Geographic and Genomic Distribution of SARS-CoV-2 mutations

Daniele Mercatelli and Federico M. Giorgi*

Department of Pharmacy and Biotechnology, University of Bologna, Via Selmi 3, 40126, Bologna, Italy

* Corresponding author email: federico.giorgi@unibo.it

ORCID 0000-0003-3228-0580 (Daniele Mercatelli)

ORCID 0000-0002-7325-9908 (Federico M. Giorgi)

Abstract

The novel respiratory disease COVID-19 has reached the status of worldwide pandemic and large efforts are currently being undertaken in molecularly characterizing the virus causing it, SARS-CoV-2. The genomic variability of SARS-CoV-2 specimens scattered across the globe can underly geographically specific etiological effects. In the present study, we gather the 10,014 SARS-CoV-2 complete genomes currently available thanks to the collection endeavor of the GISAID consortium and thousands of contributing laboratories. We analyze and annotate all SARS-CoV-2 mutations compared with the reference Wuhan genome NC_045512.2. Our analysis shows the prevalence of single nucleotide transitions as the major mutational type across the world. There exist at least three clades characterized by geographic and genomic specificity. In particular, the clade G, prevalent in Europe, carries a D614G mutation in the Spike protein, which is responsible for the initial interaction of the virus with the host human cell. Our analysis may drive local modulation of antiviral strategies based on the molecular specificities of this novel virus.

Keywords

SARS-CoV-2; genomics coronavirus; COVID-19 evolution

Abbreviations

AA: aminoacid
COVID-19: Coronavirus Disease 2019
GISAID: Global Initiative on Sharing All Influenza Data
Indel: insertion/deletion event
NSP: non-structural protein
ORF: open reading frame
S: SARS-CoV-2 spike protein
SARS-CoV-2: Severe Acute Respiratory Syndrome, Coronavirus 2
SNP: single nucleotide polymorphism

Introduction

Initially reported in mid-December 2019 in the Chinese city of Wuhan, the newly emerged severe acute respiratory syndrome virus (SARS-CoV-2) is a single-stranded RNA beta-coronavirus with a very compact 29,903 nucleotides-long genome. This virus causes a serious disease known as Coronavirus Disease 2019 (COVID-19), which has spread in over 210 countries in less than four months, counting more than 2,5 million confirmed cases and almost 180,000 deaths reported worldwide as of April 22, 2020 (source: World Health Organization). A difference in case fatality rates across countries was observed, possibly due to a diverse demographic composition and the type of measures that have been taken in different countries to limit viral spreading [1]. According to data from the public database of the Global Initiative on Sharing All

Influenza Data (GISAID), three major clades of SARS-CoV-2 can be identified [2], that have been subsequently named as clade G (variant of the spike protein S-D614G), clade V (variant of the ORF3a coding protein NS3-G251), and clade S (variant ORF8-L84S). However, as more complete sequences become available, the need to define specific geographic distributions of virus variants becomes of practical importance to define clinical and political strategies at the local level. Despite several reports having confirmed a relatively low variability of SARS-CoV-2 genomes [3,4], it is still unclear if different fatality rates disease spreading speed in different countries may be the consequence of clade's differences in virulence, as discussed by a recent commentary comparing different strains in the USA [5]. It is therefore possible that more insights into the pathogenesis and virulence of this virus may come from comparative genomic analysis linked to epidemiologic data coming from different countries.

Genetic variance analyses must now play a crucial role in expanding knowledge on this new virus to adopt measures to contain its outbreak. Complete viral genome sequences have been made rapidly publicly available to the research community and have recently surpassed the 10,000 units, thanks to the worldwide effort of scientists and to the GISAID consortium. This data avalanche will result in an unprecedently rapid effort to analyze data to understand genome diversity [6,7], to hypothesize targetable targets for drug repositioning [8,9] and to develop prevention strategies [10]. In the present study, we performed the largest comparative study so far by analyzing more than 10,000 complete SARS-CoV-2 genomes. We will report all mutations and stratify them genomically and geographically, also highlighting insurgence of sub-clades and genomic highly variable spots. These finding may be extremely useful to design and think about the efficacy of measures that have been taken on a regional basis to limit SARS-CoV-2 spreading.

Methods

10,014 SARS-CoV-2 genomic sequences were downloaded from GISAID (Supplementary File 1) on April 20, 2020. Only viruses affecting human hosts were selected, excluding low coverage sequences and incomplete (<29,000 nucleotides) genomes.

The reference NC_045512.2 SARS-CoV-2 Wuhan genome [11], 29,903 nucleotides long, was obtained from NCBI GenBank. A GFF3 annotation associated to the refence, showing genomic coordinates for all protein sequences of SARS-CoV-2, is provided as Supplementary File 2. The large ORF1 polyprotein was split into its constituent Non-structural proteins (NSPs). The NSP12, encoding for the viral RNA-dependent RNA polymerase, was considered in the annotation as two regions, NSP12a and NSP12b, corresponding to the regions before and after a ribosomal frameshift, occurring as nucleotide 13468 is translated as both the last nucleotide of a codon and the first of the next codon.

Nucmer version 3.1 [12] was used to align all 10,014 genome sequences over the NC_045512.2 reference. The output of the alignment was converted to an annotated list of all mutational events using an internally developed R SARS-CoV-2 annotation algorithm provided as Supplementary File 3.

