Mutation landscape of SARS-CoV-2 reveals three mutually exclusive clusters of leading and trailing single nucleotide substitutions

Akhilesh Mishra[$1], Ashutosh Kumar Pandey[$1], Parul Gupta[$1], Prashant Pradhan[$1,2], Sonam Dhamija[1], James Gomes[1] , Bishwajit Kundu[1], Perumal Vivekanandan*[1] and Manoj B. Menon*[1]

[1]Kusuma School of biological sciences, Indian institute of technology, New Delhi-110016, India. [2]Acharya Narendra Dev College, University of Delhi, New Delhi-110019, India

[$]Equal contribution

* Address for correspondence: vperumal@bioschool.iitd.ac.in or menon@bioschool.iitd.ac.in

Abstract

The COVID-19 pandemic has spread across the globe at an alarming rate in the last four months. However, unlike any of the previous global outbreaks the availability of hundreds of SARS-CoV-2 sequences provides us with a unique opportunity to understand viral evolution in real time. We analysed 480 full-length (>29000 nt) sequences from the 1575 SARS-CoV-2 sequences available and identified 37 single-nucleotide substitutions occurring in >1% of the genomes. Majority of the substitutions were C to T or G to A. We identify C/Gs with an upstream TTT trinucleotide motif as hotspots for mutations in the SARS-CoV-2 genome. Interestingly, two of the 37 substitutions occur within highly conserved secondary structures in the 5' and 3' untranslated regions that are critical for the virus life cycle. Furthermore, clustering analysis revealed unique geographical distribution of SARS-CoV-2 variants defined by their mutation profile. Of note, we observed several co-occurring mutations that almost never occur individually. We define 3 mutually exclusive lineages (A1, B1 and C1) of SARS-CoV-2 which account for about three quarters of the genomes analysed. We identify lineage-defining leading mutations in the SARS-CoV-2 genome which precede the occurrence of sub-lineage defining trailing mutations. The identification of mutually exclusive lineage-defining mutations with geographically restricted patterns of distribution has potential implications for diagnosis, pathogenesis and vaccine design. Our work provides novel insights on the temporal evolution of SARS-CoV-2.

Key Words: COVID-19; SARS-CoV-2; single-nucleotide substitutions; mutations; hotspot; lineage-defining mutations; leading and trailing mutations

Introduction

The COVID-19 pandemic caused by the novel coronavirus SARS-CoV-2 (SARS-like coronavirus 2) is rapidly spreading across the globe with over 3.5 million cases within a period of about 4 months (https://www.who.int/emergencies/diseases/novel-coronavirus-2019). This is far more alarming than the SARS (Severe Acute Respiratory Syndrome) epidemic of 2003 which affected 26 countries, killed 774 people, but was contained within 6 months (https://www.who.int/csr/sars/country/table2003_09_23/en/). Understanding the unique features of the SARS-CoV-2 is key to containing the COVID-19 pandemic (Wu et al., 2020). SARS-CoV and the SARS-CoV-2 are closely related RNA viruses with plus strand RNA genomes. RNA viruses usually display high mutation rates that facilitate adaptability and virulence (Sanjuan et al., 2010). The moderate mutation rates seen in SARS-CoV (Zhao et al., 2004) have been attributed to the presence of a 3'-5' exonuclease or proof-reading like activity (Minskaia et al., 2006).

The rapid global spread of SARS-CoV-2 in a short period of time and the availability of a large number of fully sequenced genomes provide us with a unique opportunity of understanding the short-term temporal evolution of this virus in humans in a near real-time scale. While several studies have used phylogenetic analysis to reveal mutation hotspots in the viral genome and identify variants (Forster et al., 2020; Koyama et al., 2020; Yeh and Contreras, 2020), we have utilized an alternative approach by focusing only on mutations present in >1% genomes followed by clustering. By this approach we propose the classification of the SARS-CoV-2 virus genomes into 3 mutually exclusive lineages with unique set of co-occurring mutations and geographic distribution.

Results and discussion

Mutation landscape of SARS-CoV-2

To understand the evolution of SARS-CoV-2 over the first three months of the pandemic, we performed a detailed analysis of full-length genomes available from the GISAID database from December 24, 2019 to March 24, 2020. A total of 1575 SARS-CoV-2 sequences were retrieved from GISAID (as on 24th March 2020). Sorting for full-length genomes resulted in a dataset of 481 sequences including the RefSeq sequence. A multiple sequence alignment was performed to visualize the variations in SARS-CoV-2 genomes. To analyse the single nucleotide substitutions in detail, we decided to focus on those which were observed in >1% of the genomes. As expected, the 5' and 3' ends of the sequences contained gaps and inconclusive nucleotide positions (Ns) attributable at least in part to the technical shortcomings of sequencing. The substitutions in the 3' and 5' UTRs were considered only when the quality of the sequences in the area is greater than 95%. Our analysis revealed a total of 37 nucleotide substitutions which occurred at > 1% in the SARS-CoV-2 genomes (Table 1 and Figure 1A). This includes 13 synonymous mutations, 21 missense mutations and 3 substitutions in the non-coding regions. ORF9/N gene encoding for the nucleocapsid phosphoprotein seems to have accumulated the maximum number (n=6) of mutations. This is also evident when substitution frequencies for each gene are normalized for the gene length (Figure 1B). Among the non-structural proteins (Nsps), the viral RNA polymerase-helicase pair (nsp12/13) accumulated seven mutations. These observations are important considering that the widely used diagnostic tests for the detection of SARS-CoV-2 target nsp12 and N genes. While, we found 5 mutations in the Nsp3 region, the substitution frequencies per unit length are rather low as Nsp3 is the longest among the non-structural genes. It is interesting to note than no mutations were observed in the Nsp5 region coding for the main protease (Mpro), a widely investigated anti-viral target for SARS-CoV-2 (Jin et al., 2020; Zhang et al., 2020). A quick scan along the 30kb genome of SARS-CoV-2 clearly suggests clustering of mutations at the 3' end of the genome (Figure 1). It would be

worth investigating differences if any, in the fidelity of the virus polymerase or exonuclease pertaining to the 3' end of the genome. All subsequent analysis presented in this study pertain to the 37 mutations occurring at >1% frequency.

Table 1. Summary of single nucleotide substitutions with >1% frequency in SARS-CoV-2 genomes

Sl. N	Refseq Positio	Substitution %	Substitution	Gene	Product	Amino Acid Substitutions
1	241	34.8	C > T	5'UTR	-	-
2	514	1.4	T > C	ORF1ab	leader protein/ nsp1	synonymous
3	1059	4.9	C > T		nsp2	T > I
4	1397	3.3	G > A			V > I
5	1440	1.2	G > A			G > D
6	2416	1.2	C > T			synonymous
7	2891	1.2	G > A		nsp3	A > T
8	3037	35.2	C > T			synonymous
9	3145	1.6	G > T			L > F
10	4402	2.4	T > C			synonymous
11	5062	2.4	G > T			L > F
12	8782	29.1	C > T		nsp4	synonymous
13	11083	12	G > T		nsp6	L > F
14	14408	34.8	C > T		RNA-dependent RNA polymerase (nsp12)	P > L
15	14805	5.9	C > T			synonymous
16	15324	2.4	C > T			synonymous
17	17247	3.5	T > C		helicase (nsp13)	synonymous
18	17373	1.4	C > T			synonymous
19	17747	16.1	C > T			P > L
20	17858	16.3	A > G			Y > C
21	18060	17.5	C > T		3'-to-5' exonuclease	synonymous
22	22277	1.2	C > A	S	surface glycoprotein	Q > K
23	22661	1.2	G > T			V > F
24	23403	35.2	A > G			D > G
25	24034	1.2	C > T			synonymous
26	25563	5.3	G > T	ORF3a	ORF3a protein	Q > H
27	26144	9.4	G > T			G > V
28	27046	10.6	C > T	M	membrane glycoprotein	T > M
29	28144	29.5	T > C	ORF8	ORF8 protein	L > S
30	28688	3.3	T > C	ORF9/ N	nucleo-capsid phosphoprotein	synonymous
31	28878	1.8	G > A			S > N
32	28881	16.5	G > A			R > K
33	28882	16.5	G > A			R > K*
34	28883	16.5	G > C			G > R
35	29095	2.4	C > T			synonymous
36	29553	1.2	G > A	intergenic	-	-
37	29742	3.2	G > A/T	3'UTR	-	-

*G28881A and G28882A occur within the same codon and they always co-occur. Amino acid annotation (R > K) is based on the co-occurrence of these mutations

The non-coding mutations fall within conserved structural elements of SARS-CoV-2 genome

The mutations identified in the non-coding regions include one each from the 5' UTR (C241T), 3' UTR (G29742A/T) and the intergenic region between ORF9 and ORF10 (G29553A) (Table 1). These mutations may have implications on viral RNA folding. Group I and II coronaviruses harbour three sub-structural hairpins SL5a, SL5b & SL5c with conserved hexameric "UUYCGU" loop motifs, which when present in multiple copies are presumed to act as packaging signals for viral encapsidation (Chen and Olsthoorn, 2010). Similar structures have been predicted in the 5'UTR of SARS-CoV-2 (Andrews et al., 2020; Chan et al., 2020). The C241T mutation changes the C-residue in the hexameric loop motif of SARS-CoV-2 SL5b (Supplementary Figure S1A & S1B). This is one of the earliest and the most prevalent (34.8%) of the SARS-CoV-2 mutations and may impact viral packaging and titres.

