Trends of mutation accumulation across global SARS-CoV-2 genomes: implications for the 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:** rcnbusiliguri@gmail.com and wriman@jcbose.ac.in

Running Title: Microevolution of SARS-CoV-2

Abstract

To understand SARS-CoV-2 microevolution, this study explored the genome-wide frequency, gene-wise distribution, and molecular nature of all point-mutations detected across its 71,703 RNA-genomes deposited in the GISAID repository, till 21 August 2020. Globally, *nsp1*/*nsp2*/*nsp3*/ *nsp11* and *orf7a*/*orf3a*/*S* were the most mutation-ridden non-structural and structural genes respectively. Phylogeny based on 4,618 spatiotemporally-representative genomes revealed that entities belonging to the early lineages are mostly spread over Asian countries (including India, the biggest hotspot of the pandemic) whereas the recently-derived lineages are more globally distributed. Of the total 16,602 polymorphism-bearing sites in the pan-genome, 11,037 and 4,965 involved transitions and transversions, which in turn were predominated by cytidine-to-uridine and guanosine-to-uridine conversions, respectively. Positive selection of nonsynonymous mutations (dN/dS >1) in most of the structural, but not non-structural, genes indicated that SARS-CoV-2 has already harmonized its replication/transcription machineries with the host's metabolic system, while it is still redefining virulence/transmissibility strategies at the molecular level.

Key words: SARS-CoV-2, genome-wide mutations, transition, transversion, nonsynonymous and synonymous mutations, microevolution.

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 [1] ; that has spread to hundreds of countries, infecting tens of million people people, and killing approximately a million (https://covid19.who.int). The first whole-genome sequence of SARS-CoV-2 was deposited in GenBank (NC_045512.2) on January 5 by researchers of Shanghai Public Health Clinical Center and School of Public Health, Fudan University, Shanghai, China [2]. 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 [2]. High gene-arrangement similitudes of SARS-CoV-2 with coronaviruses found in bats (*Rhinolophus sinicus*) [3,4] and Sunda Pangolin (*Manis javanica*) [5] indicate SARS-CoV-2 to be a zoonotic disease [6]. 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 [7], the virus on its part is busy accumulating mutations across its pan-genome, some of which may well help it evade clinical interventions [8-11]. In this context of SARS-CoV-2 evolution, the present study analyzes 71,703 whole-genome sequences of this novel coronavirus, isolated from 108 different countries, to reconstruct the phylogeny and reveal the global trends of point-mutation accumulation. Besides identifying the genome-wide frequency, gene-wise distribution, and molecular characteristics of all point-mutations detected across global SARS-CoV-2 genomes, the ratio between the recruitment rates of nonsynonymous (dN) and synonymous (dS) mutations (dN/dS) was determined to understand the selection pressures on the different genes. Potential molecular biological and chemical mechanisms that could be instrumental in accelerating mutation recruitment were also envisaged.

2. Methods and algorithms

2.1. Comparative genomics

Of the 83,475 SARS-CoV-2 whole-genome sequences available in the repository of Global Initiative on Sharing All Influenza Data (GISAID) till 21 August 2020, an overwhelming percentage (42.22%) were from UK, followed by those from USA (22.39%), Australia (3.46%), Spain (3.25%), India (>2.88%), Portugal (1.98%), The Netherlands (1.92%), South Africa (1.35%), Canada (1.32%), Switzerland (1.21%), Belgium (1.17%) and China (1.07%%). Genome sequences of this novel coronavirus were also found to have been deposited in the GISAID collection from 96 other countries. All the 83,475 genome sequences were downloaded from the GISAID website on 21 August 2020 together with the metadata associated with the depositions. The dataset was filtered using the Augur tool kit [12] to eliminate undesired sequences - 11,723 entries were removed based on the minimum 29,000 nucleotide length cut-off, another 49 were removed because they originated from non-human sources. In this way, 71,703 GISAID entries remained in the final dataset used for further study. For all the present analyses of comparative genomics, the 29,903-nucleotide-long complete whole-genome of the earliest-sequenced SARS-CoV-2 strain from Wuhan, China (accession number NC_045512.2) was used as the reference sequence. The software package called MicroGMT or Microbial Genomics Mutation Tracker [13] was used to identify modifications in the SARS-CoV-2 genome sequences analyzed. This package essentially uses Minimap2 [14] and Bcftools [15] to map individual genomes against the reference and store the results in a Variant Call Format (VCF) table. It further utilizes the SnpEff tool [16] to characterize all the detected mutations at the level of the nucleotide as well as the amino acid in the translated sequence. Although MicroGMT also reports instances of insertion and deletion in the sequences compared, the current study focused only on the point-mutation data, which in turn were further verified as follows. The software MAFFT [17] 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* [18], which specifically identifies single nucleotide polymorphisms (SNPs) from aligned multi-fasta sequence files. Subsequently, the VCF file generated from the *SNP-site* analysis was processed using the software *VCFtools* [19] to enumerate all transition and transversion events within the entire dataset of aligned whole-genome sequences.. Frequency of point mutations (M_f) in the SARS-CoV-2 pan-genome, or a given segment (locus) of the pan-genome was calculated as $P_i \div (L_n \times N_s)$. In this equation giving a measure of polymorphisms per nucleotide of the genome/locus aligned per sequence entity present in the dataset, P_i is the number of instances of polymorphism detected within the genome or locus under consideration, L_n is the nucleotide length of the genome or locus considered, and N_s is the number of sequenced entities present in the dataset. dN/dS (also known as ω or Ka/Ks) value, which is the ratio between the recruitment rates of nonsynonymous (dN) and synonymous (dS) mutations, was determined for all the individual genes of SARS-CoV-2, based on likelihood analysis using the software package HyPhy [20]. Sequence similarity between SARS-CoV-2 genomes was computed in a pairwise manner involving all the combinations possible, using using the software FastANI, which uses a high throughput method for average nucleotide identity analysis [21].