Results

Our analysis of 10,014 SARS-CoV-2 highlights a total of 67,364 mutation events compared to the NC_045512.2 Wuhan reference genome. Our results, event by event, are available as Supplementary File 4. While 130 samples, mostly originated from Asia, did not have any difference from the reference, 9,884 samples possessed at least one mutation. The number of mutations is relatively low, with a mode of 7 per sample and very few samples having more than 15 events (Figure 1 A). Overall, the least mutated samples (on average, less than 3 events/sample) originated from Asian samples (909 in total), while we observed a higher mutation rate (higher than 5 events/sample) in all other continents (Figure 1 B). Country-wise, the lowest divergence from the reference could be seen in Asian countries such as China, where the virus originated, Singapore and Japan (Figure 1 C).

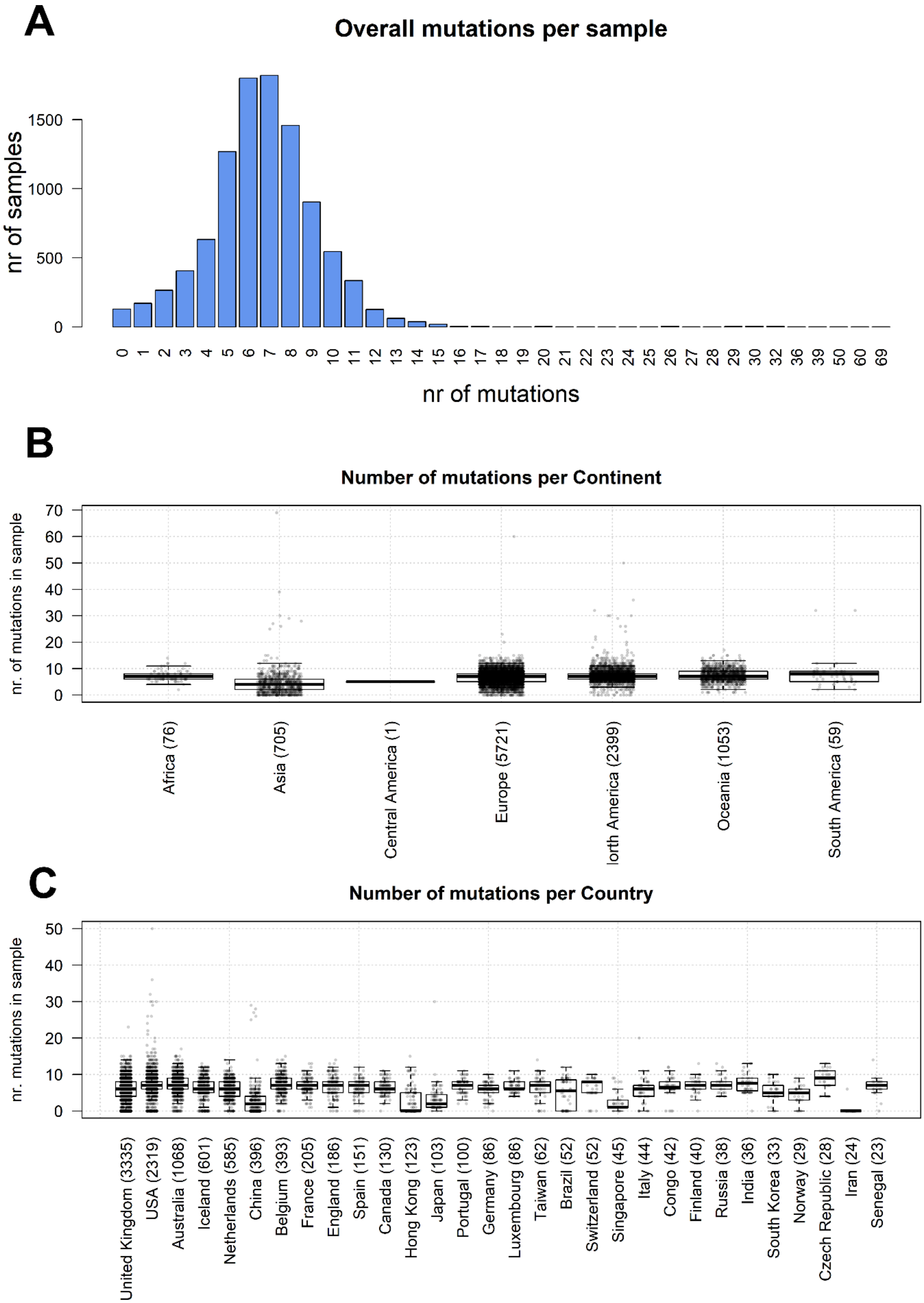

Figure 1. (A) distribution of number of mutational events for all SARS-CoV-2 genome samples analyzed. (B) Distributions of number of mutations for each sample, stratified per continent. The main boxplot rectangles are drawn between the 1st and 3rd quartile, with the median value indicated as a thick line. Boxplot whiskers fall on the closest point to the 1st/3rd quartile + 1.5 interquartile range as described in the R boxplot() function. The number in brackets after the continent name indicates the number of sequenced genomes. (C) As in B, with stratification performed country-wise, using the 30 countries with the highest number of sequenced genomes.