Interestingly, G29742 in the 3' UTR is the only position among the 37 mutations which shows two different variants, G29742T and G29742A (Table 1). This residue is part of the stem-loop II-like motif (s2m) highly conserved in SARS-like coronaviruses (Supplementary Figure S1C). SARS-CoV s2m sequence assumes a unique RNA fold similar to that seen in ribosomal RNA and is thought to be of regulatory function (Robertson et al., 2005). The G29742 occupies a conserved position critical for the unique kinked structure and the substitutions at this position are expected to alter the s2m stem-loop (Supplementary Figure S1D). Recent studies have also reported other low frequency mutations in this region (Yeh and Contreras, 2020).

C to T and G to A substitutions predominate in the SARS-CoV-2 genomes

Transitions are more common than transversions across mammalian genomes as well as viruses (Duchene et al., 2015; Petrov and Hartl, 1999; Rosenberg et al., 2003). Analysis of substitutions in Influenza A virus and HIV-1 suggest that transversions are more deleterious than transitions (Lyons and Lauring, 2017). Hence, it is not surprising that 28 of the 37 substitutions we report here for SARS-CoV-2 are transitions. Substitutions at one of the nucleotide positions include both transition (G29742A) and transversion (G29742T). Thus the transitional mutation bias in SARS-CoV-2 is in keeping with that reported for other viruses (Duchene et al., 2015).

C to T substitutions account for about 40% (n=14) followed by G to A substitution which account for about 20% (n=8) of the 37 substitutions (Figure 2A & B). The predominance of C to T (U) substitutions in SARS-CoV-2 has been documented recently (Koyama et al., 2020; Wang et al., 2020). Amongst host RNA editing enzymes, APOBECs (Apolipoprotein B mRNA editing complex) and ADARs (Adenosine Deaminases Acting on RNA) have been studied for their ability to edit virus genomes (Cheng et al., 2019; Piontkivska et al., 2017). Both negative-strand intermediates and genomic RNA (positive sense) may be available as ssRNA inside coronavirus infected host cells. The negative-strand intermediates serve as templates for the synthesis of the positive strand (genomic RNA) as well as the sub-genomic transcripts. In this process the negative-strand intermediates will progressively lose the ssRNA conformation and gain dsRNA conformation. In essence, both ssRNA and dsRNA forms will be available for RNA editing by host enzymes. This is important considering that RNA editing APOBECs specifically act on ssRNA to induce C>U transitions and ADARs target dsRNA resulting in A>I (inosine; read as G) editing. APOBEC-mediated editing of the genomic RNA (positive strand) will result in C to U changes, while the editing of the negative-strand intermediates (C to U) will result in G to A changes in the genomic RNA (positive strand). It is well documented that the abundance of the positive sense RNA outnumbers that of the negative strand RNA in infected host cell (Sethna et al., 1991). Our finding of predominance of C to T (C to U) mutations (38.9%) over G to A mutations (19.4%) is consistent with the abundance of genomic RNA (positive sense) over the negative-strand intermediates.

Similarly, ADAR-mediated editing of the genomic RNA (positive sense) will result in A to G changes and the editing of the negative-strand (A to G) will result in U to C changes in the genomic RNA (positive sense). Our findings indicate that the substitutions which are consistent with ADAR editing (A to G and U to C), constitute about 21.1 % (8 of 38 substitutions) of those observed in SARS-CoV-2 genomes analysed. The predominance of APOBEC-like RNA editing (i.e C to U and G to A) over ADAR-like RNA editing (i.e. A to G and U to C) in SARS-CoV-2 may be explained at least in part by the transient nature of virus dsRNA available inside infected host cells. Editing of SARS-CoV-2 RNA by APOBECs and ADARs has been recently reported (Di Giorgio et al., 2020). In addition, APOBEC3-mediated virus restriction has been documented for the human coronavirus HCoV-NL63 (Milewska et al., 2018).

Substitutions in SARS-CoV-2 lead to the loss of GG dinucleotides and gain of TT dinucleotides

We also analysed the dinucleotide context of the mutations described. For this purpose, we considered nucleotides flanking the substitutions described in table 1. For example, a C>T mutation in the trinucleotide sequence 5`-ACG-3` sequence will result in 5`-ATG-3` trinucleotide. In this context an AC and a CG dinucleotide are lost and an AT and a TG dinucleotide are gained. In other words, each substitution will lead to a loss of 2 dinucleotides and a gain of 2 dinucleotides. We analysed the 36 substitutions (G29742A/T was excluded) to assess the net loss or gain for each of the 16 dinucleotides. Interestingly, the 36 substitutions were associated with a loss of 10 GG dinucleotides and a gain of 13 TT dinucleotides (Figure 2C). The dinucleotides lost or gained are influenced not only by the nature of the substitutions (eg. C to T or G to A) but also by the nucleotides flanking the substitution site. Furthermore, dinucleotides provide important clues about virus pathogenesis (Sankar et al., 2018; Wasson et al., 2017). In particular, CpG dinucleotides have been associated with virus replication, immune response and virus pathogenesis (Fros et al., 2017; Simmonds et al., 2015). A recent paper suggests that SARS-CoV-2 is extremely CpG depleted to avoid host defences (Xia, 2020). Virus evolution has been also linked to variations in other dinucleotides including TpA and GpT (Simmonds et al., 2013; Upadhyay et al., 2013). Previous reports suggest that coronaviruses have excess of TT dinucleotides and are depleted for GG dinucleotides (Cheng et al., 2013); this is consistent with major dinucleotide gain/loss in SARS-CoV-2 genomes (Figure 2C) The biological implications of these findings remain to be determined.

Identification of TTTC and TTTG as hotspots for mutations in the SARS-CoV-2 genome

Signatures of APOBEC editing (C to T and G to A) dominate the substitutions observed in the SARS-CoV-2 genomes (Figure 2A). A recent study on RNA editing of SARS-CoV-2 genome shows evidence for APOBEC editing (Di Giorgio et al., 2020). Nonetheless, the specific motifs for APOBEC-mediated editing of virus genomes remain poorly understood. The presence of 14 C to T mutations and 8 G to A mutation among the 37 mutations we report gave us an opportunity to analyse preferences, if any for specific dinucleotide and trinucleotide motifs flanking these substitutions. While we found no such preferential motifs flanking the C to T or G to A substitution sites, we found interesting upstream trinucleotide motifs when all possible C (C>T, C>A and C>G) & G (G>A, G>T and G>C) substitution sites were considered together We found TTT trinucleotides upstream of 26.7 % of all substitutions (8 out of 30) occurring at C or G residues (C/Gs) (Figure 2D). We did not find any specific trinucleotide preference downstream to these sites (data not shown). There were no specific di or trinucleotide preferences upstream of substitutions occurring at A/Ts (Figure 2E).

Intrigued by the preference for TTT trinucleotide upstream of substitutions occurring at C/Gs, we analysed the statistical significance of this finding. Briefly, we generated 10000 random numbers corresponding to

nucleotide positions in the SARS-CoV-2 genome and found 3826 positions with C/G. We then analysed the frequency of TTT trinucleotides upstream of these 3826 positions. The TTT trinucleotide frequency immediately upstream of substitutions originating at C/Gs is significantly higher than that upstream of randomly generated nucleotide positions (26.7 % vs 3.6 %; $p<0.00001$; Figure 2F). This finding indicates that TTT trinucleotides followed by C/G (i.e. TTTC or TTTG) in the SARS-CoV-2 genome represent hotspots for substitutions. To the best of our knowledge, tetranucleotide motifs predisposed to higher substitution rates have not been reported for SARS-CoV-2. Furthermore, TTT trinucleotides were not detected upstream of any substitution occurring at As or Ts (n=7) (Figure 2E).