2.2. Phylogenomic Analyses

Evolutionary relationship between the existing SARS-CoV-2 lineages was inferred from a phylogenetic tree constructed based on a subset of the 71,703 whole-genome sequences used for studying mutation accumulation trends. Sub-sampling was necessary because it is not possible to meaningfully display 71,703 sequences in a single phylogenetic tree. This sub-dataset, comprising 4,618 complete whole-genome sequences, was created using the software package Augur [12], and by means of including (in an unbiased way) 150 genomes per geographical region (continent) per month since the first Wuhan strain was sequenced (NC_045512). Multiple sequence alignment was also created using the Augur tool kit of the Nextstrain package. Further alignment was carried out using the software IQ-TREE 2 [22], and the Generalised Time Reversible (GTR) model was followed to construct the phylogenetic tree, which was finally visualized in the software Auspice (https://auspice.us). For the labeling of clades in the phylogenetic tree, type defining marker mutations were downloaded from the Nextstrain github repository which comes as a package within the Nextstrain tool (https://github.com/nextstrain/ncov). Rules of clade-labeling followed were those mentioned in the website located at https://nextstrain.github.io/ncov/naming_clades.html. Thus, clades were labeled based on the geographical origin of the sequences, plus three different concepts of clade nomenclature that are in use for the ongoing COVID-19 outbreak, namely (i) the dynamic clade nomenclature system PANGOLIN [23] (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. [24], and followed by GISAID, which names major clades based on nine distinct marker mutations spread over 95% of the known SARS-Cov-2 diversity.

In order to elucidate the biogeography and microevolution of SARS-CoV-2 in India, the latest super hotspot of the COVID-19 pandemic, we reconstructed the phylogeny using a separate sub-dataset (derived from the same 71,703 GISAID sequences) that included a large number of sequences from Indian strains, alongside representative sequences from all other geographical areas to enable understanding of the whole dynamics from a global perspective. This sub-dataset building involved 'focal' sampling for India and 'selective' sampling for other geographical areas, both following custom rules laid down in Nextstrain: for the 'focal' country (India), up to 300 sequences, or whatever maximum number (<300) is available, per month for each year under consideration; for contextual sampling, 50 such whole-genome sequences per month per country that are genetically associated to the 'focal' samples based on the priority call criterion called 'Proximity'. This approach short-listed 5,778 whole-genome sequences, of which 1,148 belonged to the 'focal' country India. These 5,778 sequences were analyzed using the same methodology as the one described above for the global phylogentic tree, following which the Indian sequences were mapped as per their clade affiliation and indicated using the GISAID and Year-Letter clade nomenclature systems.

