,028 combinations possible, using FastANI, a high throughput method for average nucleotide identity analysis (Jain et al., 2018). The software MAFFT (Katoh et al., 2002) was used with default options to align all the whole genome sequences included in the dataset. Nucleotide positions involving polymorphisms (base substitutions) were identified in the individual genomes using the software *SNP-sites* (Page et al., 2016), which specifically identifies single nucleotide polymorphisms (SNPs) from aligned multi-fasta sequence files. Subsequently, the variant call format (VCF) file

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

generated from the *SNP-site* analysis was processed using the software *VCFtools* (Danecek et al., 2011) to enumerate all transition and transversion events within the entire dataset of aligned whole genome sequences. The results obtained from *VCFtools* analysis were further compared and validated with those obtained from the alignment-free SNP-detection tool kSNP3 (Gardner et al., 2015).

Evolutionary relationship between the existing SARS-CoV-2 lineages was delineated in the form of a phylogenetic tree constructed and downloaded alongside the associated metadata, on 15 July 2020, from the website https://nextstrain.org/ncov/, which contains all epidemiological records for the ongoing COVID-19 outbreak. The genomic data utilized in this service is sourced from Global Initiative on Sharing All Influenza Data (GISAID; https://www.gisaid.org/) which enables rapid data sharing for different contagious diseases. The data downloaded contained information for 3,746 representative SARS-CoV-2 genomic sequences derived from a total of 63,894 sequences stored in the GISAID repository, as on 15 July, 2020. Out of the 3,746 representative sequences encompassed in the downloaded data, six were from non-human sources, so were removed from the final dataset. Branches of the phylogenetic tree obtained were labeled based on the geographical origin of the sequences in addition to three different concepts of clade nomenclature that are in use for the ongoing COVID-19 outbreak, namely (i) the dynamic clade nomenclature system PANGOLIN (Rambaut et al., 2020), (ii) Year-Letter nomenclature system proposed by Hodcroft et al. (https://nextstrain.org/blog/2020-06-02-SARSCoV2-clade-naming), and (iii) the system proposed by Tang et al. (2020) (https://academic.oup.com/nsr/article/7/6/1012/5775463), and followed by GISAID, which names major clades based on nine distinct marker mutations spread over 95% of the known SARS-Cov-2 diversity.

Results and Discussions

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

Small but phylogenetically significant divergences accumulating in global SARS-CoV-2 genomes

To delineate mutual sequence similarities existing between the 3,608 SARS-Cov-2 genomes, we calculated average nucleotide identity (ANI for Kmer size of 16 over a 1,000 base fragment size) and coverage level for all the pairwise combinations possible. In the process, an overwhelming number of genome-pairs (2,890,861 out of the total 6,507,028 pair-wise combinations possible) were found to possess 99.97-100.00% sequence similarities alongside 99-100% coverage. On the other hand, 2,437,727 and 1,178,440 genome-pairs had 99.93-99.96% and ≤ 99.93% sequence similarities respectively.

Potential evolutionary paths traced thus far by SARS-CoV-2 in time and space were delineated by labeling the phylogenetic tree obtained from https://nextstrain.org/ncov/, on the basis of three different concepts of clade nomenclature (Figure 1A-1C). Information regarding geographical origin of the sequences analyzed was also used to label the tree (Figure 1D). Figure 1A, where the tree topology was labeled according to the dynamic clade nomenclature system (Rambaut et al. 2020) called **P**hylogenetic **A**ssignment of **N**amed **G**lobal **O**utbreak **LIN**eages (PANGOLIN), reflected the global preponderance and ancestral status of the members of the SARS-CoV-2 clade A by virtue of the association of its sub-lineages (namely A3 and A6) with the lineage O encompassing other betacoronaviruses (Clade A is epitomized by the genome LR757995, which was isolated from Wuhan on 26 December 2019, sequenced, and submitted to GenBank on 30 January 2020). This nomenclatural approach also illustrated the clear divergence of Clade A from the other SARS-CoV-2 major-clade named B, the typical representative (MN908947.3) of which was also isolated from Wuhan on 26 December 2019, but submitted to GenBank on 12 January 2020 as MN908947.1. Albeit the genome sequence MN908947 was deposited at an earlier date, the clade it represents (B) has apparently diverged at a later stage of evolution from Clade A alongside the other A-derived linages A1a and A7.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