We analyzed the nature of each mutation, highlighting a prevalence of single-nucleotide polymorphisms (SNPs) over short insertion/deletion events (indels) (Figure 2 A and Supplementary File 5). Worldwide, we observed 39,036 aminoacid(aa)-changing SNP events, with less than half silent SNPs (19,629). Short in-frame deletions (3x deletions reducing the viral protein length without introducing stop codons) are the next largest class (542 total events), followed by 154 SNPs introducing a stop codon. We observe only 135 frameshift deletions, 35 frameshift insertions, 24 in-frame deletions (inserting multiples of 3 nucleotides) and 3 deletions introducing a stop codon in the mutated frame. Overall, 7806 mutations were located outside gene regions, prevalently in the untranslated regions (UTRs) of SARS-CoV-2 genome.

The distribution of these events is largely uniform across all continents (Figure 2 B).

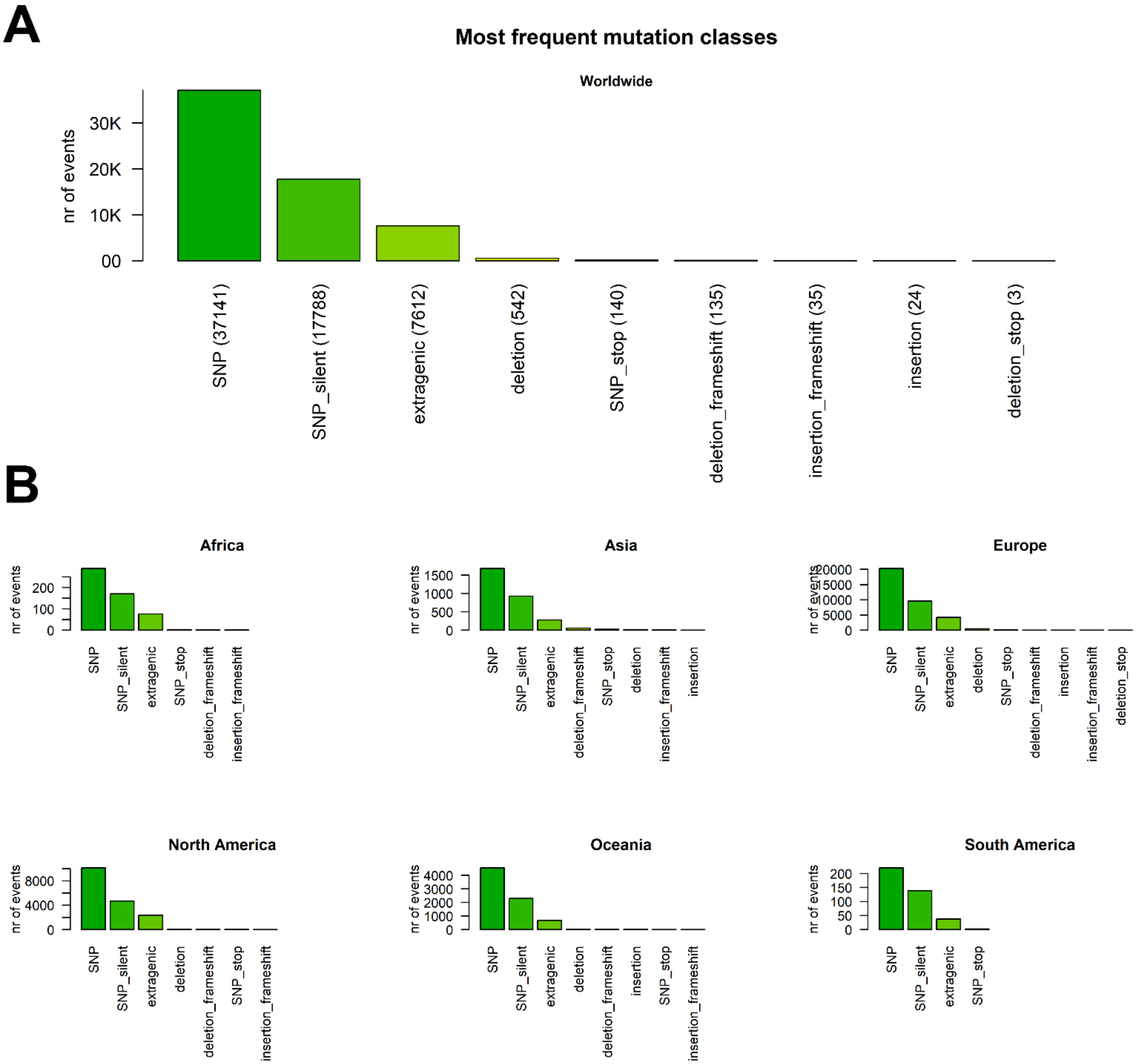

Figure 2. (A) Worldwide distribution of SARS-CoV-2 mutation classes. “SNP”, “deletion” and “insertion” terms without further specifications are intended as frameshift-preserving aa-changing events. (B) Distribution of mutation classes in continents.