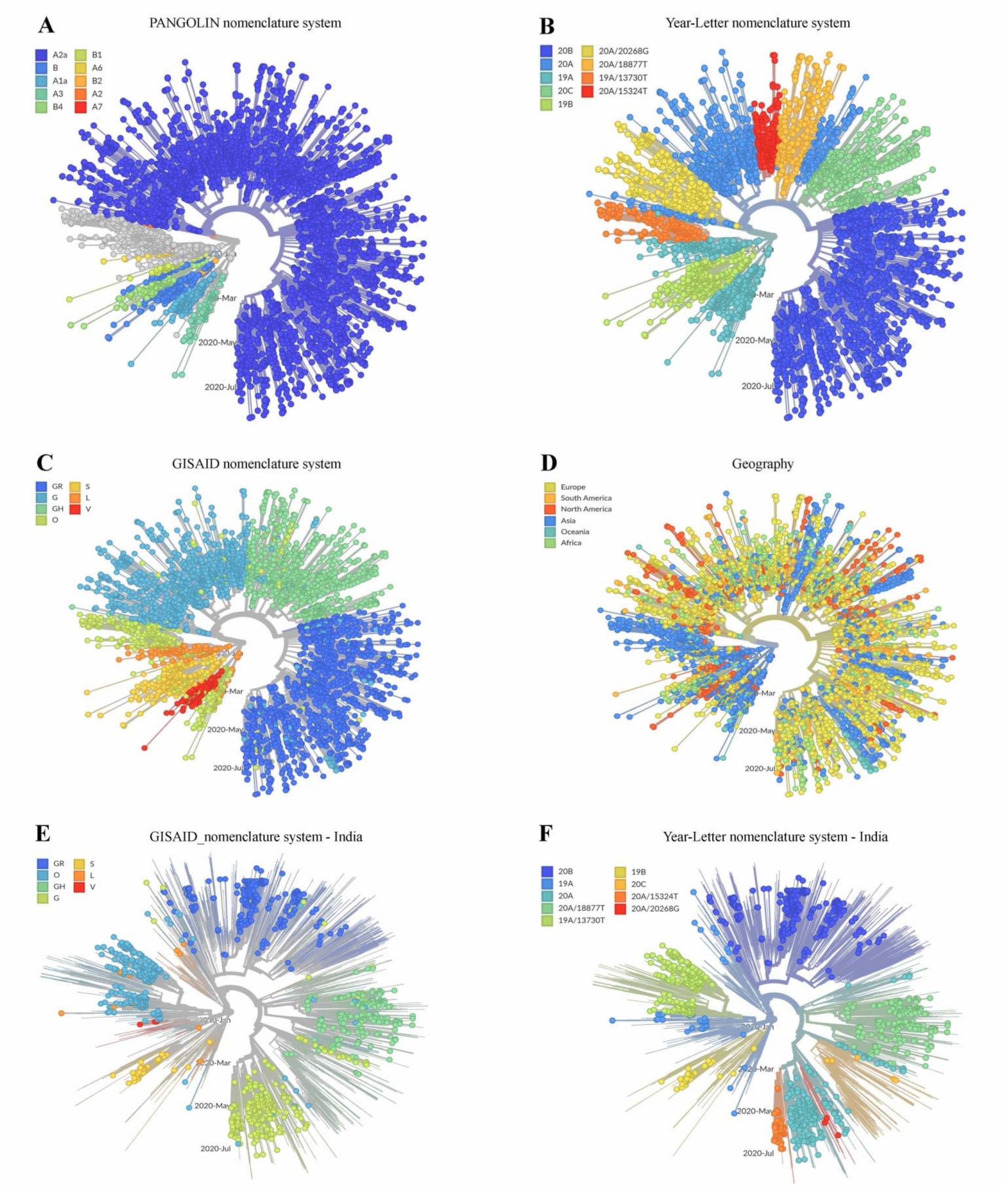
3. Results and Discussions

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

A cursory estimation of average nucleotide identity (ANI, for a Kmer size of 16, over a fragment size of 1,000 nucleotides), and sequence length coverage for all the pairwise alignments possible between the 11,189 complete whole-genome sequences available simultaneously in GISAID as well as NCBI SARS-CoV-2 database (https://www.ncbi.nlm.nih.gov/sars-cov-2/) on 21 August 2020, showed that in all the cases both identity and coverage were within 99 and 100% (notably, ANI calculation was not possible for all the 71,703 GISAID genomes retrieved on 21 August 2020, so this sample-survey was carried out). Whilst individual SARS-CoV-2 genomes differed only in terms of a few nucleotides, the small but rampant sequence divergences across geographies indicated that within the short time span of the current pandemic, the pan-genome has diversified, and the quasispecies reservoir has expanded, rapidly for this novel coronavirus. This holds major implications for the adaptation of the virus within human hosts, and in doing so have serious consequences on the resultant pathogenesis, disease complications, and control [25].

The overall evolutionary paths traced thus far by SARS-CoV-2 was delineated by labeling the 4,618 global (GISAID) sequences on the phylogenetic tree using three different concepts of clade nomenclature defined in the web-based resoure https://nextstrain.github.io/ncov/ (Figures 1A-1C). Information regarding the 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 [23] called **P**hylogenetic **A**ssignment of **N**amed **G**lobal **O**utbreak **LIN**eages (PANGOLIN), reflected the global preponderance of the ancestral SARS-CoV-2 lineage identified as Clade A. Notably, this ancestral clade [23] is epitomized by the 29,872-nucleotide-long genome LR757995, which was isolated from Wuhan on 26 December 2019, sequenced, and submitted to GenBank on 30 January 2020. The PANGOLIN is nomenclatural approach also illustrated the clear divergence of Clade A from the other SARS-CoV-2 major-clade named B, the typical representative (NC_045512.2) of which was also isolated from Wuhan on 26 December 2019, but submitted to GenBank on 12 January 2020. Albeit the genome sequence NC_045512.2 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.

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.



Consistent with the above phylogenetic interpretations, labeling of the tree with the third clade-nomenclature convention proposed by Tang et al. [24] 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 21 August 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). India being the latest super hotspot of the COVID-19 pandemic, recording >50,000 cases of infection and >750 cases of fatality daily since the last week of July 2020 (https://www.worldometers.info/coronavirus/country/india/), the phylogeny and biogeography of Indian SARS-CoV-2 isolates was analyzed using the specialized (GISAID-derived) dataset encompassing 1,148 and 4,630 genome srquences of Indian and global origins respectively. The phylogenetic tree topology obtained with this India-focused dataset (Figures 1E and 1F) was essentially congruent with that obtained for the global dataset of 4,618 GISAID sequences (Figures 1A-1D). Mapping of the Indian sequences on this tree topology using the GISAID (Figure 1E) and Year-Letter (Figure 1F) clade nomenclature systems showed that all the mutational types which epitomize the major clades of global SARS-CoV-2 evolution are also present in India, albeit at potentially different frequencies of distribution within the country's viral population. For instance, the relatively lower number of sequences populating the two emerging lineages 20A/20268G and 20A/15324T can be clearly seen in Figure 1F which, in turn, corroborated the hypothesis that in the Asian countries the ancestral lineages are still more prevalent than the recently-derived mutational groups.

3.2. Gene-wise mapping of the substitution mutations recruited in global SARS-CoV-2 genomes

Multiple alignment of the 71,703 SARS-CoV-2 whole-genome sequences investigated in this study (29,903 completely aligned nucleotide positions, with reference to the 5' to 3' sequence of NC_045512.2, the earliest-sequenced strain from Wuhan, China), revealed 20,163 instances of single nucleotide substitution (polymorphism) across the genomes participating in the alignment (Supplementary File 1, Table S1). Overall, these point mutations have taken place at a frequency (M_f) of 9.4 × 10^{-6} , i.e. [20,163 ÷ (29,903 × 71,703)] polymorphisms per nucleotide of the SARS-CoV-2 genome aligned per sequence entity present in the dataset,. On the other hand, frequency of point mutations (M_f) in the 21,290 nucleotide long SARS-CoV-2 genomic locus coding for non-structural proteins was found to be 8.78 × 10^{-6}, i.e. [13,417 ÷ (21,290 × 71,703)], as across the global dataset of 71,703 genomes, 13,417 instances of polymorphism were detected within this locus. M_f for the 8,112 nucleotide long genomic locus encoding structural proteins was considerably higher, i.e. 1.06 × 10^{-5} = [6,196 ÷ (8,112 × 71,703)]. Notably, frequency of point mutations in the 493 nucleotide long total-UTR of the SARS-CoV-2 genome was highest, i.e. 1.54 × 10^{-5} = [547 ÷ (493 ×71,703)]. Genes-wise, the loci for *nsp1* and *orf7a*, happened to be the most mutation-prone non-structural and structural gene respectively, as their M_f values were 8.74 × 10^{-6} [339 ÷ (541 ×71,703)] and 9.83 × 10^{-6} [258 ÷ (366 × 71,703)] respectively; M_f was also comparably high for *nsp2* (8.64 × 10^{-6}) and *orf3a* (9.69 × 10^{-6}).