On the other hand, Figure 1B, where branches of the phylogenetic tree have been labeled according to the Year-Letter nomenclature system (i.e., with the year of identification followed by an alphabet) of Hodcroft et al., 2020 (https://nextstrain.org/blog/2020-06-02-SARSCoV2-clade-naming), showed that the largest lineage A2 identified by PANGOLIN clade-nomenclature system, emerged in the year 2020 and evolved further into a number of sub-lineages characterized by mutations in specific nucleotide positions (these have been designated in XB as branches 20A, 20B, 20C, etc.). This system, which names new major clades only when the frequency of a clade exceeds 20% in a representative global sample and that clade differs in at least two nucleotide positions from its parent clade, also corroborated the early (i.e. 2019) advent of the ancestral lineages of the PANGOLIN clade A, alongside their derivatives which formed PANGOLIN Clade B.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

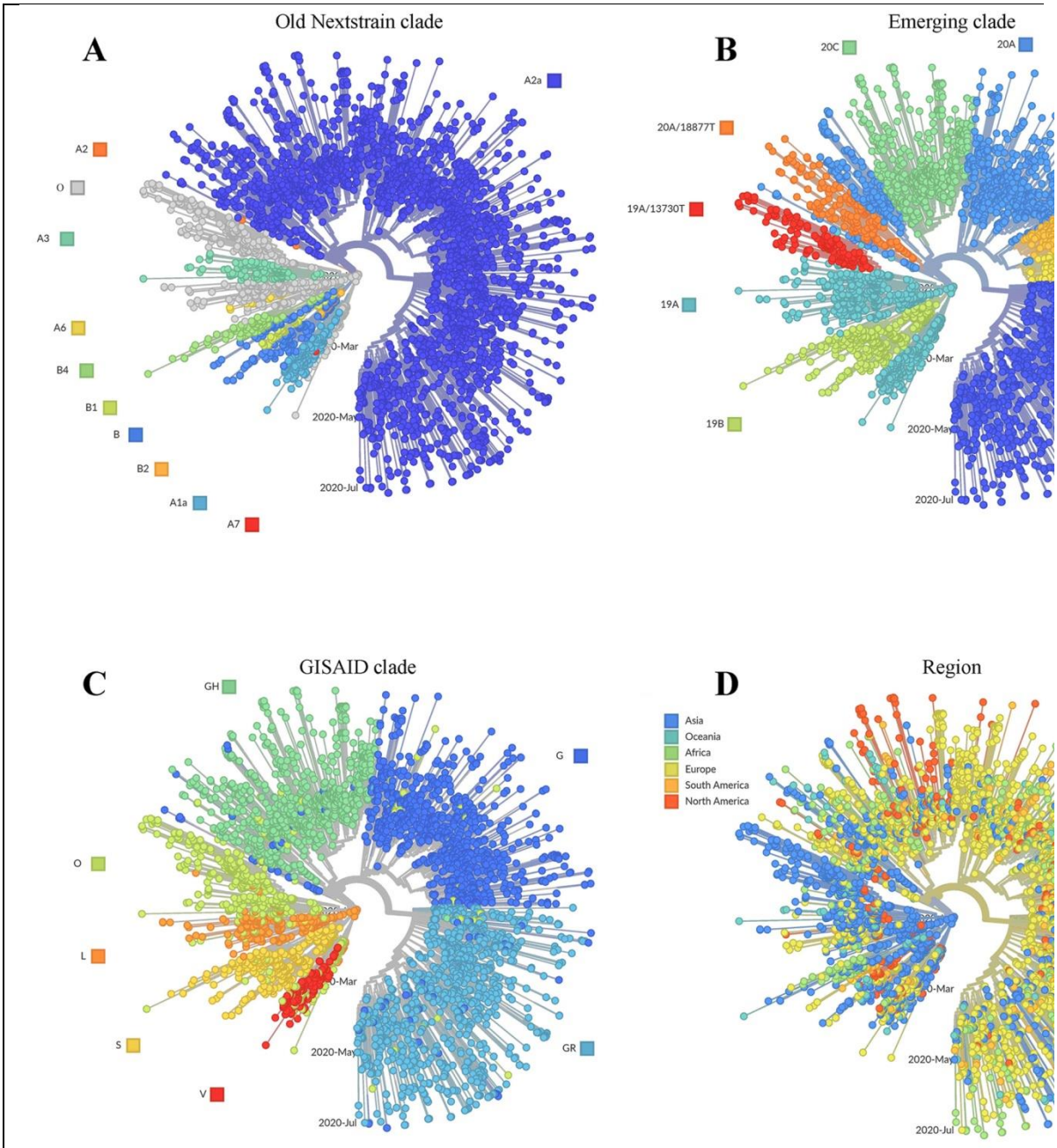

Figure 1. Phylogenetic relationship between SARS-COV-2 and related betacoronavirus entities related to human (n = 3,740). (**A)** identifies and labels the clades based on the dynamic clade nomenclature system PANGOLIN (Rambaut et al. 2020). This convention currently defines 57 evolved lineages based on shared mutations, of which 10 initially-described lineages (old Nextstrian Clades) have been shown. (**B)** identifies and labels the clades based on Year-Letter naming as per the nomenclature system proposed by Hodcroft et al. (https://nextstrain.org/blog/2020-06-02-SARSCoV2-clade-naming). (**C)**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

identifies and labels the clades based on the nomenclature system proposed by Tang et al. (https://academic.oup.com/nsr/article/7/6/1012/5775463) and which is also followed by GISAID. (**D)** labels the entities analyzed based on the geographical origin.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