While TTTG/TTTC have not been reported as hotspots among viruses, TTTG has been reported as a hotspot for G to T substitutions in yeast deficient in nucleotide excision repair (Abdulovic et al., 2008). TTC has been identified as a hotspot for APOBEC editing among gamma herpesviruses, which are dsDNA viruses (Martinez et al., 2019). The specific mechanism associated with increased substitutions at TTTG/TTTC motifs in the SARS-CoV-2 genome is unclear. As negative-strand intermediates of SARS-CoV-2 serve as the template for the synthesis of positive-sense genomic RNA, we speculate that the virus RdRP (RNA-dependent RNA polymerase, nsp12) may be more error prone at C/Gs following a homopolymeric stretch of Ts (i.e. TTT). We cannot rule out a role for other proteins involved in genomic RNA replication including the nsp13 helicase and the nsp14 exonuclease in this process. Nonetheless, the identification of TTTG and TTTC as hotspots for substitutions opens up a plethora of opportunities for research on identification of the underlying virus /host factors.

The longest homopolymeric stretch in the entire SARS-COV-2 genome is an octamer of Ts (i.e.TTTTTTTT) and the G at the end of this octamer (i.e. TTTTTTTTG) is one of the substitution sites (G11083T; present in 12% of sequences analyzed)(refer Table 1). More interestingly, a homopolymeric stretch of GGGG (nt.28881 to 28884 in the Refseq) is associated with a triple mutant (i.e. GGGG to AACG; present in 16% of sequences analysed). All the three Gs are either mutated together or they remain wild-type in the 480 genomes analysed. In addition, among the 37 substitutions we report, only the G28881A and the G28882A represent mutations within a single codon (R>K) in SARS-CoV-2. Multiple substitutions within a single codon represent positive selection of an amino acid (Bazykin et al., 2006). Further, positively selected amino acids are frequently present in virus proteins involved in critical functions such as receptor binding (Bush et al., 1999).

Nucleotide substitutions define several clusters of SARS-CoV-2 genomes

Several studies have looked into the aspect of real-time SARS-CoV-2 evolution and strain diversification by using phylogenetic analyses (Forster et al., 2020; Koyama et al., 2020; Yeh and Contreras, 2020). In contrast to this approach, we utilized our catalogued set of 38 single nucleotide substitutions (36 substitutions + G29742T and G29742A) to understand the emergence of SARS-CoV-2 variants. A clustering analysis was performed to systematically group all 480 full-length genomes into subtypes which are defined by the set of mutations they harbor. After removing groups which consisted of <5 genomes, the analysis revealed 24 distinct clusters (Figure 3A). This analysis identified very interesting geographical distribution of the clusters across the globe (Figure 3B). For example, cluster-1 (T28144C, C8782T, C18060T, A17858G, C17747T) accounted for more than 15% of the genomes analysed and was almost exclusively reported from North America (NA). A similar pattern was observed for cluster 16, which was also restricted to NA. Several clusters including cluster -2, -3, -5, -13, -18, -19, -20 and -23 were unique to the European countries (EU). A large number of EU specific clusters could be partially due to the over representation of genomic sequences from EU (196 out of 480).

Identification of three mutually exclusive lineages and sequential mutation trails of SARS-CoV-2

Our clustering analysis revealed several mutations which almost always co-occur. We define co-occurring mutations as those which occur together in >90% of their individual occurrences. For example, if mutation A occurs 50 times and mutation B occurs 51 times in the dataset and mutations A and B are present together in the same genomes (i.e. co-occur) 50 times, they are defined as co-occurring mutations as 50/51 is >90%. Accumulation of mutations is key to the diversification of viral lineages. In addition, the co-occurrence of mutations in virus genomes is often suggestive of compensatory mutations. Individual mutations as well as co-occurring mutations can act as lineage defining mutations. We consider a specific mutation or a set of co-occurring mutations as "lineage-defining" for SARS-CoV-2, only when they are present in at least 5% (n=24) of the sequences analysed. Each lineage-defining mutation(s) are mutually exclusive and are not present along with another lineage-defining mutation. Our analyses reveal three mutually exclusive lineages for SARS-CoV-2. We refer to these three lineages as A1, B1 and C1 in the chronological order of their appearance (Figure 4 and Supplementary Figure 3. The A1 lineage (co-occurrence of C8782T, T28144C) appeared as early as 5th January 2020 in Asia. The presence of G26144T defines our B1 lineage. This mutation was first observed in samples collected from North America on 22nd January 2020. The co-occurrence of C241T, C3037T, C14408T, A23403G defines our lineage C1 that was first detected on 20th February 2020 in Italy, Europe.

Leading and trailing mutations have been previously described in virus genomes (Kryazhimskiy et al., 2011; Neverov et al., 2015). We define leading mutations as mutations that are documented in SARS-CoV-2 at an earlier time point and their presence is a pre-requisite for the origination of trailing mutations. In other words, trailing mutations never occur in the absence of leading mutations. The lineage-defining mutations A1, B1 and C1 represent leading mutations in the evolution of SARS-CoV-2. The chronology of the first appearance of all the 37 substitutions in the SARS-CoV-2 genome is summarised in Supplementary table 1. The A1 lineage accumulates two independent trailing mutation sets giving rise to A2 (A1+T4402C, G5062T) and A3 (A1+ C17747T, A17858G, C18060T) variants of SARS-CoV-2. While A2 was reported from South Korea by the end of January, A3 first appeared on 20th February 2020 from the USA and accounts for >15% of the 480 genomes. The B1 lineage acquired a trailing synonymous mutation in the Nsp13 helicase (T17247C) to give rise to the B2 sub-lineage (B1+ T17247C) reported from South America in the end of February, 2020. While the A1 lineage is primarily restricted to Asia and North America, majority of the sequences in the B1 lineage were reported from Europe (Supplementary Figure S2).

The most diverse pattern of trailing mutations has appeared in the C1 lineage, which incidentally is the most recent of the lineage-defining mutations of SARS-CoV-2. Moreover, the lineage-defining mutations in the C1 variant include missense mutations in the Nsp12/RNA polymerase (C14408T, P>L) and the S gene (A23403G, D>G), in addition to the SL5b loop mutation (C241T, Figure S1). While C2 (C1+G28881A, G28882A, G28883C), C5 (C1+G25563T), C6 (C1+C1059T) and C7 (C1+G29553A) emerged from C1 when the virus acquired specific set of trailing mutations, C3 (C2+C27046T) originated from the C2 variant upon the gain of a missense mutation (C27046T, T>M) in the viral M protein. A small number of genomes (n=6) also display a co-occurring mutation profile designated as the C4 variant (C3+C22277A) first reported in early March 2020 in Europe. Interestingly, the C1 lineage is mostly restricted to Europe and North America (Supplementary Figure S2) and involves incorporation of missense mutations in the viral N (C2 trailing mutations C28881A, G28882A, G28883C), M (C3 trailing mutation C27046T), S (C4 trailing mutations C22277A), ORF3a (C5 trailing mutation G25563T) and Nsp2 (C6 trailing mutation C1059T) genes (Figure 4 and Table 1). Within the C1 lineage, we could observe interesting bifurcation based on geographical distribution with C2 and C3

prevalent in Europe and South America and C5-C6 in Europe and North America. C4 and C7 were exclusively found in Europe and North America respectively (Supplementary Figure S2).