The 20,163 instances of single nucleotide substitution (polymorphism) detected across 71,703 SARS-CoV-2 genomes corresponded to only 16,002 nucleotide positions of the global alignment. This has happened in such a way that 12,203 positions each involved one specific substitution in one particular strain; 3,437 positions each involved two different substitutions in two different strains; and 362 positions each involved three different substitutions in three different strains. This distribution showed that 53.5% (i.e. 16,002 ÷ 29,903) of the SARS-CoV-2 pan-genome has developed polymorphism via generation of small but definite mutations across the plethora of strains disseminated globally since the COVID outbreak in December 2019. Table 1 shows the genetic locus-wise distribution of the 16,002 polymorphism-bearing nucleotide positions of the SARS-CoV-2 pan-genome. This mapping revealed that all the 25 genes of SARS-CoV-2, its two untranscribed regions (UTRs), and also the intergenic regions, have recruited mutations in one or more sequenced genome(s). Out of these 16,002 polymorphism-bearing nucleotide positions, 11,046 were found to be located between nucleotide positions 266 and 21,555 (with reference to the 1 - 29,903 positions of NC_045512.2), within the foremost locus of the SARS-CoV-2 genome that encodes the 16 non-structural proteins, Nsp1 through Nsp16. All the Nsp-encoding SARS-CoV-2 genes, except *nsp11*, were found to have more than 130 point mutation-bearing positions each (*nsp11* has only 21 such positions globally); numerically, maximum number of polymorphic positions were in the gene encoding Nsp3 (3,135). On the other hand, 4,457 polymorphic positions were found to occur 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 NC_045512.2); maximum number of point mutation-bearing positions (1,966) were detected in the gene encoding spike protein S. Furthermore, 350 and 149 polymorphic positions were also identified within the two UTRs (located in the 5' and 3' ends of the SARS-CoV-2 genome) and the intergenic regions (between different structural genes), respectively.

Table 1. Locus-wise distribution of polymorphism-bearing nucleotide positions of the SARS-CoV-2 pan-genome, based on 71,703 complete whole-genomes sequenced globally till 21 August 2020.

Loci (length in bp)	**Number of transitions detected (Ts)**				**Total Ts detected**	**Number of transversions detected (Tv)**								**Total Tv detected**	**Ts+Tv**	**Rate of mutation accumulation across genomes (Mu_r)**	**No. of non-synonymous mutations detected**	**No. of synonymous mutations detected**	**dN/dS**
	A>G	**G>A**	**C>U**	**U>C**		**A>U**	**U>A**	**C>A**	**A>C**	**C>G**	**G>C**	**G>U**	**U>G**						
5' UTR (265)	21	19	42	29	111	16	13	7	8	1	3	20	7	75	186	9.78×10^{-6}	ND	ND	ND
nsp1 (541)	55	60	79	60	254	12	12	4	4	3	4	36	10	85	339	8.74×10^{-6}	213	121	0.7398
nsp2 (1914)	239	181	286	188	894	34	50	19	48	2	3	97	39	292	1186	8.64×10^{-6}	780	399	0.9479
nsp3 (5836)	644	356	682	595	2277	111	122	65	145	10	26	289	90	858	3135	7.49×10^{-6}	1982	1127	0.5803
nsp4 (1500)	134	84	188	166	572	24	38	10	11	1	11	55	24	174	746	6.94×10^{-6}	429	313	0.5126
nsp5 (918)	81	43	110	91	325	12	20	10	21	1	5	43	16	128	453	6.88×10^{-6}	276	174	0.6417
nsp6 (870)	78	44	95	93	310	19	24	9	10	5	8	52	20	147	457	7.33×10^{-6}	274	175	0.7000
nsp7 (249)	27	11	34	21	93	2	7	2	6	0	1	13	8	39	132	7.39×10^{-6}	73	58	0.4999
nsp8 (594)	54	37	70	56	217	7	7	2	13	1	4	27	6	67	284	6.67×10^{-6}	166	111	0.4892
nsp9 (339)	33	24	45	24	126	2	6	4	3	0	0	12	8	35	161	6.62×10^{-6}	87	74	0.4933
nsp10 (417)	29	18	49	41	137	7	7	5	7	0	5	13	6	50	187	6.25×10^{-6}	103	82	0.4187
nsp11 (39)	2	3	5	3	13	1	2	0	1	1	0	3	0	8	21	7.51×10^{-6}	17	4	1.132
nsp12 (2847)	240	110	308	273	931	39	44	22	41	9	9	178	35	377	1308	6.41×10^{-6}	744	552	0.6057
nsp13 (1713)	170	58	191	164	583	24	29	19	29	2	2	114	22	241	824	6.71×10^{-6}	481	340	0.4500
nsp14 (1581)	134	57	186	157	534	15	29	15	36	7	6	110	24	242	776	6.85×10^{-6}	454	314	0.4024