We then classified the SARS-CoV-2 mutations according to their type, observing a prevalence of SNP transitions (purine->purine and pyrimidine->pyrimidine) over SNP transversions (purine->pyrimidine and vice versa). The most common event, both worldwide and continent-wise, is by far the C>T transition, accounting with roughly 50% of all observed worldwide viral mutations (Figure 3 A), followed by the A>G transition. The most common transversion, G>T, is the third most common event, with 8348 occurrences (Supplementary File 5). The most common indel, the deletion of the ATG codon, is the 10th most common event, with a total of 412 occurrences. The G>T transversion is however the second most common event overall in Asia and Australia, while the G>A transversion is the second most common event in South America (Figure 3B).

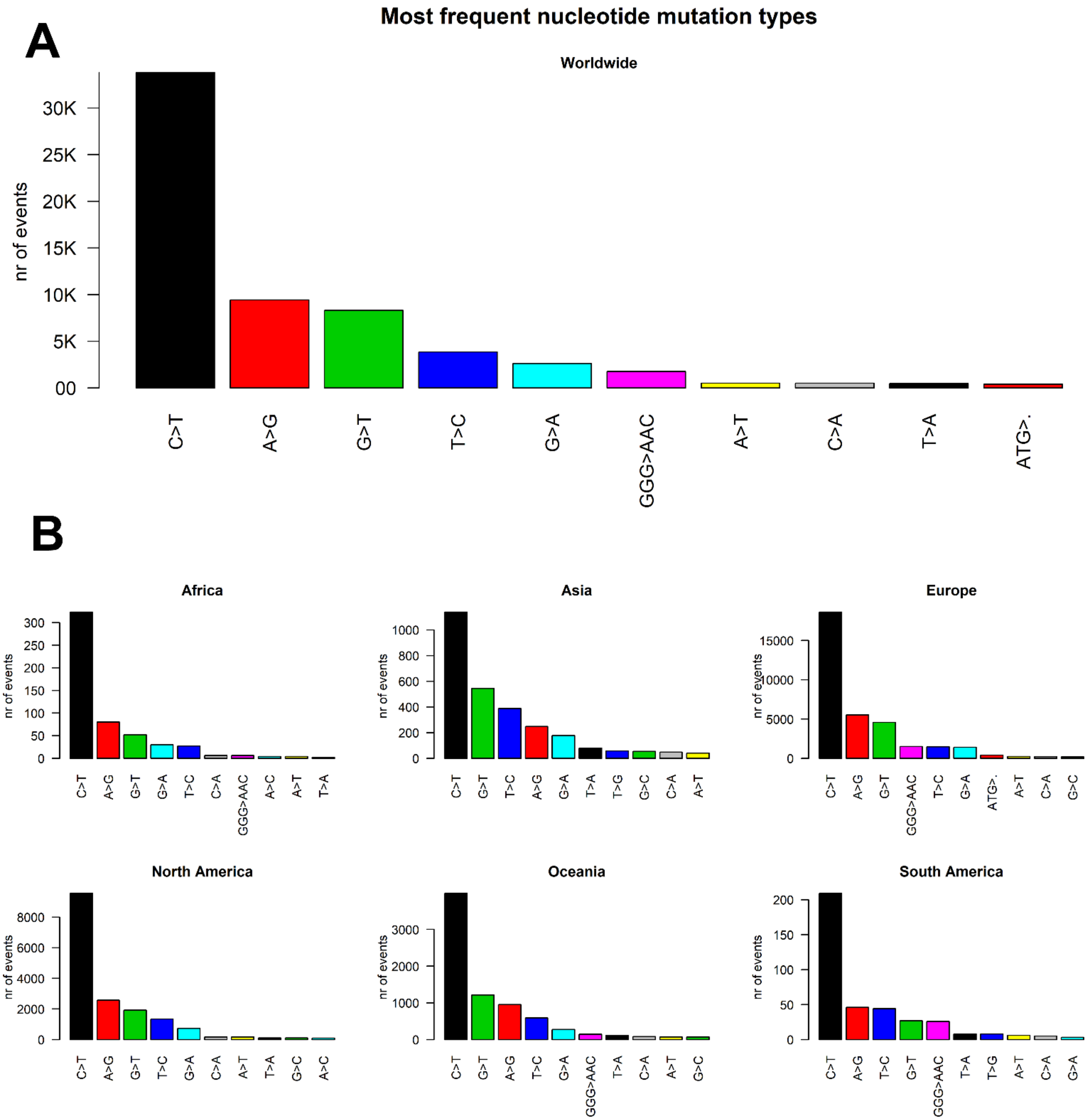

Figure 3. (A) Worldwide and (B) continent-stratified distribution of SARS-CoV-2 mutation types. Colors are assigned randomly but preserved across panels to facilitate tracking of identical types across continents.

We went into higher detail and analyzed the effects of each mutation on the protein sequences of SARS-CoV-2. Our results in this case start to address major differences across continents. The most prevalent mutation in sequenced genomes worldwide is a transversion affecting the 23,403rd nucleotide adenosine (Figure 4 A and Supplementary File 5), transformed into a guanosine (A23403G), defining the so-called G-clade of SARS-CoV-2 genomes, prevalent in Europe (where overall the highest sequencing effort has been undertaken, and therefore the highest number of samples), Oceania, South America and Africa (Figure 4 B). This mutation causes a D614G (aspartate to glycine in protein position 614) aa-change of the Spike (S) protein, which is responsible for the initial entry of the virus in the cell via the ACE2 human receptor [13]. The most frequent mutation in Asia is G11083T, causing a L37F aa-change in Non-structural protein 6 (Table 1).