Consistent with the above phylogenetic interpretations, labeling of the tree with the third clade-nomenclature convention proposed by Tang et al. (2020) and also followed by GISAID, indicated that the two original lineages, named as S and L (essentially equivalent to 19A and 19B of the Year-Letter nomenclature system), has diversified and thus far given rise to a total of seven clades, based on nine distinct marker mutations spread over 95% of the known SARS-Cov-2 diversity (Figure 1C). As per the data available till 15 July 2020, Clade L is apparently more populous than Clade S, and has diversified further into V and G, with G splitting further into G, GH and GR (essentially equivalent to the old A2a clade of PANGOLIN, or the 20A, 20C and 20B of Year-Letter, nomenclature systems).

Labeling of the phylogenetic tree on the basis of the geographical origin of the sequences showed that members of the original and early-diverged clades (S and L, and V, respectively) are still mostly spread over Asian countries, whereas the recently derived clades (G, GH and GR) are distributed across the globe, especially in Europe and North America (Figure 1D).

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

Alignment of the 3,608 SARS-CoV-2 whole genome sequences investigated in this study revealed that over the 29,903 completely aligned nucleotide positions, 2,452 instances of single nucleotide substitution have occurred in one or more genomes at a frequency of $[2,452 \div (29,903 \times 3,608)] = 2.27 \times 10^{-5}$ mutated nucleotide positions per nucleotide participating in the alignment. Tables S1 and S2 show the nucleotide positions of all the 2,452 global mutations with reference to the 5' to 3' sequence of the 29,903-nucleotide-long genome (MN908947.3) of the earliest-sequenced reference strain from Wuhan, China. Gene-wise mapping of these global mutations encountered across the dataset of 3,608 genomes (Tables 1 and 2) revealed that all 25 SARS-CoV-2 genes, as well as the two untranscribed regions (UTRs), contained mutation(s) in one or more reference genomes. Out of the 2,452 global substitution mutations, 1,581 mutations were

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

found to be located between nucleotide positions 266 and 21,555 (with reference to the 1 - 29,903 positions of MN908947.3), within the foremost locus of SARS-CoV-2 genome that encodes the 16 non-structural proteins Nsp1 through Nsp16. On the other hand, 781 mutations have occurred within the nine structural protein-encoding genes *S*, *orf3a*, *E*, *M*, *orf6*, *orf7a*, *orf8*, *N* and *orf10*, which are located between nucleotide positions 21,563 to 29,674 (with reference to MN908947.3); the rest of the mutations are within the two UTRs located in the 5' and 3' ends of the SARS-CoV-2 genome. Furthermore, all the Nsp-encoding genes, except *nsp11*, have at least 12 mutations each identified across the 3,608 global SARS-CoV-2 genomes. *nsp11*, in turn, was found to be devoid of any mutation across the 3,608 genomes investigated, except one G→U transversion in a single strain from USA. The frequency of mutated (base substituted) positions within the SARS-CoV-2 genomic locus dedicated to encoding non-structural proteins was found to be 0.07 (1,581 ÷ 21,290 nucleotide positions), while that of mutated positions within the genomic locus dedicated to encoding structural proteins was 0.1 (781 ÷ 8,112 nucleotide positions). In contrast, the frequency of mutated positions within the two UTRs was double, i.e, 0.18 (90 ÷ 493 nucleotide positions). Genes-wise, the loci for *nsp1* and *orf3a*, happened to be the most mutation-prone non-structural and structural gene, as they had 0.06 and 0.08 mutations per nucleotide position encompassed, respectively. Numerically, maximum mutations were found to be located in the genes encoding Nsp3 (total 435 mutations in the SARS-CoV-2 pan genome) and the spike protein S (300 mutations in the SARS-CoV-2 pan genome).