nsp15 (1038)	134	61	107	102	404	23	24	13	27	1	8	74	16	186	590	7.93×10^{-6}	392	195	0.3937
nsp16 (894)	80	47	86	95	308	22	15	7	16	4	1	58	16	139	447	6.97×10^{-6}	273	167	0.4554
geneS (3822)	303	156	402	385	1246	133	77	53	79	18	53	239	68	720	1966	7.17×10^{-6}	1216	725	0.6193
orf3a (828)	76	35	127	100	338	29	25	18	28	4	13	94	26	237	575	9.69×10^{-6}	413	147	1.5013
geneE (228)	12	13	25	28	78	5	8	5	5	3	2	18	5	51	129	7.89×10^{-6}	79	47	1.0206
gene M (669)	43	17	80	53	193	18	13	9	7	3	5	47	13	115	308	6.42×10^{-6}	162	142	0.6548
orf6 (186)	18	7	17	32	74	14	5	4	2	0	0	16	5	46	120	9.0×10^{-6}	78	36	1.3944
orf7a (366)	35	17	56	47	155	18	16	8	13	4	2	34	8	103	258	9.83×10^{-6}	172	72	1.1946
orf8 (366)	31	15	42	54	142	19	9	9	6	2	7	39	10	101	243	9.26×10^{-6}	153	73	1.4522
geneN (1260)	125	72	203	94	494	67	16	26	34	7	17	124	14	305	799	8.84×10^{-6}	505	284	1.2633
orf10 (117)	8	3	16	13	40	5	2	0	3	0	1	6	2	19	59	7.03×10^{-6}	39	16	1.2981
3' UTR (229)	28	11	31	24	94	12	6	4	8	0	3	32	5	70	164	9.99×10^{-6}	ND	ND	ND
Intergenic	22	6	35	31	94	11	8	2	7	0	3	20	4	55	149	ND	ND	ND	ND
Grand Total	2856	1565	3597	3019	11037	701	634	351	618	89	202	1863	507	4965	16002	7.46×10^{-6}	9561	5748	-

ND = not determined

dN = Rate of non-synonymous mutation accumulation (ratio between the number of non-synonymous mutations and non-synonymous sites)

dS = Rate of synonymous mutation accumulation (ratio between the number of synonymous mutations and synonymous sites)

3.3. High rate of nonsynonymous mutations in the structural protein-coding genes

SARS-CoV-2, with its typically long, positive single-stranded RNA genome (that dedicates almost two-third of its length to encoding non-structural proteins), has experienced strong selection pressure over a short period of time. For animal viruses, in general, forces of selection (fitness constraints) emanate from host immunogenic responses, and also during replication and transmission between hosts. Evolutionarily fit (selected) strains develop tropism, and infect different cell types or tissues of the host, reproduce within them, and in turn give rise to a variety of new strains having diverse chronic to acute infectious characteristics [26,27]. The advent of affordable high-throughput nucleotide sequencing techniques has enabled the generation of large scale genomic data, which in turn can reveal where, when, and (sometimes) how viral pathogens have responded to various forces of natural selection. In the context of codon models, natural selection of any genomic locus is typically measured using the parameter dN/dS (also referred to as ω or Ka/Ks), which represents the ratio between the recruitment rates of nonsynonymous (dN) and synonymous (dS) mutations. As the present study seeked to understand the trends of sequence evolution across a global dataset of SARS-CoV-2 genomes, a likelihood-based analysis was carried out to determine the selection pressures on the different genes of this novel coronavirus. For any genetic locus, trends of Darwinian selection yield dN/dS values >1, whereas tendencies of negative selection, or selective removal of alleles that are deleterious, result in dN/dS values <1 [28]. In our analysis, all the SARS-CoV-2 genomic loci encoding non-structural proteins (Nsps), except *nsp11*, were found to have dN/dS values <1; among the structural genes, the same was true for *S* and *M* (genes for the structural proteins ORF3a, E, ORF6, ORF 7a, ORF8, N, and ORF10 had dN/dS >1; see Table 1). These numbers indicated that in the Nsp-coding genes of SARS-CoV-2 (except *nsp11*) nonsynonymous point mutations are under purifying selection; in contrast, for the structural protein-coding genes (except *S* and *M*), nonsynonymous point mutations tend to result in positive selection, thereby becoming potent drivers of evolution of this virus. Interestingly, most of the structutural protein-coding genes that are under positive selection (i.e. the ones having dN/dS >1) confer abilities to infect host cells via evading the immune system (specifically, the innate immune system), and eventually induce apoptotic pathways [29-35]. Consequently, brisk amino acid changes in these protein sequences may well be instrumental in allowing the virus innovate newer techniques to fulfil its pathogenic objectives. From a holistic evolutionary perspective based on the above considerations, SARS-CoV-2 seems to have already succeeded in stably synchronizing its replication and transcription machineries with the host’s metabolic environment (as its non-structural genes are clearly recruiting less nonsynonymous mutations). The virus, however, by means of actively recruiting more nonsynonymous mutations in its structural genes, is still testing newer biophysical options to increase the efficiency of its molecular contrivances for virulence and transmissibility (pathogenicity).

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

Of the 16,002 polymorphism-bearing nucleotide positions of the SARS-CoV-2 pan-genome, 11,037 and 4,965 involved transition and transversion mutations respectively. In this way, a transition:transversion ratio of 2.22 characterized the nucleotide substitution bias of SARS-CoV-2. In other words, the rate of transition mutations in SARS-CoV-2 is higher than what is expected if transition and transversion events took place randomly. Individually also, all the SARS-CoV-2 genes had transition:transversion ratios >1.

The ratio of transition-bearing sites and locus length, for the non-structural and structural protein-coding regions, was 0.37 (i.e. 7,978 ÷ 21,290) and 0.34 (i.e. 2760 ÷ 8,112) respectively. In contrast, the ratio of transition-bearing sites and locus length, for the total UTR of the SARS-CoV-2 pan-genome was higher, i.e. 0.41 (= 205 ÷ 493). *nsp1* and *orf7a* were found to have the highest (transition-site count) : (locus length) ratios, 0.47 (i.e. 254 ÷ 541) and 0.42 (i.e. 155 ÷ 366) among the non-structural and structural genes respectively. In terms of the number of nucleotide positions mutated, the loci coding for the Nsp3 and S proteins were the most transition mutation affected (2,277 and 1,246 transition affected sites in *nsp3* and *S* respectively). Of the 11,037 positions of the SARS-CoV-2 pan-genome involving transition mutations, an overwhelming number (count: 3,597) featured C→U conversion; this was 32.6% of the total number of transition mutation-bearing sites of the pan-genome (Table 1). Individually, again, all the SARS-CoV-2 genes had C→U conversion as the most predominant transition type across the global genomes analyzed.