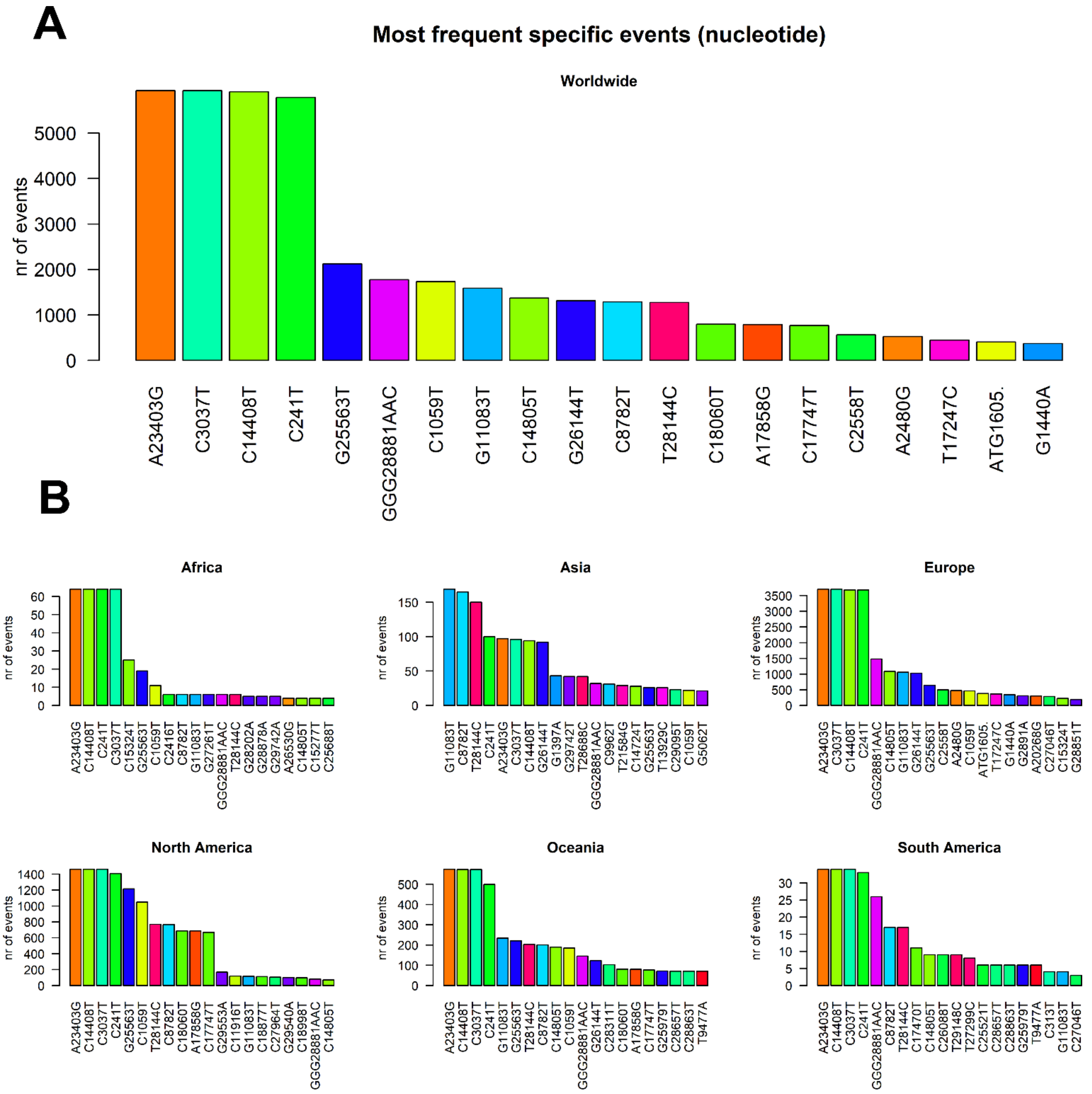

Figure 4. (A) Worldwide and (B) continent-stratified distribution of SARS-CoV-2 most frequent specific events, annotated as nucleotide coordinates over the reference genome NC_045512.2. Colors are assigned randomly but preserved across panels to facilitate tracking of identical types across continents.