High frequency of C→U (transition) and G→U (transversion) mutations across global SARS-CoV-2 genomes

Of the 2,452 single nucleotide substitution mutations encountered globally in SARS-CoV-2 genomes, 1,797 and 655 were transition and transversion mutations respectively. In this way, a transition:transversion ratio of 2.74 characterized the nucleotide substitution bias of the SARS-CoV-2. In other words, the rate of transition mutations in SARS-CoV-2 is higher than what is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

expected if transition and transversion events took place randomly. Of the 1,797 positions of the SARS-CoV-2 pan genome undergoing transition mutation, an overwhelming number (count: 987) featured C→U conversion; this was 54.9% of the total number of transition mutations encountered in the pan genome (Table 1). Interestingly again, the frequency of transition mutations within the SARS-CoV-2 pan genome locus harboring the non-structural protein-coding genes was 0.058 (1,239 ÷ 21,290 nucleotide positions), while the corresponding statistic for the locus harboring the structural protein-encoding genes was 0.062 (503 ÷ 8,112 nucleotide positions). In contrast, the frequency of transition mutations within the two UTRs of the SARS-CoV-2 pan genome was almost double, i.e, 0.11 (55 ÷ 493 nucleotide positions). Genes-wise, the loci encoding the Nsp3 and S proteins were most transition mutation-affected (349 and 192 respectively); these events in turn were replete with C→U substitutions (189 and 106 respectively) (Table 1). On the other hand, frequency of transition mutation per nucleotide of a gene was highest for *orf7a* (0.09), *orf3a* (0.09) and *geneN* (0.08). Out of the total number of transition events that have taken place in these five genes, 54.15% (i.e., 189 out of 349), 55.2% (106 out 0f 192), 67.6% (i.e. 23 out of 34), 56.57% (43 out of 76) and 55.14% (59 out of 107) were C→U substitutions respectively (Table 1).

Table 1. Gene-/locus-wise frequency of the different types of transition mutation encountered in the SARS-CoV-2 pan genome.

Genomic locus (nucleotide positions[1] between which the locus spans)	Number of A→G substituted positions detected	Number of G→A substituted positions detected	Number of C→U substituted positions detected	Number of U→C substituted positions detected	Total number of nucleotide positions involving transition mutations (N)	N / length of the locus
5/ UTR (1-265)	03	04	19	04	30	0.11
nsp1 (266-806)	07	14	16	10	47	0.09
nsp2 (807-2720)	23	33	80	16	152	0.08
nsp3 (2720-8555)	61	43	189	56	349	0.06
nsp4 (8556-10055)	20	11	47	14	92	0.06
nsp5 (10056-10973)	12	05	31	09	57	0.06

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

nsp6 (10974-11843)	09	09	24	10	52	0.06
nsp7 (11844-12092)	04	02	06	05	17	0.07
nsp8 (12093-12686)	05	07	20	02	34	0.06
nsp9 (12687-13025)	02	04	15	02	23	0.07
nsp10 (13026-13442)	02	00	10	04	16	0.04
nsp11 (13443-13481)	00	00	00	00	00	0.00
nsp12 (13482-16328)	22	15	73	27	137	0.05
nsp13 (16329-18041)	21	08	55	12	96	0.06
nsp14 (18042-19622)	10	03	52	14	79	0.05
nsp15 (19623-20660)	05	07	26	10	48	0.05
nsp16 (20661-21554)	06	05	22	07	40	0.04
geneS (21563-25384)	32	17	106	37	192	0.05
orf3a (25393-26220)	11	06	43	16	76	0.09
geneE (26245-26472)	01	01	07	01	10	0.04
geneM (26523-27191)	02	05	29	05	41	0.06
orf6 (27202-27387)	03	00	07	04	14	0.08
orf7a (27394-27759)	06	01	23	04	34	0.09
orf8 (27894-28259)	06	03	06	05	20	0.05
geneN (28274-29533)	10	19	59	19	107	0.08
orf10 (29558-29674)	00	01	08	00	09	0.08
3′ UTR (29675-29903)	02	04	14	05	25	0.11
Entire pan genome	285	227	987	298	1797	-

[1] Nucleotide positions between which the locus spans have been designated based on the 5’ to 3’ sequence of the 29,903 nucleotide RNA-genome (MN908947) of the reference strain from Wuhan, China.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

Table 2. Gene-/locus-wise frequency of the different types of transversion mutation encountered in the SARS-CoV-2 pan genome.