The ratio of transversion-bearing sites and locus length, for the non-structural and structural protein-coding regions, was 0.14 (i.e. 3,068 ÷ 21,290) and 0.21 (i.e. 1,697 ÷ 8,112) respectively. In contrast, the ratio of transversion-bearing sites and locus length, for the total UTR of the SARS-CoV-2 pan-genome was higher, i.e. 0.31 (= 155 ÷ 493). *Nsp11* and *orf3a* were found to have the highest (transversion-site count) : (locus length) ratios, 0.21 (i.e. 8 ÷ 39) and 0.29 (i.e. 237 ÷ 828) among the non-structural and structural genes respectively. In terms of the number of nucleotide positions mutated, the loci coding for the Nsp3 and S proteins were the most transversion mutation affected (858 and 720 transversion affected sites in *nsp3* and *S* respectively). Of the 4,965 positions of the SARS-CoV-2 pan-genome involving transversion mutations, 37.5% (i.e. 1,863) featured G→U conversion (Table 1). Individually, again, all the SARS-CoV-2 genes had G→U conversion as the most predominant transversion type across the global genomes analyzed.

3.5. Evolutionary/pathogenic significance of copious mutations in non-structural genes 1, 2, 3 and 11, and most of the structural genes

Pace of mutation accumulation due to replication errors is generally higher in the RNA genomes of viruses than the spontaneous mutation rates in the DNA genomes of other living entities. Since RNA viruses encode their own genome replication machineries (and do not depend on the hosts' replication systems as the DNA viruses do), they can optimize their mutation rates to achieve evolutionary fitness. This leads to an unrelenting generation of genomic variants 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 [36]. 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 (quasispecies) are expected to possess equal abilities to replicate and complete the infection cycle [36]. In this context, the divergence of several lineages and sub-lineages of SARS-CoV-2 since the December-2019 outbreak (via generation of small mutations across its world-wide strains) - alongside the more or less efficient circulation of its two original major-lineages (clades indicated as S and L in Figure 1) across distinct geographies - reflects the equivalent pathological and evolutionary fitness of all its extant quasispecies. 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 [25].

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 [37]. Coronavirus-encoded accessory proteins, in general, play critical roles in virus-host interactions and modulation of host-immune responses, thereby contributing to their pathogenicity [38]. The clinical prognosis of SARS-CoV-2 infection [39], in conjunction with the gene content of its precisely-mapped RNA genome [2,5], indicates that this novel coronavirus also possesses sophisticated molecular mechanisms designed to subvert human immune system, thereby facilitating high transmission.

nsp1 and *nsp2* are the most mutation-prone non-structural genes of SARS-CoV-2, as they have the highest M_f values among all such genes. *nsp1* also has the highest (transition-site count) : (locus length) ratios among all the non-structural genes. Nsp1 is known to inhibit translation by binding to the host's 40S ribosome, and also inhibit IFN signaling, while Nsp2 inhibits the two host proteins proinhibitin1 and proinhibitin2 to disrupt the cellular environment [33]. Copious mutations in these two genes, therefore, can help the virus innovate novel molecular routes to evade host immunogenic response. 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 [40]. Furthermore, Nsp3 defies host innate immunity by its de-ATP-ribosylating, de-ubiquitinating, and de-ISGylating activities [40]. These attributes have currently made Nsp3, especially its papain-like protease component, a lucrative target for new antiviral drugs [41]. In this scenario our discovery of 3,135 polymorphic nucleotide positions in the *nsp3* locus of the SARS-CoV-2 pan-genome (2,277 transitions with 682 C→U substitutions, and 858 transversions with 289 G→U substitutions; see Table 1) calls for a re-evaluation of the molecular worthiness of Nsp3 as a faithful drug target (Supplementary File 1, Table S1 documents the specific nucleotide positions where all the transitions and transversions have occurred in *nsp3*). With regard to the 16 non-structural genes of SARS-CoV-2 it is remarkable that only *nsp11* has a dN/dS value >1. The exact function of Nsp11 is not known. However, in Arterivirus, this protein has been characterized as a Nidoviral uridylate-specific endoribonuclease (NendoU) that is associated with RNA processing [29]. So, a dN/dS vaue >1 for *nsp11* could be indicative of an intrinsic versatility of this gene in contriving newer ways of shielding the genetic material from the host's innate-immune system.

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 [42]. Earlier studies had reported that the accessory protein ORF3a of SARS-CoVs has pro-apoptotic activity [43]; very recent studies further implicated this protein of SARS-CoV-2 in inducing extrinsic apoptotic pathway through a unique membrane-anchoring strategy [34]. In view of these key roles of ORF3a in SARS-CoV-2 pathogenicity, and thereby transmissibility, the existence of 575 point mutations (338 transitions with 127 C→U substitutions and 237 transversions with 94 G→U substitutions) in the *orf3a* locus of the pan-genome (Table 1 and Supplementary File 1, Table S1) 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 575 global polymorphic positions in this locus did not hamper the pathogenic aptitude of the virus. Furthermore, in this context it is noteworthy that *orf3a* is one of the most mutation-prone structural genes (M_f second highest among all such genes); its dN/dS is value >1; the locus also has the highest (transversion-site count) : (locus length) ratio among all the SARS-CoV-2 structural genes.