Dynamics of single-nucleotide substitutions in the SARS-CoV-2 genome

The constantly updating list of SARS-CoV-2 sequences provides a unique opportunity for us to understand the evolution of this virus in humans in the course of the global outbreak. To understand the dynamics of different mutations over the evolution of the virus, we compared full-length genome datasets downloaded from GISAID on 31st January (n=50), 5th March (n=171), 9th March (n=213) and 24th March 2020 (n= 480). The mutations which show significant change in their frequency over the first three months of SARS-CoV-2 evolution are shown in Figure 5. The data is filtered for mutations which appeared in >5% of the sequences in at least one of the datasets. Interesting patterns emerge for different mutations spread over the four datasets (Figure 5). Interestingly, the frequency of the lineage-defining A1 mutation increases to about one third of the sequences in the second dataset, but plateau from the second to fourth dataset. This would mean that the A1 lineage is consistently contributing to about a third of all newly reported sequences. The C1 lineage profile follows a logarithmic rise between the 3rd and 4th datasets. The first case of this leading mutation appeared on the 20th of February and by the 18th of March it was among the most frequent mutations (164 out of 480; >34%). Consistent with the appearance of the trailing mutations later in the timeline, trailing mutations for A3 (denoted as A3'), C2 (C2'), C3 (C3') and C5 (C5') sharply increase in numbers between the third and fourth dataset. The G11083T variant displays a profile similar to A1, despite with much lower prevalence and plateauing at approximately 12%. C14805T (Nsp12/RNA polymerase synonymous) is another mutation which also show an upward tendency reaching 5.9% in the fourth dataset.

The B1 leading mutation shows a marginal downward trend in frequency over time (Figure 5B). Interestingly, we also identified substitutions (C29095T & C21707T) which appear early in the timeline (Jan 2020) in >5% of the genomes, but their frequencies rapidly decline to <1% in the final dataset.

Conclusions

We focused our analyses on single nucleotide substitutions and identified 37 substitutions with >1% frequency in the SARS-CoV-2 genomes. Our analysis show that the nucleotide changes in the identified 37 substitution sites are non-random (predominant C to T/U and G to A) and are partially determined by the upstream sequence motifs. A significant proportion of analysed substitutions are consistent with APOBECs/ADARs editing. Clustering analysis revealed unique geographic distributions of SARS-CoV-2 variants defined by their mutation profile. Interestingly, we observed several co-occurring mutations that almost never occur individually. Our analyses reveal 3 mutually exclusive lineages of SARS-CoV-2 which account for about three quarters of the genomes analysed. Lineage-defining leading mutations in the SARS-CoV-2 genome precede the occurrence of trailing mutations that further define sub-lineages. The lineage-defining leading mutations together with the subsequently acquired trailing set of mutations provides a novel perspective on the temporal evolution of SARS-CoV-2

Methods

Sequence alignment and refinement of the SARS-CoV-2 genome sequences

The genomic Sequences of SARS-CoV-2 from clinical samples were downloaded from the GISAID database (https://www.gisaid.org/) {Elbe, 2017 #182}. As of 24th March 2020, GISAID consisted of 1575 SARS-CoV-2 genomes. Only samples from human host with nearly full-length SARS-CoV-2 genomic sequences (>29000 nt) were included in the analyses. The sequences flagged for low coverage / sequence quality (example: high frequency of Ns) by the GISAID server were excluded. The sequences were downloaded at different time periods in the past three months and fall into four serial datasets with batch download dates of 31/01/2020 (n=50), 05/03/2020 (n=171), 09/03/2020 (n=213) and 24/03/2020 (n=480). The accession IDs of the genome sequences are provided as Supplementary table 2. The Alignment and refinement of the 480 sequences with the SARS-CoV-2 reference genome were performed by using MUSCLE multiple sequence alignment software (Edgar, 2004). The average alignment scores for the genomic sequences in the dataset 1 (n= 50+1 reference sequence genome), dataset 2 (n=172), dataset 3 (n=214) and dataset 4 (n=481) were 99.83%, 99.65%, 99.50% and 99.00 % respectively.

Identification of single nucleotide substitutions in the SARS-CoV-2 genome

Alignment positions which harbor defined A/T/G/C residues (rather than gaps or `n`) in 95% of the aligned genomes were only considered for identifying nucleotide substitutions. These filtered nucleotide positions were analysed and scored for A/T/G/C occupancy in comparison to the SARS-CoV-2 reference genome (Accession: NC_045512.2). We then calculated the percentage of sequences with single-nucleotide variations for each nucleotide position in the SARS-CoV-2 genome. Substitutions with >1% frequency in the 480 genome dataset were included for further analysis.

Nucleotide, dinucleotide and trinucleotide variation analysis at substitution sites

The nucleotide changes at the 37 positions were analysed and tabulated. Similar analyses were done to calculate the gain and loss in frequencies of specific dinucleotide sequences at the substitution sites. We also analysed the trinucleotide context upstream and downstream of the substitution sites (-3 to +3 genomic positions). To understand the probability of finding TTT trinucleotide upstream of a random C/G position, we first mapped 10000 random positions (based on random numbers generated in MS Excel) on the SARS-CoV-2 genome and identified 3826 G/C residues. We the calculated the TTT trinucleotide frequency upstream to these 3826 sites and compared it to the actual frequencies at the substitution sites using a *chi*-squared test.

Clustering analysis and defining leading and trailing mutations

Clustering was performed on the 480 SARS-CoV-2 sequences with a custom script written using Python programming language and the data was visualized using Seaborn Statistical Visualization Tool (https://seaborn.pydata.org/). The 38 substitutions (36 substitutions + G29742T and G29742A) were used for the generation of mutation groups which co-occurred in many genomes. There were also sequences in which none of the selected mutations were present and hence formed the 'no-mutation group'. In the second step, sequences harboring the same mutation groups were clustered together. Clusters with less than 5 genome sequences were excluded to obtain a list of 24 clusters represented as a heat map. The clusters were correlated with their geographical location (obtained from GISAID) and represented.

The datasets and clusters were analysed to identify co-occurring, lineage-defining, leading and trailing mutations. We classify mutations into the following categories:

- Co-occurring mutations: We define two (or more) mutations as co-occurring when they occur together in >90% of their individual occurrences.
- Leading and trailing mutations: we define leading mutations as mutations that are documented in SARS-CoV-2 genomes at an earlier time point and their presence is a pre-requisite for the development of trailing mutations. The trailing mutations never occur in the absence of the leading mutations. In view of possible sequencing errors, a maximum of 1 appearance of the trailing mutation(s) in the absence of the leading mutation(s) is tolerated.
- Lineage-defining mutations include both co-occurring mutations and singlet mutations that are present in >5% (n=24) of sequences analysed. Lineage-defining mutations are mutually exclusive (absolutely never occur together).

RNA sequence alignment and structure analysis

RNA sequences were aligned using Multalin software (http://multalin.toulouse.inra.fr/multalin) (Corpet, 1988). The 5` UTR stem-loop structures were predicted using the Forna server (Kerpedjiev et al., 2015)(http://rna.tbi.univie.ac.at/forna/). The PDB structure for s2m stem-loop motif from SARS-CoV was downloaded from the Protein Data Bank (http://www.rcsb.org/pdb)(PDB ID: 1XJR) and was visualized using UCSF Chimera software (version 1.14)(Pettersen et al., 2004).

Acknowledgements

We gratefully acknowledge the Authors from the Originating laboratories responsible for obtaining the specimens and the Submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based. We acknowledge Kusuma School of Biological Sciences, Indian Institute of Technology Delhi for infrastructural support. The authors thank IIT Delhi HPC facility for computational resources.

References

Abdulovic, A. L., Minesinger, B. K., and Jinks-Robertson, S. (2008). The effect of sequence context on spontaneous Polzeta-dependent mutagenesis in Saccharomyces cerevisiae. Nucleic acids research *36*, 2082-2093.

Andrews, R. J., Peterson, J. M., Haniff, H. F., Chen, J., Williams, C., Greffe, M., Disney, M. D., and Moss, W. N. (2020). An in silico map of the SARS-CoV-2 RNA Structurome. bioRxiv, 2020.2004.2017.045161.

Bazykin, G. A., Dushoff, J., Levin, S. A., and Kondrashov, A. S. (2006). Bursts of nonsynonymous substitutions in HIV-1 evolution reveal instances of positive selection at conservative protein sites. Proceedings of the National Academy of Sciences of the United States of America *103*, 19396-19401.

Bush, R. M., Fitch, W. M., Bender, C. A., and Cox, N. J. (1999). Positive selection on the H3 hemagglutinin gene of human influenza virus A. Molecular biology and evolution *16*, 1457-1465.

Chan, J. F.-W., Kok, K.-H., Zhu, Z., Chu, H., To, K. K.-W., Yuan, S., and Yuen, K.-Y. (2020). Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerging Microbes & Infections *9*, 221-236.

Chen, S. C., and Olsthoorn, R. C. (2010). Group-specific structural features of the 5'-proximal sequences of coronavirus genomic RNAs. Virology *401*, 29-41.