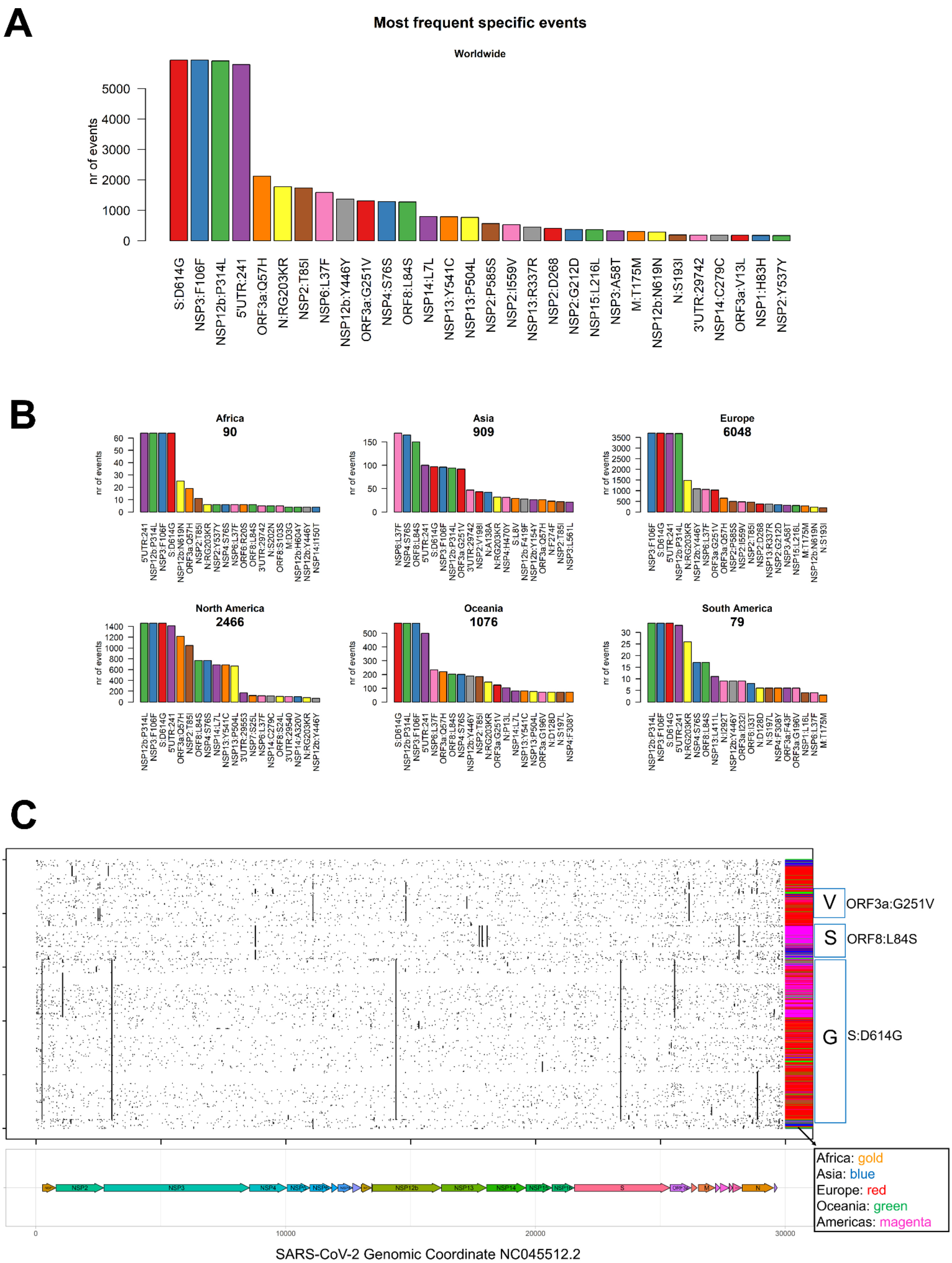

Figure 5. (A) Worldwide and (B) continent-stratified distribution of SARS-CoV-2 most frequent specific events, annotated protein changes using the format protein:mutation. Colors are assigned randomly but preserved across panels to facilitate tracking of identical types across continents. (C) Dot mat showing as X-axis the 29,903 nucleotide positions (sorted from left, 5' to right, 3') of SARS-CoV-2, and as Y axis the 10,014 genomes analyzed in this study. The

genomic sequences were clustered using the "complete" clustering algorithm. Coding sequence regions are shown at the bottom of panel B. To the right of the plot, we assigned a color to each sample according to the continent of origin. Further right, we manually annotated the groups according to the known GISAID clades (V, S and G) and the mutations that named them. The ORF8, where the S-clade T28144C/ORF8:L84S mutation is located, is situated directly to the left of the N gene.

We summarized the 20 most frequent events in Table 1. Apart from the already cited S:D614G mutation in the Spike protein, we observed a silent mutation (F106F) in NSP6, a proline-to-leucine in NSP12b (P314L) and a mutation in the 5'UTR at genomic coordinates 241 (Figure 5 A). All these four mutations are prevalent in Europe, Africa, Americas and Oceania (Figure 5 B), and almost always co-occurring in the sequenced genomes (Figure 5 C), determining the strongest signature for the G clade of SARS-CoV-2 genomes. The 5'UTR:241 mutations seems slightly less frequent than the other three G-clade events, but this could be due to the lower coverage of genomic tails in sequencing experiments.

The second most represented clade, V, is characterized and named by a mutation in ORF3a, G251V (Table 1 and Figure 5). We detected that this mutation is often (>95% of the times) co-occurring with the silent mutation at genomic location C14805T and by the more frequent V-clade associated NSP6-L37F. Clade V seems prevalent in a subset of European sequences (Figure 5). We identified a V-subclade, characterized by the three prevalent V variants, plus two extra aa-changing SNPs, both affecting NSP2: P585S and I559V (Table 1).

Finally, we observed a clade most represented in genomes sequenced in Asia and Americas, dubbed "S" by the GISAID consortium due to a L84S mutation in the ORF8, already identified in early February, 2020 [4]. The L84S mutation is often co-occurring with a silent mutation in the NSP4 gene, C8782T, and by other less frequent events in NSP13 and NSP14 (Table 1).