Furthermore, in the context of the structural genes of SARS-CoV-2 it is noteworthy that *orf7a* is the most mutation-prone ($M_f = 9.83 \times 10^{-6}$), has the highest (transition-site count) : (locus length) ratio, and a dN/dS value of 1.2. In all SARS-CoVs, the type I membrane protein encoded by this gene (i.e. ORF7a) is known to interact with bone marrow stromal antigen-2 (BST-2) and may play a role in viral assembly or budding events unique to SARS-CoVs [33]. Budding events are central to the transmissibility of SARS-CoV-2, so recruitment of copious mutations, especially nonsynonymous ones, in this structural gene affords novel molecular options to increase the efficiency of virulence (pathogenicity) 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 [44-46]. However, the detection of 1,966 point mutations (1,246 transitions with 402 C→U substitutions and 720 transversions with 239 G→U substitutions), distributed almost evenly across the total length of the *S* locus in the SARS-Cov-2 pan-genome (Table 1, and Supplementary File 1, Table S1) seriously questions the prospects of eventual effectiveness of S-targeting vaccines. Effects of the above mentioned mutations on the structures and functions of S protein need to be studied in-depth so as to ensure that the protein product of the right alleles are chosen as antigenic epitopes for vaccine development.

3.6. 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 processes akin to hydrolytic deamination under the action of ultra-violet (UV) irradiation, which is well established in the context of DNA [47]. C→U conversion is also possible chemically under the mediation of bisulfite reagents [48] that are frequently used as disinfectants, antioxidants and preservative agents. Incidentally, several control techniques involving heating, sterilization, ultraviolet germicidal irradiation (UVGI) [49] and/or chemical disinfectants [50] 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 [51], 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 [52]. 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 (NC_045512.2) 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 [53]. 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 [54,55]. 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 [56,57]; 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 [55]. 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 [58,59]. 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) [60].

4. Conclusion

The current investigation of 71,703 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 4,965 transversion mutations, notwithstanding the fact that natural selection disfavors transversion mutations because they are often nonsynonymous, so less likely to conserve the structural biological properties of the original amino acids. Likewise, positive selection of nonsynonymous mutations (reflected in dN/dS values >1) in most of the structural genes of SARS-CoV-2 is indicative of vigorous molecular maneuvering by virus to augment its virulence potentials, escape human immunity, and ensure enhanced global transmissibility. 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.

Acknowledgement

The authors are grateful to all the researchers worldwide who sequenced and shared SARS-CoV-2 whole-genome sequence data via the NCBI (https://www.ncbi.nlm.nih.gov/sars-cov-2/) and GISAID (https://www.gisaid.org/) respositories. The research has benefited from the valuable data available at the websites https://nextstrain.org/ncov/ and https://nextstrain.org/blog/2020-06-02-SARSCoV2-clade-naming.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary Material

Supplementary data to this article includes one Supplementary File (MS Excel) that contains only one Supplementary Table.

Author Contributions

RC conceived and designed the study. RC and WG interpreted the results and wrote the paper. CR brought in the methodology, performed the experiments, and contributed to the manuscript. 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.

References

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

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

[3] YX-L. P. Zhou, X-G. Wang, B. Hu, et al., 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.

[4] D. Paraskevis, E. G. Kostaki, G. Magiorkinis, et al., 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 (2020) 104212.

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

[6] R. Lu, X. Zhao, J. Li, et al., Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, Lancet 395 (2020) 565-74.

[7] P. M. Folegatti, K. J. Ewer, P. K. Aley, et al., 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.

[8] D. Benvenuto, S. Angeletti, M. Giovanetti, et al., Evolutionary analysis of SARS-CoV-2: How mutation of non-structural protein 6 (NSP6) could affect viral autophagy, J. Infect. S0163-4453 (20) (2020) 30186-30189.

[9] B. Korber, W. M. Fischer, S. Gnanakaran, et al., 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.

[10] P. Saha, R. Majumder, S. Chakraborty, et al., 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.

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

[12] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T, Bedford, R. A. Neher, Nextstrain: real-time tracking of pathogen evolution, Bioinformatics. 34 (2018) 4121-4123.

[13] Y. Xing, X. Li, X. Gao, Q. Dong, MicroGMT: A mutation tracker for SARS-CoV-2 and other microbial genome sequences, Frontier. Microbiol. 11(2020) 1502.

[14] H. Li, Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics. 34(2018) 3094-3100.

[15] H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinformatics. 27 (2011) 2987-2993.

[16] P. Cingolani, P. Adrian, W. Le Lily, C. Melissa, N. Tung, W. Luan, et al., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3, Fly 6, no. 2 (2012) 80-92.

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

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

[19] P. Danecek, A. Auton, G. Abecasis, et al., The variant call format and VCFtools, Bioinformatics 27 (2011) 2156-2158.

[20] S. L. K. Pond, S. D. W. Frost, S. V. Muse, HyPhy: hypothesis testing using phylogenies, Bioinformatics 21(2005) 676-679.

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

[22] B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, et al., IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era, Mol. Biol. Evol. 37(2020) 1530-1534.

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

[24] X. Tang, C. Wu, X. Li, Y. Song, et al., On the origin and continuing evolution of SARS-CoV-2, Natl. Sci. Rev. 7 (2020) 1012–1023.

[25] E. Domingo, J. J. Holland, 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.

[26] B. Aiamkitsumrit, N.T. Sullivan, M. R. Nonnemacher, V. Pirrone, B. Wigdahl, Human immunodeficiency virus type 1 cellular entry and exit in the T lymphocytic and monocytic compartments: mechanisms and target opportunities during viral disease, Adv. Virus. Res. 93 (2015) 257-311.

[27] S. J. Spielman, S. Weaver, S. D. Shank, B. R. Magalis, M. Li, S.L.K. Pond, Evolution of viral genomes: interplay between selection, recombination, and other forces, Methods. Mol. Biol. 910(2019) 427-468.