Cheng, A. Z., Yockteng-Melgar, J., Jarvis, M. C., Malik-Soni, N., Borozan, I., Carpenter, M. A., McCann, J. L., Ebrahimi, D., Shaban, N. M., Marcon, E.*, et al.* (2019). Epstein-Barr virus BORF2 inhibits cellular APOBEC3B to preserve viral genome integrity. Nature microbiology *4*, 78-88.

Cheng, X., Virk, N., Chen, W., Ji, S., Ji, S., Sun, Y., and Wu, X. (2013). CpG usage in RNA viruses: data and hypotheses. PloS one *8*, e74109.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. Nucleic acids research *16*, 10881-10890.

Di Giorgio, S., Martignano, F., Torcia, M. G., Mattiuz, G., and Conticello, S. G. (2020). Evidence for RNA editing in the transcriptome of 2019 Novel Coronavirus. bioRxiv, 2020.2003.2002.973255.

Duchene, S., Ho, S. Y., and Holmes, E. C. (2015). Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. BMC evolutionary biology *15*, 36.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics *5*, 113.

Forster, P., Forster, L., Renfrew, C., and Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences *117*, 9241-9243.

Fros, J. J., Dietrich, I., Alshaikhahmed, K., Passchier, T. C., Evans, D. J., and Simmonds, P. (2017). CpG and UpA dinucleotides in both coding and non-coding regions of echovirus 7 inhibit replication initiation post-entry. eLife *6*.

Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C.*, et al.* (2020). Structure of M(pro) from COVID-19 virus and discovery of its inhibitors. Nature.

Kerpedjiev, P., Hammer, S., and Hofacker, I. L. (2015). Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics *31*, 3377-3379.

Koyama, T., Platt, D., and Parida, L. (2020). Variant analysis of COVID-19 genomes. World Health Organ *Preprint*.

Kryazhimskiy, S., Dushoff, J., Bazykin, G. A., and Plotkin, J. B. (2011). Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins. PLoS genetics *7*, e1001301.

Lyons, D. M., and Lauring, A. S. (2017). Evidence for the Selective Basis of Transition-to-Transversion Substitution Bias in Two RNA Viruses. Molecular biology and evolution *34*, 3205-3215.
Martinez, T., Shapiro, M., Bhaduri-McIntosh, S., and MacCarthy, T. (2019). Evolutionary effects of the AID/APOBEC family of mutagenic enzymes on human gamma-herpesviruses. Virus evolution *5*, vey040.
Milewska, A., Kindler, E., Vkovski, P., Zeglen, S., Ochman, M., Thiel, V., Rajfur, Z., and Pyrc, K. (2018). APOBEC3-mediated restriction of RNA virus replication. Scientific reports *8*, 5960.
Minskaia, E., Hertzig, T., Gorbalenya, A. E., Campanacci, V., Cambillau, C., Canard, B., and Ziebuhr, J. (2006). Discovery of an RNA virus 3'->5' exoribonuclease that is critically involved in coronavirus RNA synthesis. Proceedings of the National Academy of Sciences of the United States of America *103*, 5108-5113.
Neverov, A. D., Kryazhimskiy, S., Plotkin, J. B., and Bazykin, G. A. (2015). Coordinated Evolution of Influenza A Surface Proteins. PLoS genetics *11*, e1005404.
Petrov, D. A., and Hartl, D. L. (1999). Patterns of nucleotide substitution in Drosophila and mammalian genomes. Proceedings of the National Academy of Sciences of the United States of America *96*, 1475-1479.
Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. Journal of computational chemistry *25*, 1605-1612.
Piontkivska, H., Frederick, M., Miyamoto, M. M., and Wayne, M. L. (2017). RNA editing by the host ADAR system affects the molecular evolution of the Zika virus. Ecology and evolution *7*, 4475-4485.
Robertson, M. P., Igel, H., Baertsch, R., Haussler, D., Ares, M., Jr., and Scott, W. G. (2005). The structure of a rigorously conserved RNA element within the SARS virus genome. PLoS biology *3*, e5.
Rosenberg, M. S., Subramanian, S., and Kumar, S. (2003). Patterns of transitional mutation biases within and among mammalian genomes. Molecular biology and evolution *20*, 988-993.
Sanjuan, R., Nebot, M. R., Chirico, N., Mansky, L. M., and Belshaw, R. (2010). Viral mutation rates. Journal of virology *84*, 9733-9748.
Sankar, S., Borkakoti, J., Ramamurthy, M., Nandagopal, B., Vivekanandan, P., and Gopalan, S. (2018). Identification of tell-tale patterns in the 3' non-coding region of hantaviruses that distinguish HCPS-causing hantaviruses from HFRS-causing hantaviruses. Emerging microbes & infections *7*, 32.
Sethna, P. B., Hofmann, M. A., and Brian, D. A. (1991). Minus-strand copies of replicating coronavirus mRNAs contain antileaders. Journal of virology *65*, 320-325.
Simmonds, P., Tulloch, F., Evans, D. J., and Ryan, M. D. (2015). Attenuation of dengue (and other RNA viruses) with codon pair recoding can be explained by increased CpG/UpA dinucleotide frequencies. Proceedings of the National Academy of Sciences of the United States of America *112*, E3633-3634.
Simmonds, P., Xia, W., Baillie, J. K., and McKinnon, K. (2013). Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla –selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. BMC Genomics *14*, 610.
Upadhyay, M., Samal, J., Kandpal, M., Vasaikar, S., Biswas, B., Gomes, J., and Vivekanandan, P. (2013). CpG Dinucleotide Frequencies Reveal the Role of Host Methylation Capabilities in Parvovirus Evolution. Journal of virology *87*, 13816-13824.
Wang, C., Liu, Z., Chen, Z., Huang, X., Xu, M., He, T., and Zhang, Z. (2020). The establishment of reference sequence for SARS-CoV-2 and variation analysis. Journal of medical virology.
Wasson, M. K., Borkakoti, J., Kumar, A., Biswas, B., and Vivekanandan, P. (2017). The CpG dinucleotide content of the HIV-1 envelope gene may predict disease progression. Scientific reports *7*, 8162.
Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J.*, et al.* (2020). Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell host & microbe *27*, 325-328.

Xia, X. (2020). Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. Molecular biology and evolution.
Yeh, T. Y., and Contreras, G. P. (2020). Faster de novo mutation of SARS-CoV-2 in shipboard quarantine. Bull World Health Organ *Preprint*.
Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K., and Hilgenfeld, R. (2020). Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors. Science *368*, 409-412.
Zhao, Z., Li, H., Wu, X., Zhong, Y., Zhang, K., Zhang, Y. P., Boerwinkle, E., and Fu, Y. X. (2004). Moderate mutation rate in the SARS coronavirus genome and its implications. BMC evolutionary biology *4*, 21.

Figure Legends

Figure 1. Mutation landscape of SARS-CoV-2. A. A schematic representation of the SARS-CoV-2 genome with gene/polypeptide annotations based on the NCBI reference sequence. The 37 substitution (>1%) sites are indicated with grey bars. The height of the bars indicates the percentage frequency of each mutation in the 480 genome dataset and mutations that occur at >25% are labelled. **B.** Substitutions per unit length (1000 nt) per gene are calculated and plotted.

Figure 2. Nucleotide substitutions in the SARS-CoV-2 genomes are non-random. A. Percentage of specific nucleotide changes among the 37 substitutions showing higher proportion of C>T and G>A changes. **B.** When mutations occurring at >10% (n=14) and <10% (n=23) are separately analysed, the distribution shows positive correlation between mutation frequency and enrichment of C>T substitutions. **C.** The net gain/loss of specific dinucleotides at the 36 single substitution sites (G29742T/A excluded) are calculated and plotted. **D & E.** The trinucleotide sequences upstream of C/G (D) and A/T (E) nucleotide substitution sites are shown. **F.** The frequency of appearance of TTT trinucleotides upstream of C/G substitution sites are compared to that of random C/G sites in the SARS-CoV-2 genome.