Our analysis also highlighted a curious event in the Nucleocapsid gene (N) of some genomes characterized by the S:D614G mutation and therefore belonging to the G clade. A succession of three nucleotides with genomic coordinates 28881, 28882 and 28883 is affected by almost-always (99.7%) co-occurring events, transforming the triplet GGG into AAC and causing two aminoacids of the N protein to change (RG203KR). This trinucleotide Nucleocapsid event defines a distinct G-subclade of the SARS-CoV-2 population. Two mutations on the NSP2, G212D and the deletion of Aspartate (D) 268 (Table 1), are not associated specifically to any of the three main clades, but instead we observe amongst the samples more similar to the original reference (Figure 5C, top group). In particular, the D268 deletion is observed in a minor part of S clade samples.

Finally, we include a distribution of the most common mutations (annotated as protein events) in the 25 countries with the highest number of sequenced genomes (Supplementary File 6). We noticed that all European countries share very similar profiles with each other. In Asia, we could however observe a marked difference between Japan (where the NSP6:L37F mutation prevails), China (where the ORF8:L84S and NSP4:S76S events are the most common) and Hong Kong (marked by ORF3a:G251V events).

Discussion

Our analysis confirms a low mutation rate of the virus, with an average of 6.7 mutations per sample with respect to the reference SARS-CoV-2 genome sequences. However, the existing mutations allow to group the samples into three distinct clades, G, S and V, characterized by a collection of specific mutations. The clades can be further characterized by most recent mutations and will likely be split even further in the future.

While the aa-changing SNPs are the most prevalent mutational events, we observed also silent SNPs and extragenic (mostly 5'UTR) SNPs. The silent events may not determine an immediate effect on the protein sequences, but they have repercussions as they may change the codon usage and translation efficiency. In the case of the 5’UTR SNPs, mutations may affect the transcription and replication rates of the virus, or the folding of the genomic ssRNA, processes that have been only recently and only partially elucidated [14].

The early studies currently performed on SARS-CoV-2 transcriptome dynamics may also suggest mechanisms for mutation onset, which our study shows being prevalently single-nucleotide transitions. This phenomenon can be associated to defective efficiency of the viral RNA-depedent RNA polymerase or, as recently suggested, by mechanisms of RNA editing triggered by the host cell as a defense mechanism [15]. Whatever the origin, SARS-CoV-2 tends to retain its genomic integrity across propagation, with almost no reported large indels across sequenced genomes (the largest reported being a unique 80-nucleotide deletion in ORF7a, in Arizona sample EPI_ISL_424669 – Supplementary Files 1 and 4).

Table 1. The 20 most frequent mutation events observed in sequenced SARS-CoV-2 genomes.

Genomic Coordinate	Effect on protein/UTR	Nr of samples	Class	Genomic Region	Clade
A23403G	S:D614G	5935	AA-changing SNP	Spike protein	G
C3037T	NSP3:F106F	5933	silent SNP	Non-Structural protein 3 (predicted phosphoesterase)	G
C14408T	NSP12b:P314L	5908	AA-changing SNP	Non-Structural protein 12, post-ribosomal frameshift (RNA-dependent RNA polymerase)	G
C241T	5'UTR:241	5787	5'UTR SNP	5' UnTranslated Region	G
G25563T	ORF3a:Q57H	2126	AA-changing SNP	ORF3a protein	G
GGG28881AAC	N:RG203KR	1773	AA-changing SNP triplet	Nucleocapsid protein	G
C1059T	NSP2:T85I	1731	AA-changing SNP	Non-Structural protein 2	G
G11083T	NSP6:L37F	1585	AA-changing SNP	Non-Structural protein 6 (transmembrane protein)	V
C14805T	NSP12b:Y446Y	1372	silent SNP	Non-Structural protein 12, post-ribosomal frameshift (RNA-dependent RNA polymerase)	V
G26144T	ORF3a:G251V	1312	AA-changing SNP	ORF3a protein	V
C8782T	NSP4:S76S	1289	silent SNP	Non-Structural protein 4	S
T28144C	ORF8:L84S	1276	AA-changing SNP	ORF8 protein	S
C18060T	NSP14:L7L	796	silent SNP	Non-Structural protein 14 (3'-to-5' exonuclease)	S
A17858G	NSP13:Y541C	786	AA-changing SNP	Non-Structural protein 13	S
C17747T	NSP13:P504L	765	AA-changing SNP	Non-Structural protein 13	S
C2558T	NSP2:P585S	564	AA-changing SNP	Non-Structural protein 2	V
A2480G	NSP2:I559V	524	AA-changing SNP	Non-Structural protein 2	V
T17247C	NSP13:R337R	447	silent SNP	Non-Structural protein 13	S
ATG1605del	NSP2:D268	406	deletion	Non-Structural protein 2	
G1440A	NSP2:G212D	370	AA-changing SNP	Non-Structural protein 2	

Further studies combining genomic details with epidemiological information and clinical features of COVID-19 patients may be extremely useful to identify strategies and therapies that can help to reduce the burden of this disease. One important effect of mapping mutations is the development of antiviral therapies targeting specific regions. For example, the development of protein-based and RNA-based vaccines based on the SARS-CoV-2 Spike region [16] will have to take into account the observed diversity of the Spike protein, affected by a mutation D614G in the G clade, which are the most common viruses observed in European samples.

Supplementary Material Legends

Supplementary File 1: GISAID acknowledgment table reporting the geographic origin and contributions of all genomes analyzed in this study.