Genomic locus (nucleotide positions[1] between which the locus spans)	No. of U→A substituted positions detected	No. of U→G substituted positions detected	No. of C→A substituted positions detected	No. of C→G substituted positions detected	No. of A→U substituted positions detected	No. of A→C substituted positions detected	No. of G→U substituted positions detected	No. of G→C substituted positions detected	Total number of nucleotide positions involving transversion mutations (N)	N / length of the locus
5′ UTR (1-265)	02	01	04	00	01	02	06	00	16	0.06
nsp1 (266-806)	01	00	02	00	01	02	04	00	10	0.02
nsp2 (807-2720)	03	01	04	00	04	02	11	00	25	0.01
nsp3 (2720-8555)	04	04	13	01	03	10	48	03	86	0.02
nsp4 (8556-10055)	02	03	03	01	01	00	06	00	16	0.01
nsp5 (10056-10973)	00	00	01	00	01	01	03	02	08	0.008
nsp6 (10974-11843)	03	01	00	00	01	00	05	02	12	0.01
nsp7 (11844-12092)	00	00	00	00	00	00	01	02	03	0.01
nsp8 (12093-12686)	00	00	00	00	01	01	03	05	10	0.02
nsp9 (12687-13025)	01	00	01	00	00	00	03	00	05	0.01
nsp10 (13026-13442)	00	00	01	00	02	00	03	00	06	0.01
nsp11 (13443-13481)	00	00	00	00	00	00	01	00	01	0.02
nsp12 (13482-16328)	03	03	03	00	05	05	22	02	43	0.02
nsp13 (16329-18041)	03	01	05	00	01	01	30	00	41	0.02
nsp14 (18042-19622)	05	03	02	00	02	01	29	01	43	0.03
nsp15 (19623-20660)	02	00	02	00	00	01	14	00	19	0.02
nsp16 (20661-21554)	01	01	00	00	01	00	08	03	14	0.02
geneS (21563-25384)	06	06	07	05	11	07	55	11	108	0.03
orf3a (25393-26220)	00	07	05	00	03	02	31	04	52	0.06
geneE (26245-26472)	02	00	00	01	00	00	01	00	04	0.02
geneM (26523-27191)	01	01	01	01	00	00	10	01	15	0.02
orf6 (27202-27387)	01	00	00	00	00	00	07	00	08	0.04
orf7a (27394-27759)	00	01	00	00	01	00	05	01	08	0.02
orf8 (27894-28259)	00	01	01	00	00	01	10	03	16	0.04
geneN (28274-29533)	05	01	04	02	05	02	42	03	64	0.005
orf10 (29558-29674)	00	00	00	00	00	01	01	01	03	0.03
3′ UTR (29675-29903)	01	01	02	00	03	02	08	02	19	0.08
Entire pan genome	46	36	61	11	47	41	367	46	655	-

[1] Nucleotide positions between which the locus spans have been designated based on the 5' to 3' sequence of the 29,903 nucleotide RNA-genome (MN908947) of the reference strain from Wuhan, China.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

So far as the 655 positions of the SARS-CoV-2 pan genome undergoing transversion mutations are concerned, an overwhelming number (367) featured G→U conversion; this was 56.03% of the total number of transversion mutations encountered in the pan genome (Table 2). Furthermore, the frequency of transversion mutations within the SARS-CoV-2 pan genome locus harboring the non-structural protein-coding genes was 0.02 (342 ÷ 21,290 nucleotide positions), while the corresponding statistic for the locus harboring the structural protein-encoding genes was 0.03 (278 ÷ 8,112 nucleotide positions). On the other hand, the frequency of transversion mutations within the two UTRs was far higher, i.e, 0.07 (35 ÷ 493 nucleotide positions). Genes-wise, the loci encoding the S, Nsp3 and N proteins were most transversion mutation-affected (108, 86 and 64 respectively); these events again were replete with G→U substitutions (55, 48 and 42 respectively) (Table 2). On the other hand, the frequency of transversion mutation per nucleotide of a gene was highest for *orf3a* (0.06). Out of the total number of transversion events that have taken place in these four genes, 50.92% (i.e., 55 out of 108), 55.81% (48 out of 86), 65.62% (i.e. 42 out of 64) and 59.61% (31 out of 52) were G→U substitutions respectively (Table 2).

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 generation 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). In the context of the highly dynamic epidemiology of SARS-CoV-2, knowledge on its genome evolution becomes all important for the surveillance and containment of the outbreak. In fact, progressive diversification of the SARS-CoV-2 genome is taking place in sync with the pace at which it is undergoing transmission over geographies and anthropologies; and in doing so, it is playing out a ‘hide and seek’ game with the promises of antiviral drugs and vaccines innovated over time. Furthermore, all active genomic variants maintained within global/local RNA virus populations (i.e., quasispecies) are expected to possess equal abilities to replicate and complete the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

infection cycle (Eigen et al., 1988). Therefore, the more or less efficient circulation (and diversification into potential quasispecies) of the two original major-lineages of SARS-CoV-2 (clades S and L) across distinct geographies (Figure 1) reflects their equivalent pathological as well as evolutionary fitness.

In the present study, mutation frequency in the genetic material of SARS-CoV-2 was found to be 2.27×10^{-5} nucleotide positions mutated per nucleotide analyzed. This indicated that a majority of the individual genomes differed from the consensus sequence of the pan genome at one or more nucleotide positions. These small, but definite and widespread, genomic divergences further imply 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, 1994).