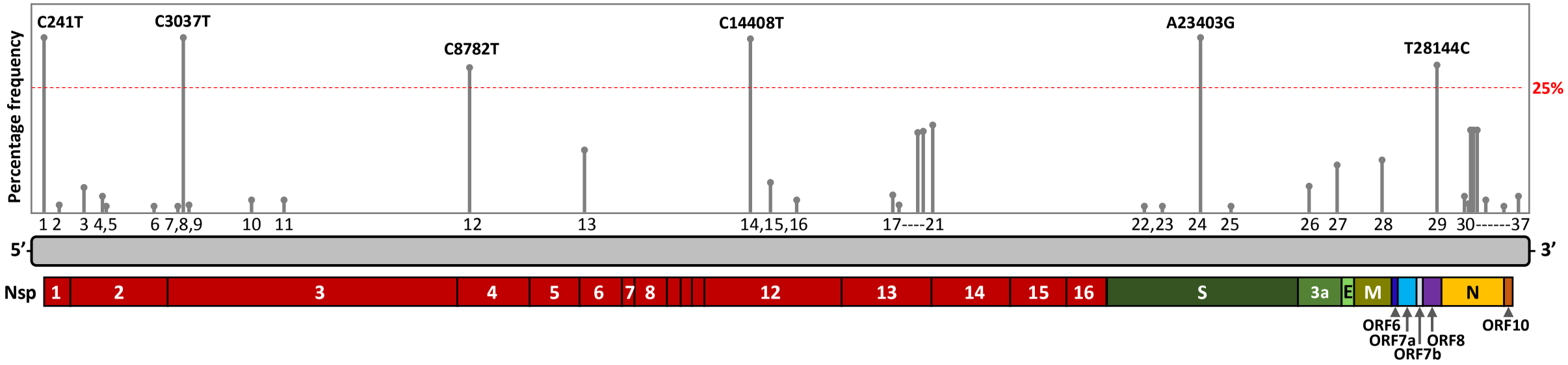
Figure 3. Clusters of SARS-CoV-2 genomes and their geographical distribution. A. Heat-map of the 24 clusters, generated from 480 full-length genome sequences based on 38 single nucleotide substitutions. Color of the plot indicates the number of sequences with a given nucleotide substitutions present within the cluster. A specific mutation cluster is represented only if there are at least 5 full-length sequences. **B.** Heat-map showing the percentage distribution of clusters across different geographical regions (AS = Asia, EU= Europe, NA = North America, SA = South America and OC = Oceania).

Figure 4. Three mutually exclusive lineages of SARS-CoV-2 and temporal acquisition of sequential mutations. Schematic representation of the COVID-19 pandemic time-line showing the leading mutations (A1, B1 and C1) and the sequential appearance of trailing mutations giving rise to additional SARS-CoV-2 variants (A2-A3,B2, C2-C7). The date and country/continent of origin of sequences in which the indicated set of co-occurring mutations were first reported are shown. *First reported on the same date from Netherlands (EU), Finland (EU) and Brazil (S. America).

Figure 5. SARS-CoV-2 mutation frequencies over the first 3 months of the pandemic. We compared the percentage fraction of high frequency mutations (>5% in at least one dataset) in four sequential datasets consisting of 51, 172, 214 and 481 genomes of SARS-CoV-2 (including the reference genome sequence). **A.** Some mutations including the A1 and C1 leading mutations appear very early in the virus evolution and are present in over 60% of the sequences in the most recent dataset. The trailing mutations of A3, C2, C3, C5 are denoted as A3', C2', C3', C5' respectively and these do not include their leading mutations. For example, A3' consists of A3 defining mutations alone and does not include its leading mutation set A1. A3', C2', C3' and C5' as well as the C14805T and G11083T mutations are increasing in frequency over time. **B.** C29095T and C21707T (<1% in the 481 sequence dataset, hence not described earlier in the manuscript) substitutions rapidly declined with time. The B1 leading mutation shows significant presence in all the four datasets.

Figure 1

A


C241T
C3037T
C8782T
C14408T
A23403G
T28144C
25%
Percentage frequency
1 2
3 4,5
6 7,8,9
10
11
12
13
14,15,16
17----21
22,23
24
25
26
27
28
29
30------37
5'
3'
Nsp
1
2
3
4
5
6
7
8
12
13
14
15
16
S
3a
E
M
N
ORF6
ORF7a
ORF7b
ORF8
ORF10

B

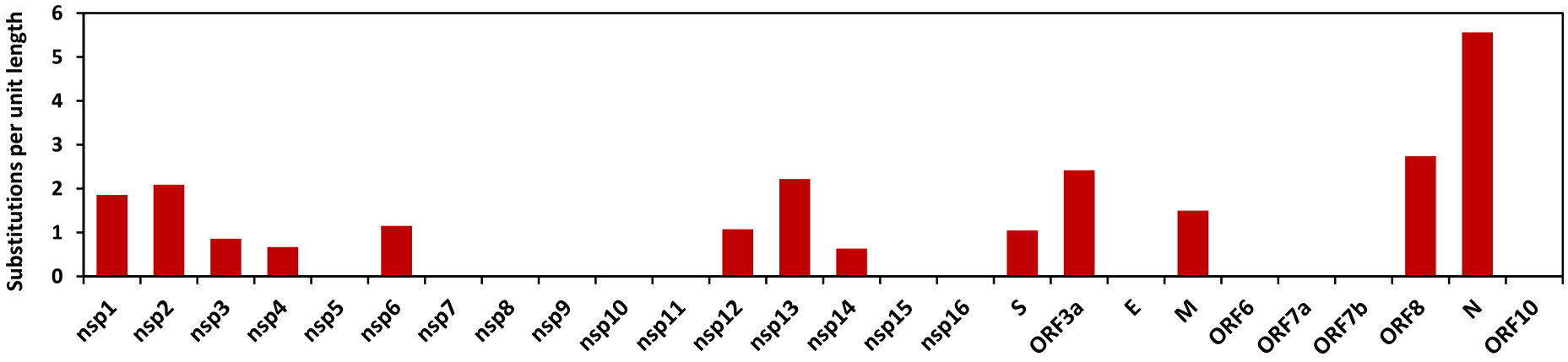

Substitutions per unit length
0
1
2
3
4
5
6
nsp1
nsp2
nsp3
nsp4
nsp5
nsp6
nsp7
nsp8
nsp9
nsp10
nsp11
nsp12
nsp13
nsp14
nsp15
nsp16
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
ORF10