Supplementary File 2: annotation of NC_045512.2 SARS-CoV-2 Wuhan genome sequence (GFF3 format).

Supplementary File 3: bash/R scripts used to generate and annotate genome variants.

Supplementary File 4: full annotation of all mutations identified by this study. Columns are described here. Sample: GISAID sample id; refpos: position in the NC_045512.2 reference genome; refvar: nucleotide composition of the reference at refpos coordinate (a "." Indicates an insertion); qvar: variant in the query sample (a "." indicates a deletion); qlength: length of the query genome (reference genome is always 29,903 nucleotides long); region: region annotated in the event position (coding sequence, intergenic or UTR); variant: either a protein change (shown as aminoacid code) or the genomic position (if the event affects a noncoding region); varclass: variant class (as in Figure 2); annotation: full name of the protein coded by the affected region (if coding).

Supplementary File 5: table indicating the number of events associated to all mutation classes and most frequent mutation types and specific events.

Supplementary File 6: country-stratified distribution of SARS-CoV-2 most frequent specific events, annotated protein changes using the format protein:mutation.

Author Contributions

F.M.G. designed the study. F.M.G. and D.M. performed research, analyzed data and wrote the manuscript.

Acknowledgments

The authors wish to thank Michele Morgante, Davide Scaglione and Luca Triboli for the fruitful discussions.

Conflict of Interest

The authors declare no conflict of interest.

Funding

This work was supported by the Italian Ministry of University and Research, Montalcini Grant 2016.

References

1. Dowd, J.B.; Andriano, L.; Brazel, D.M.; Rotondi, V.; Block, P.; Ding, X.; Liu, Y.; Mills, M.C. Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proc. Natl. Acad. Sci.* **2020**, 202004911, doi:10.1073/pnas.2004911117.
2. Forster, P.; Forster, L.; Renfrew, C.; Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci.* **2020**, 202004999, doi:10.1073/pnas.2004999117.
3. Lu, R.; Zhao, X.; Li, J.; Niu, P.; Yang, B.; Wu, H.; Wang, W.; Song, H.; Huang, B.; Zhu, N.; et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **2020**, S0140673620302518, doi:10.1016/S0140-6736(20)30251-8.
4. Ceraolo, C.; Giorgi, F.M. Genomic variance of the 2019-nCoV coronavirus. *J. Med. Virol.* **2020**, *92*, 522–528, doi:10.1002/jmv.25700.

5. Brufsky, A. Distinct Viral Clades of SARS-CoV-2: Implications for Modeling of Viral Spread. *J. Med. Virol.* **2020**, jmv.25902, doi:10.1002/jmv.25902.
6. Andersen, K.G.; Rambaut, A.; Lipkin, W.I.; Holmes, E.C.; Garry, R.F. The proximal origin of SARS-CoV-2. *Nat. Med.* **2020**, *26*, 450–452, doi:10.1038/s41591-020-0820-9.
7. Shen, Z.; Xiao, Y.; Kang, L.; Ma, W.; Shi, L.; Zhang, L.; Zhou, Z.; Yang, J.; Zhong, J.; Yang, D.; et al. Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. *Clin. Infect. Dis.* **2020**, ciaa203, doi:10.1093/cid/ciaa203.
8. Wu, C.; Liu, Y.; Yang, Y.; Zhang, P.; Zhong, W.; Wang, Y.; Wang, Q.; Xu, Y.; Li, M.; Li, X.; et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm. Sin. B* **2020**, S2211383520302999, doi:10.1016/j.apsb.2020.02.008.
9. Zhou, Y.; Hou, Y.; Shen, J.; Huang, Y.; Martin, W.; Cheng, F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov.* **2020**, *6*, 14, doi:10.1038/s41421-020-0153-3.
10. Zhao, S.; Chen, H. Modeling the epidemic dynamics and control of COVID-19 outbreak in China. *Quant. Biol.* **2020**, *8*, 11–19, doi:10.1007/s40484-020-0199-0.
11. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **2020**, *5*, 536–544, doi:10.1038/s41564-020-0695-z.
12. Delcher, A.L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **2002**, *30*, 2478–2483, doi:10.1093/nar/30.11.2478.
13. Guzzi, P.H.; Mercatelli, D.; Ceraolo, C.; Giorgi, F.M. Master Regulator Analysis of the SARS-CoV-2/Human Interactome. *J. Clin. Med.* **2020**, *9*, 982, doi:10.3390/jcm9040982.
14. Kim, D.; Joo-Yeon, L.; Jeong-Sun, Y.; Jun Won, K.; Narry, K.; Hyeshik, C. The architecture of SARS-CoV-2 transcriptome. *Cell In press*, doi:https://doi.org/10.1016/j.cell.2020.04.011.
15. Milewska, A.; Kindler, E.; Vkovski, P.; Zeglen, S.; Ochman, M.; Thiel, V.; Rajfur, Z.; Pyrc, K. APOBEC3-mediated restriction of RNA virus replication. *Sci. Rep.* **2018**, *8*, 5960, doi:10.1038/s41598-018-24448-2.
16. Callaway, E. Coronavirus vaccines: five key questions as trials begin. *Nature* **2020**, *579*, 481–481, doi:10.1038/d41586-020-00798-8.