Significance of the abundance of mutations in genes encoding Nsp3, ORF3a, S and N

Viruses that have evolved to survive via changing their hosts are extremely skilled molecular manipulators; the key to their ecological fitness is attributed to their ability to subvert host defense systems to ensure survival, replication and proliferation (Bowie and Unterholzner, 2008). Coronavirus-encoded accessory proteins, in general, play critical roles in virus-host interactions and modulation of host-immune responses, thereby contributing to their pathogenicity (Narayanan et al., 2008). The clinical prognosis of SARS-CoV-2 infection (Prete et al., 2020), in conjunction with the gene content of its precisely-mapped RNA genome (Liu et al., 2019; Wu et al., 2020), indicates that this novel coronavirus also possesses sophisticated molecular mechanisms designed to subvert human immune system, thereby facilitating high transmission. The multi-domain accessory protein Nsp3, which is the largest among all SARS-CoV proteins, binds to viral RNA, nucleocapsid protein (N), and other viral proteins; in addition, it participates in polyprotein processing (Qiu and Xu, 2020). Furthermore, Nsp3 defies host innate immunity by its de-ATP-ribosylating, de-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

ubiquitinating, and de-ISGylating activities (Qiu and Xu, 2020). These attributes have currently made Nsp3, especially its papain-like protease component, a lucrative target for new antiviral drugs (Báez-Santos et al., 2015). In this scenario our discovery of 435 point mutations in the *nsp3* locus of the SARS-CoV-2 pan genome (349 transitions with 189 C→U substitutions, and 86 transversions with 48 G→U substitutions) calls for a re-evaluation of the molecular worthiness of Nsp3 as a faithful drug target (Tables S1 and S2 document the specific nucleotide positions where all the transitions and transversions have occurred in the pan genomic *nsp3* respectively).

There is a clearcut distinction in the cell-death related consequences of viral infection. While Herpesviruses, Poxviruses, Adenoviruses, and Baculoviruses bring about reduction of cell death, SARS-CoV (Coronaviruses), Ebola (Filoviruses), Poliovirus (Picornaviruses), West Nile virus (Flaviviridae) and Hepatitis B virus (Hepadnaviruses) are capable of increasing cell death (Clarke and Tyler, 2009). Earlier studies had reported that the accessory protein ORF3a of SARS-CoVs has pro-apoptotic activity (Padhan et al., 2008); very recent studies further implicated this protein of SARS-CoV-2 in inducing extrinsic apoptotic pathway through a unique membrane-anchoring strategy (Ren et al., 2020). In view of these key roles of ORF3a in SARS-CoV-2 pathogenicity, and thereby transmissibility, the existence of 128 point mutations (76 transitions with 43 C→U substitutions and 52 transversions with 31 G→U substitutions) in the *orf3a* locus of the pan genome (Tables S1 and S2) appears to be a part the insidious strategies of the virus towards successful completion of its life cycle and killing of host cells. The intrinsic molecular plasticity of *orf3a* activity is underscored by the fact that the 128 pan genomic mutations in this locus did not hamper the pathogenic aptitude of the virus.

The envelope spike protein S, and the unexposed nucleocapsid protein N, are among the most promising targets for vaccine development against SARS-Cov-2 (Chen et al., 2020; Pang et al., 2020; Salvatori et al., 2020). However, the detection of 300 point mutations (192 transitions with 106 C→U substitutions and 108 transversions with 55 G→U substitutions), distributed almost evenly across the total length of the *S* locus in the SARS-Cov-2 pan genome (Tables S1 and S2) seriously questions the prospects of eventual effectiveness of S-targeting vaccines. On the other hand, the *N* gene which was initially thought to be relatively more conserved and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

mutation-proof (Dutta et al., 2020), was found in the present study to harbor a total of 171 point mutations (107 transitions with 59 C→U substitutions and 64 transversions with 42 G→U substitutions) across global SARS-Cov-2 genomes (Tables S1 and S2). Effects of the above mentioned mutations on the structures and functions of N as well as S proteins need to be studied in-depth so as to ensure that the protein product of the right alleles are chosen as antigenic epitopes in the respective vaccination strategies.

Physicochemical underpinnings of the preponderance of C→U and G→U substitutions

In view of the overwhelming preponderance of C→U and G→U transitions in the global mutation spectrum of SARS-CoV-2 (as compared to all other transition and transversion mutations respectively) 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 S1; 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 forefront 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

(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%).