Figure 2

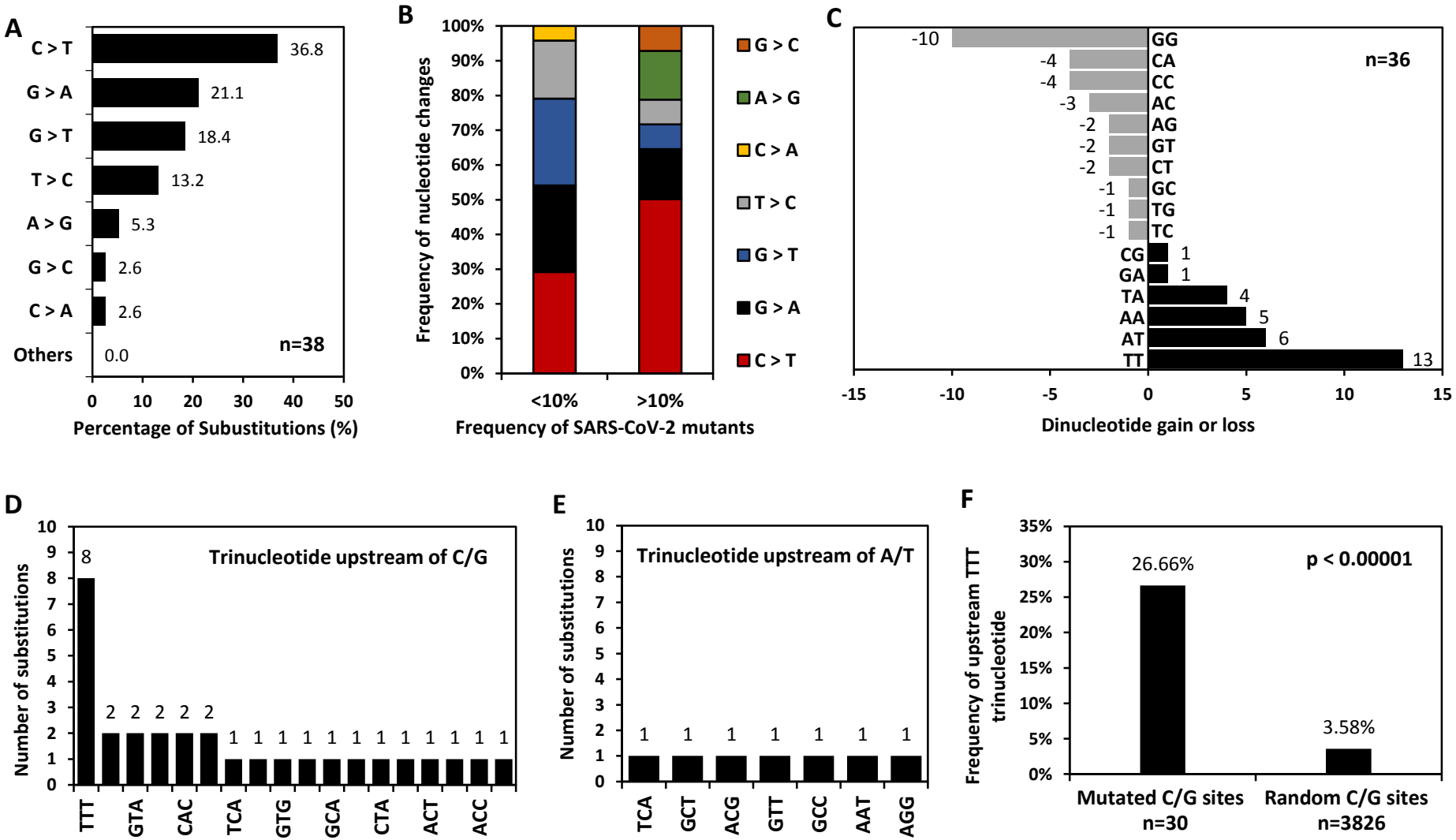

A
C > T
G > A
G > T
T > C
A > G
G > C
C > A
Others
36.8
21.1
18.4
13.2
5.3
2.6
2.6
0.0
n=38
0 10 20 30 40 50
Percentage of Subustitutions (%)
B
Frequency of nucleotide changes
100% 90% 80% 70% 60% 50% 40% 30% 20% 10% 0%
<10%
>10%
Frequency of SARS-CoV-2 mutants
G > C
A > G
C > A
T > C
G > T
G > A
C > T
C
n=36
GG -10
CA -4
CC -4
AC -3
AG -2
GT -2
CT -2
GC -1
TG -1
TC -1
CG 1
GA 1
TA 4
AA 5
AT 6
TT 13
-15 -10 -5 0 5 10 15
Dinucleotide gain or loss
D
Trinucleotide upstream of C/G
Number of substitutions
TTT 8
GTA 2
CAC 2
TCA 2
GTG 2
GCA 1
CTA 1
ACT 1
ACC 1
E
Trinucleotide upstream of A/T
Number of substitutions
TCA 1
GCT 1
ACG 1
GTT 1
GCC 1
AAT 1
AGG 1
F
Frequency of upstream TTT trinucleotide
35% 30% 25% 20% 15% 10% 5% 0%
26.66%
3.58%
p < 0.00001
Mutated C/G sites n=30
Random C/G sites n=3826

Figure 3

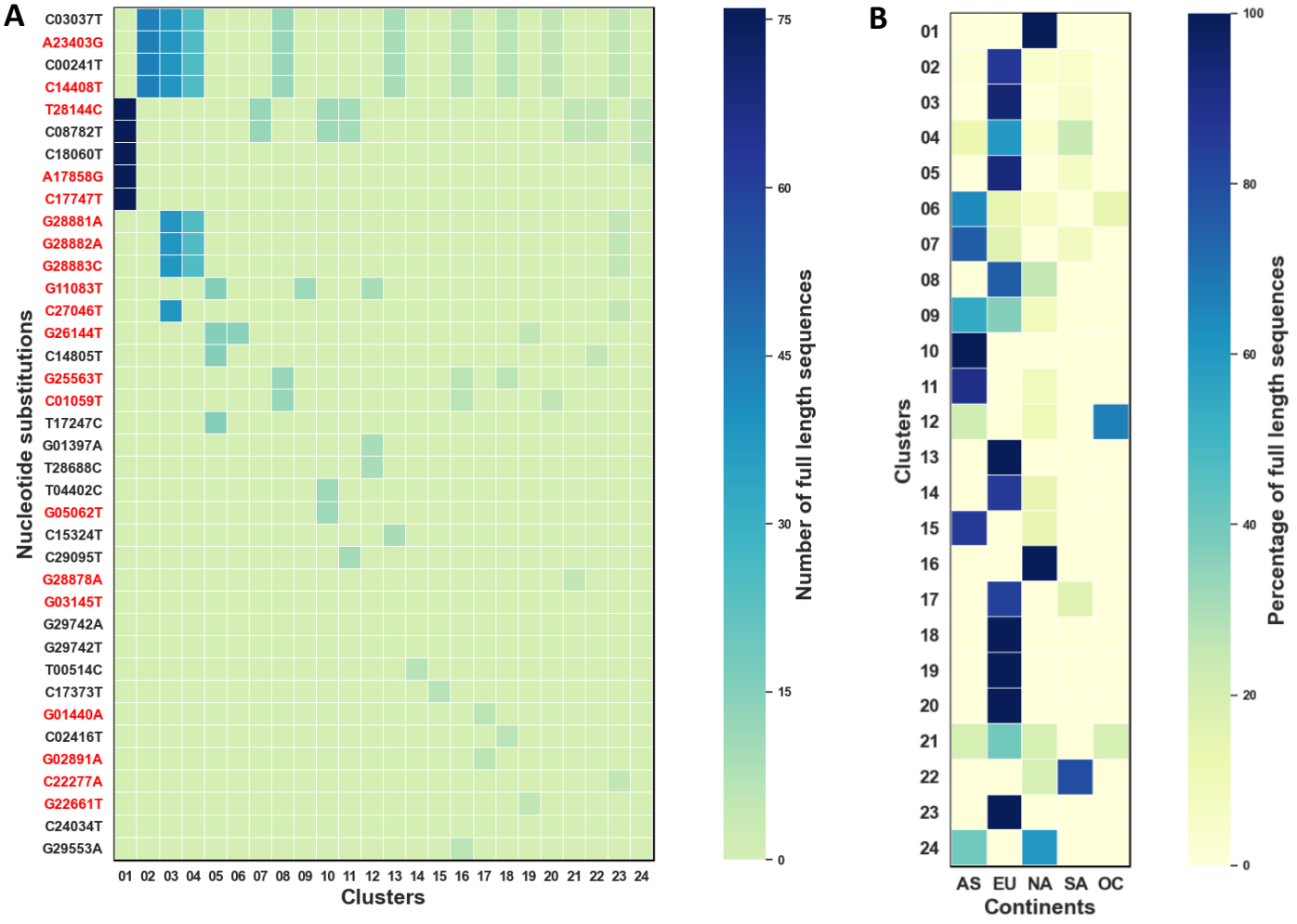

A
Nucleotide substitutions
C03037T
A23403G
C00241T
C14408T
T28144C
C08782T
C18060T
A17858G
C17747T
G28881A
G28882A
G28883C
G11083T
C27046T
G26144T
C14805T
G25563T
C01059T
T17247C
G01397A
T28688C
T04402C
G05062T
C15324T
C29095T
G28878A
G03145T
G29742A
G29742T
T00514C
C17373T
G01440A
C02416T
G02891A
C22277A
G22661T
C24034T
G29553A
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Clusters
Number of full length sequences
75
60
45
30
15
0
B
Clusters
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
AS EU NA SA OC
Continents
Percentage of full length sequences
100
80
60
40
20
0