[28] S. Kryazhimskiy, J. B. Plotkin, The population genetics of dN/dS, Plos. Genetics. 4(2008) e1000304.

[29] M. Zhang, X. Li, Z. Deng, Z. Chen, Y. Liu, Y. Gao, W. Wu, Z. Chen, Structural biology of the arterivirus nsp11 endoribonucleases, J. Virol. 91(2016) e01309-16.

[30] D. Schoeman, B. C. Fielding, Coronavirus envelope protein: current knowledge, Virol. J. 16 (2019), 1-22.

[31] W. Zeng, G. Liu, H. Ma, D. Zhao, Y. Yang, M. Liu, A. Mohammed, C. Zhao, Y. Yang, J. Xie, C. Ding, Biochemical characterization of SARS-CoV-2 nucleocapsid protein, Biochem. Biophys. Res. Commun. 527(2020) 618-623.

[32] J.Y.Li, C.H. Liao, Q. Wang, Y.J. Tan, R. Luo, Y. Qiu, X. Y. Ge, The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway, Virus. Res. 286 (2020) 198074.

[33] F. K. Yoshimoto. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19, Protein. J. 39 (2020) 198-216.

[34] Y. Ren, T. Shu, D. Wu, J. Mu, C. Wang, M. Huang, Y. Han, X. Y. Zhang, W. Zhou, Y. Qiu, X. Zhou, The ORF3a protein of SARS-CoV-2 induces apoptosis in cells, Cell. Mol. Immunol. 17(2020):881-883.

[35] R. Cagliani, D. Forni, M. Clerici, M. Sironi, Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses, Infect. Genet. Evol. 83 (2020) 104353.

[36] M. Eigen, J. McCaskill, P. Schuster, Molecular quasispecies, J. Phys. Chem. 92 (1988) 6881–6891.

[37] A. G. Bowie, L, Unterholzner, Viral evasion and subversion of pattern-recognition receptor signaling, Nat. Rev. Immunol. 8 (2008) 911–922.

[38] K. Narayanan, C. Huang, S. Makino, SARS coronavirus accessory proteins, Virus Res. 133 (2008) 113–121.[39] M. Prete, E. Favoino, G. Catacchio, et al., SARS-CoV-2 Inflammatory syndrome. clinical features and rationale for immunological treatment, Int. J. Mol. Sci. 21 (2020) 3377.

[40] Y. Qiu, K. Xu, K., Functional studies of the coronavirus nonstructural proteins, STE Medicine 1(2) (2020) e39.

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

[42] P. Clarke, K. L. Tyler, Apoptosis in animal models of virus-induced disease, Nat. Rev. Microbiol. 7 (2009) 144-155.

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

[44] W. H. Chen, U. Strych, P. J. Hotez, M. E. Bottazzi, The SARSCoV-2 vaccine pipeline: an overview, Curr. Trop. Med. Rep. 7 (2020) 61-64.

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

[46] G. Salvatori, L. Luberto, M. Maffei, et al., SARS-CoV-2 SPIKE PROTEIN: an optimal immunological target for vaccines, J. Transl. Med. 18 (2020) 222.

[47] W. Peng, B. R. Shaw, Accelerated deamination of cytosine residues in UV-induced cyclobutane pyrimidine dimers leads to CC→TT transitions. Biochemistry 35 (1996) 10172-10181..

[48] H. Hayatsu, 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 (2008) 321-330.

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

[50] S. Matallana-Surget, J. A. Meador, F. Joux, T. Douki, Effect of the GC content of DNA on the distribution of UVB-induced bipyrimidine photoproducts, Photochem. Photobiol. Sci. 7 (2008) 794–801.

[51] J. H. Miller, Mutagenic specificity of ultraviolet light, J. Mol. Biol. 182 (1985) 45–68.

[52] D. E. Brash, UV signature mutations, Photochem. Photobiol. 91 (2015) 15-26.

[53] K. Mohler, M. Ibba, Translational fidelity and mistranslation in the cellular response to stress, Nat. Microbiol. 2 (2017) 17117.

[54] L. Ribas de Pouplana, M. A. Santos, J. H. Zhu, P. J. Farabaugh, B. Javid, Protein mistranslation: friend or foe? Trends. Biochem. Sci. 39 (2014) 355–62.

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

[56] N. J. Ma, C. F. Hemez, K. W. Barber, et al., Organisms with alternative genetic codes resolve unassigned codons via mistranslation and ribosomal rescue, eLife. 7 (2018) 1–23.

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

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

[59] A. Rozov, N. Demeshkina, E, Westhof, M. Yusupov, G. Yusupova. New structural insights into translational miscoding, Trends. Biochem. Sci. 41 (2016) 798–814.

[60] I. J. Kimsey, E. S. Szymanski, W. J. Zahurancik, A. Shakya, Y. Xue, et al., Dynamic basis for dG•dT misincorporation via tautomerization and ionization, Nature 554(2018) 195-201.

Figure Legends

Figure 1. Radial trees representing the phylogenetic relationships among the different SARS-COV-2 genomes sequenced till 21 August 2020. (**A**-**D**) shows the phylogeny reconstructed based on 4,618 global sequences extracted from the universal dataset of 71,703 complete whole-genomes. (**A)** identifies and labels the clades based on the dynamic clade nomenclature system PANGOLIN (Rambaut et al. 2020). This convention currently defines 62 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)** 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 region (continent) from the sequences were obtained. (**E**-**F**) shows phylogeny based on 1,148 Indian and 4,630 global sequences extracted from the universal dataset of 71,703 complete whole-genomes. (**E**) shows only the Indian sequences, and identifies and labels the clades based on Year-Letter nomenclature system. (**F**) also shows only the Indian sequences, and identifies and labels the clades based on GISAID nomenclature system.