Errors resulting from replication as well as translation may be instrumental in rendering the G→U mutations prevalent across global SARS-CoV-2 genomes. RNA viruses mutate vastly as a result of their RNA-dependent RNA polymerases (RdRPs) being error prone. From the host's view point, a propensity for incorrect protein synthesis is ushered when cells are stressed due to viral infection, and under such circumstances the viral RNA itself becomes prone to mistranslation (Mohler and Ibba, 2017). It is therefore conceivable that SARS-CoV-2, in addition to classical mutations acquired from error-prone replication at the genomic level, uses the mistranslated replication-cum-transcription (RTC) complex for the development of diverged genomic lineages (Ribas de Pouplana et al., 2014; Ou et al., 2019). In other words, when the viral infection discharges its positively-sensed RNA-genome into the host cell, errors in the RdRP crops up via mistranslation (Ma et al. 2018; Nibert 2017); the consequent blend of wild-type and changed RdRP enzymes through its replication activities give rise to a range of viral genome-variants or quasispecies, even within a single transmission event (Ou et al., 2019). Those variants which have the best viral fitness, eventually, endure and become predominant in the population. In this context, it is further noteworthy that both tautomeric and anionic Watson-Crick(W-C)-like mismatches can increase the recruitment of replication and translation errors (Koag et al., 2014; Rozov et al., 2016). A sequence-dependent kinetic network system connects G•T/U wobbles with three particular W-C mismatches comprising of two quickly exchanging tautomeric species (Genol•T/U⇌G•Tenol/Uenol, population <0.4%) and one anionic species ($G•T^-/U^-$, population ≈0.001% at unbiased pH) (Kimsey et al. 2018).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 July 2020 doi:10.20944/preprints202005.0463.v2

Conclusion

The current investigation of 3,608 complete whole genome sequences of SARS-CoV-2 isolates from across the world brought to the fore a number of remarkable aspects of microevolution of this novel coronavirus. Phylogenomic analysis illustrated that the two major-lineages of the virus has thus far contributed almost equivalently to the pandemic, even as members of the early lineages are still mostly spread over Asian countries and those of the relatively recent lineages have undergone more global distribution. In the coming days it would be worth exploring whether this viro-geography has got any bearing on the differential death rates of COVID-19 in Asian and European/American countries (https://www.worldometers.info/coronavirus/). An overwhelming preponderance of transition mutations, and far less frequency of transversions, was observed across the pan genome of the virus, irrespective of whether the genetic locus encoded a non-structural or structural protein. In this context it is noteworthy that the 29,903-nucleotide-long SARS-CoV-2 pan genome was found to have maintained a substantive 655 transversion mutations, notwithstanding the fact that natural selection disfavors transversion mutations because they are often non-synonymous, so less likely to conserve the structural biological properties of the original amino acids. Furthermore, a molecular bias of mutations was observed in the SARS-CoV-2 pan genome involving exceedingly frequent C→U and G→U substitutions among all transitions and transversion events respectively. More comprehensive and multi-faceted surveillance of the microevolution of SARS-CoV-2 is needed so as to gain constant insights into the pathogenic dynamism of the virus, and improvise control and therapeutic strategies accordingly.

Supplementary Material

One MS Word file named “Supplementary_Information” accompany this paper.

Author Contributions

RC conceived the study, designed the experiments, interpreted the results, and wrote the paper, with the help of WG. CR performed the bioinformatic experiments and

participated in paper writing. SM, SKM and SM participated in data analysis. All authors read 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

Báez-Santos Y.M., John S.E., Mesecar, A.D., 2015. The SARS-coronavirus papain-like protease: Structure, function and inhibition by designed antiviral compounds. Antiviral. Res. 115: 21–38.

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): 30186-30189.

Bowie A.G., Unterholzner L., 2008. Viral evasion and subversion of pattern-recognition receptor signalling. Nat. Rev. Immunol. 8:911–922.

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: 4102-4107.

Brash,D.E., 2015. UV signature mutations. Photochem. Photobiol. 91: 15-26.

Chen, W.H., Strych, U., Hotez, P.J., Bottazzi, M. E., 2020. The SARSCoV-2 vaccine pipeline: an overview. Curr. Trop. Med. Rep. https://doi.org/ 10.1007/s40475-020-00201-6.

Clarke, P., Tyler, K.L., 2009. Apoptosis in animal models of virus-induced disease. Nat. Rev. Microbiol. 7: 144-155.

Danecek, P., Auton, A., Abecasis, G., et al. 2011. The variant call format and VCFtools. Bioinformatics. *27*: 2156-2158.

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.

Dutta, N.K., Mazumdar, K., Gordy, J.T., 2020. The nucleocapsid protein of SARS-CoV-2: a target for vaccine development. J. Virol. 94 e00647-20.

Eigen, M., McCaskill, J., Schuster, P., 1988. Molecular quasispecies. J. Phys. Chem. 92: 6881–6891.

Folegatti, P.M., Ewer, K.J., Aley, P.K., et al. 2020. Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. The Lancet. https://doi.org/10.1016/S0140-6736(20)31604-4.