Figure 4

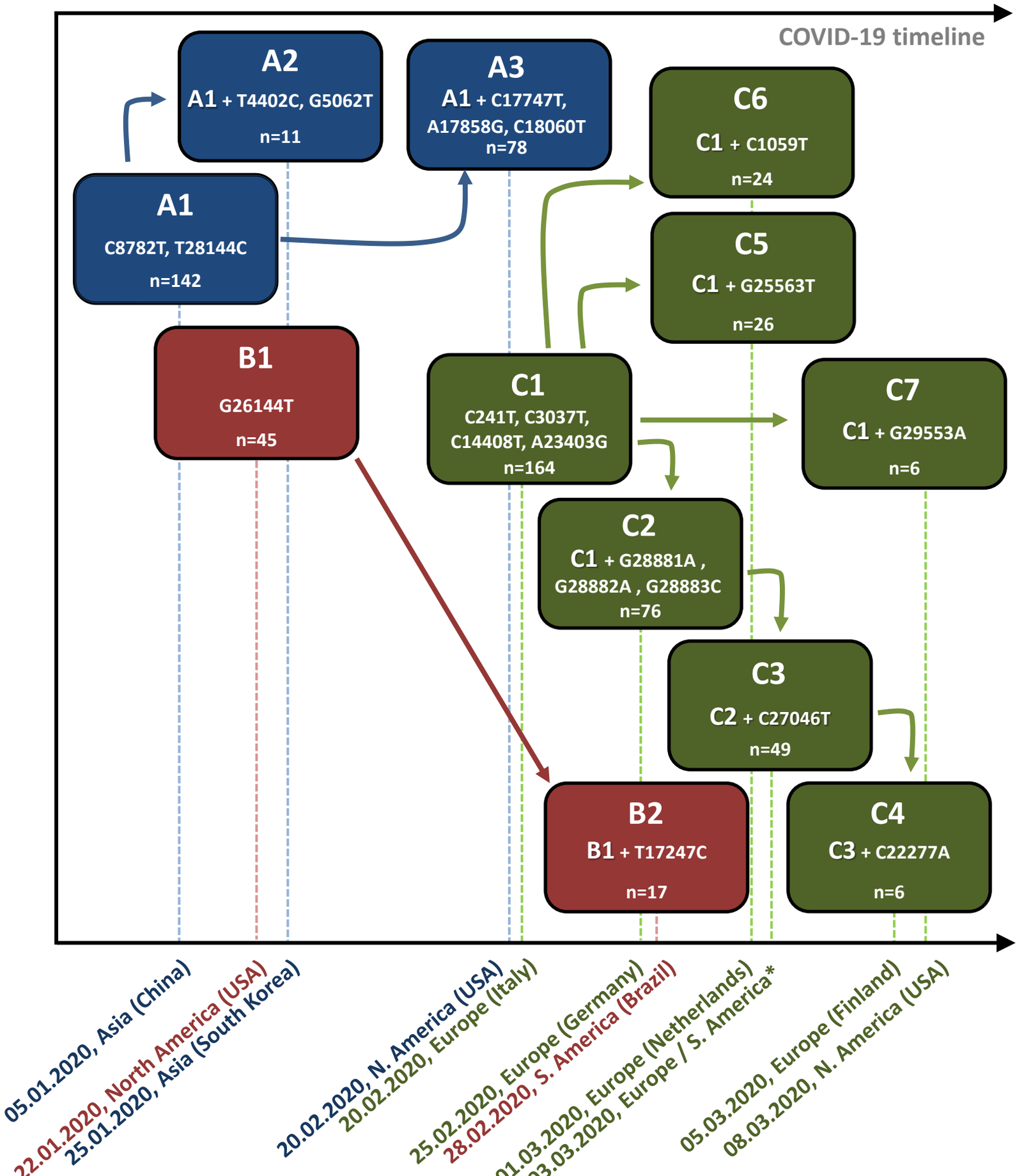
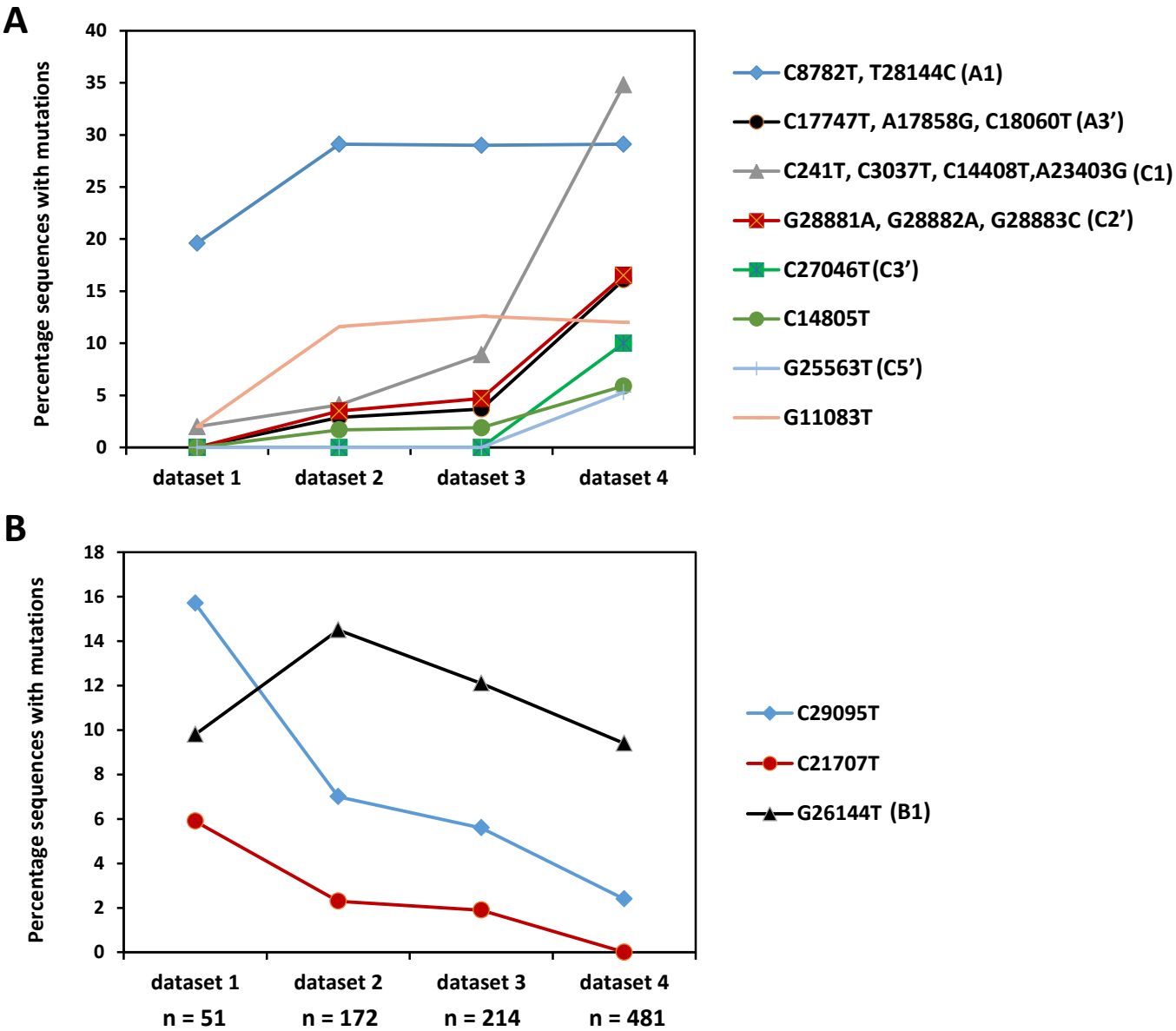

COVID-19 timeline
A2
A1 + T4402C, G5062T
n=11
A3
A1 + C17747T, A17858G, C18060T
n=78
C6
C1 + C1059T
n=24
A1
C8782T, T28144C
n=142
C5
C1 + G25563T
n=26
B1
G26144T
n=45
C1
C241T, C3037T, C14408T, A23403G
n=164
C7
C1 + G29553A
n=6
C2
C1 + G28881A , G28882A , G28883C
n=76
C3
C2 + C27046T
n=49
B2
B1 + T17247C
n=17
C4
C3 + C22277A
n=6
05.01.2020, Asia (China)
22.01.2020, North America (USA)
25.01.2020, Asia (South Korea)
20.02.2020, N. America (USA)
20.02.2020, Europe (Italy)
25.02.2020, Europe (Germany)
28.02.2020, S. America (Brazil)
01.03.2020, Europe (Netherlands)
03.03.2020, Europe / S. America*
05.03.2020, Europe (Finland)
08.03.2020, N. America (USA)

Figure 5


A
Percentage sequences with mutations
40
35
30
25
20
15
10
5
0
dataset 1
dataset 2
dataset 3
dataset 4
C8782T, T28144C (A1)
C17747T, A17858G, C18060T (A3')
C241T, C3037T, C14408T,A23403G (C1)
G28881A, G28882A, G28883C (C2')
C27046T (C3')
C14805T
G25563T (C5')
G11083T
B
Percentage sequences with mutations
18
16
14
12
10
8
6
4
2
0
dataset 1
n = 51
dataset 2
n = 172
dataset 3
n = 214
dataset 4
n = 481
C29095T
C21707T
G26144T (B1)