Gardner, S.N., Slezak, T., Hall, B.G., 2015. kSNP3. 0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics. 31: 2877-2878.

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.

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.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T. and Aluru, S., 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Comm. 9: 1-8.

Katoh, K., Misawa, K., Kuma, K.I., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl. Acids Res. 30: 3059-3066.

Kimsey, I.J., Szymanski, E.S., Zahurancik, W.J., et al. 2018. Dynamic Basis for dG•dT misincorporation via tautomerization and ionization. Nature. 554: 195–201.

Koag, M.C., Nam, K., Lee, S., 2014. The spontaneous replication error and the mismatch discrimination mechanisms of human DNA polymerase β. Nucleic Acids Res. 42:11233–11245.

Korber, B., Fischer, W.M., Gnanakaran, S., et al. 2020. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. 2020.04.29.069054; doi: https://doi.org/10.1101/2020.04.29.069054.

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

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.

Ma, N.J., Hemez, C.F., Barber, KW., et al. 2018. Organisms with alternative genetic codes resolve unassigned codons via mistranslation and ribosomal rescue. eLife. 7:1–23.

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: 794-801.

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

Mohler, K., Ibba, M., 2017. Translational fidelity and mistranslation in the cellular response to stress. Nat. Microbiol. 2:17117.

Narayanan, K., Huang, C., Makino, S., 2008. SARS coronavirus accessory proteins.Virus Res. 133: 113–121.

Nibert, M.L., 2019. Mitovirus UGA(Trp) codon usage parallels that of host mitochondria. Virology. 507:96–100.

Ou, X., Cao, J., Cheng, A., et al. 2019. Errors in translational decoding: tRNA wobbling or misincorporation? PLoS Genet 15(3): e1008017. pgen.1008017.

Padhan, K., Minakshi, R., Mohammad, M.A.B., Jameel, S., 2008. Severe acute respiratory syndrome coronavirus 3a protein activates the mitochondrial death pathway through p38 MAP kinase activation. J. Gen. Virol. 89: 1960–196.

Page, A.J., Taylor, B., Delaney, A.J., 2016. SNP-sites: rapid efficient of SNPs from multi-FASTA alignments. Microb. Genom. 2 : e000056.

Pang, J., Wang, M.X., Ang, I.Y.H., et al. 2020. Potential rapid diagnostics, vaccine and therapeutics for 2019 novel coronavirus (2019-nCoV): a systematic review. J. Clin. Med. 9: E623.

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.

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. Article ID:201868.

Prete, M., Favoino, E., Catacchio, G., et al., 2020. SARS-CoV-2 Inflammatory Syndrome. Clinical Features and Rationale for Immunological Treatment. Int J Mol Sci. 21: 3377.

Qiu, Y., Xu, K., 2020. Functional studies of the coronavirus nonstructural proteins. STEMedicine 1(2): e39

Rambaut, A., Holmes, E.C., O'Toole, Á., et al. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature Microbiol. https://doi.org/10.1038/s41564-020-0770-5.

Ren, Y., Shu, T., Wu, D. et al. 2020. The ORF3a protein of SARS-CoV-2 induces apoptosis in cells. Cell Mol. Immunol. https://doi.org/10.1038/s41423-020-0485-9.

Ribas de Pouplana L, Santos MA, Zhu JH, Farabaugh PJ, Javid B. 2014. Protein mistranslation: friend or foe? Trends Biochem. Sci. 39:355–62.

Rozov A, Demeshkina N, Westhof E, Yusupov M, Yusupova G. 2016. New Structural Insights into Translational Miscoding. Trends Biochem. Sci. 41:798–814.

Saha, P., Majumder, R., Chakraborty, S., et al. 2020. Mutations in Spike Protein of SARS-CoV-2 Modulate Receptor Binding, Membrane Fusion and Immunogenicity: An Insight into Viral Tropism and Pathogenesis of COVID-19. ChemRxiv. https://doi.org/10.26434/chemrxiv.12320567.v1.

Tang, X., Wu, C., Li, X., Song, Y., et al. 2020. On the origin and continuing evolution of SARS-CoV-2. Natl. Sci. Rev. 7: 1012–1023, https://doi.org/10.1093/nsr/nwaa036.

Tseng, C.C., Li, C-S., 2007. Inactivation of viruses on surfaces by ultraviolet germicidal irradiation. J. Occup. Environ. Hyg. 4: 400-405.

Wang, Q., Zhang, Y., Wu, L., et al., 2020. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell. 181: 894-904.e9.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., et al., 2020. A new coronavirus associated with human respiratory disease in China. Nature. 579: 265-269.

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 Jan 23. https://doi.org/10.1101/2020.01.22.914952.