Research Square Preprint: Please note that this article has not completed peer review.

Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant

CURRENT STATUS: UNDER REVIEW

Journal of Translational Medicine BMC

Maria Pachetti
Elettra Sincrotrone Trieste SCpA S.S. 14 km 163.5 34149 Basovizza (TS) Italy
ORCiD: https://orcid.org/0000-0002-1307-7475

Bruna Marini
Ulisse Biomed S.S. 14 km 163.5 34149 Basovizza (TS) Italy

Francesca Benedetti
Institute of Human Virology, Department of Biochemistry and Molecular Biology, School of Medicine, University of Maryland (Baltimore, USA)

Fabiola Giudici
University of Trieste, Department of Medicine, Surgery and Health Science, Trieste (Italy)

Elisabetta Mauro
Ulisse BioMed S.S. 14 km 163.5 34149 Basovizza (TS) Italy

Paola Storici
Elettra Sincrotrone Trieste S.S. 14 km 163.5 34149 Basovizza (TS) Italy

Claudio Masciovecchio
Elettra Sincrotrone Trieste S.S. 14 km 163.5 34149 Basovizza (TS) Italy

Silvia Angeletti
Medical Statistic and Molecular Epidemiology Unit, University of Biomedical Campus, Rome (Italy)

Massimo Ciccozzi
Medical Statistic and Molecular Epidemiology Unit, University of Biomedical Campus, Rome (Italy)

Robert Charles Gallo
Institute of Human Virology, Co-founder and International Science Advisor of Global Virus Network, Department of Medicine, School of Medicine, University of Maryland (Baltimore, USA

Davide Zella
Institute of Human Virology and Member of the Global Virus Network, Department of Biochemistry

and Molecular Biology, School of Medicine, University of Maryland (Baltimore, USA)

Rudy Ippodrino
Ulisse Biomed S.S. 14 km 163.5 34149 Basovizza (TS) Italy

r.ippodrino@ulissebiomed.com*Corresponding Author*

DOI:

10.21203/rs.3.rs-20304/v1

SUBJECT AREAS

Translational Medicine

KEYWORDS

SARS-CoV-2, COVID-19, mutation, Drug Resistance, RdRp, Europe, RNA-dependent-RNA -polymerase, pneumonia, viral mutagenesi

Abstract

Background. SARS-CoV-2 is a RNA coronavirus responsible for the pandemic of the Severe Acute Respiratory Syndrome (COVID-19). RNA viruses are characterized by a high mutation rate, up to a million times higher than that of their hosts. Virus mutagenic capability depends upon several factors, including the fidelity of viral enzymes that replicate nucleic acids, as SARS-CoV-2 RNA dependent RNA Polymerase (RdRp). Mutation rate drives viral evolution and genome variability, thereby enabling viruses to escape host immunity and to develop drug resistance.

Methods. We analyzed 220 genomic sequences from the GISAID database derived from patients infected by SARS-CoV-2 worldwide from December 2019 to mid-March 2020. SARS-CoV-2 reference genome was obtained from the GenBank database. Genomes alignment was performed using *Clustal* Omega. Mann-Whitney and Fisher-Exact tests were used to assess statistical significance.

Results. We characterized 8 novel recurrent mutations of SARS-CoV-2, located at positions 1397, 2891, 14408, 17746, 17857, 18060, 23403 and 28881. Mutations in 2891, 3036, 14408, 23403 and 28881 positions are predominantly observed in Europe, whereas those located at positions 17746, 17857 and 18060 are exclusively present in North America. We noticed that the 14408 mutation, emerged for the first time in Europe in mid-February 2020, is present in the SARS-CoV-2 RdRp gene sequence. Viruses with RdRp mutation have a median of 3 point mutations [range: 2-5], otherwise they have a median of 1 mutation [range: 0-3] (p value < 0.001).

Conclusions. These findings suggest that the virus is evolving and European, North American and Asian strains might coexist, each of them characterized by a different mutation pattern. The contribution of the mutated RdRp to this phenomenon needs to be investigated. To date, several drugs targeting RdRp enzymes are being employed for SARS-CoV-2 infection treatment. Some of them have a predicted binding moiety in a SARS-CoV-2 RdRp hydrophobic cleft, which is adjacent to the 14408 mutation we identified. Consequently, it is important to study and characterize SARS-CoV-2 RdRp mutation in order to assess possible drug-resistance viral phenotypes. It is also important to recognize whether the presence of some mutations might correlate with different SARS-CoV-2 mortality rates.

Background

The recent emergence of the novel, human pathogen Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in China and its rapid national and international spread poses a global health emergency. On March 11th 2020, WHO publicly declared the SARS-CoV-2 outbreak as a pandemic. In a few weeks, the virus caused thousands of deaths worldwide, strongly impacting the global economy and human habits. SARS-CoV-2 is an enveloped, +ssRNA virus, belonging to the Betacoronavirus genus which includes two other RNA viruses that have caused recent important epidemics: Severe Acute Respiratory Syndrome (SARS) caused by SARS-CoV, and the Middle East Respiratory Syndrome (MERS) by MERS-CoV.

Noteworthy, some evidence has been recently provided, supporting that SARS-CoV-2 mortality can significantly differ depending on the geographic area. For example, Baud and colleagues reported that mortality rate is three times higher out of China (15.2% [95% CI 12.5–17.9] out of China, compared to 5.6% [95% CI 5.4–5.8] in China)[1]. This rate has been re-estimated by dividing the number of deaths on a given day by the number of patients with confirmed SARS-CoV-2 infection 14 days before, considering the WHO data relative to the cumulative number of deaths to March 1st, 2020[1]. Differences in viral infection rates can be due to a combination of factors, including different national strategies adopted for people movement restrictions, isolation and quarantine, different genetic population herd immunity. Mortality differences are to understand, but viral mutations and evolution capability over time may be important.

RNA viruses mutation rate is dramatically high, up to a million times higher than their hosts, and this high rate is correlated with virulence modulation and evolvability, traits considered beneficial for viral adaptation[2]. Wang and coworkers have recently characterized 13 variation sites in SARS-CoV-2 ORF1ab, S, ORF3a, ORF8 and N regions, among which positions 28144 in ORF8 and 8782 in ORF1a showed a mutation rate of 30.53% and 29.47%, respectively[3]. Prior reported results show that SARS-CoV-2 is rapidly moving across countries and genomes with newly mutation hotspots are emerging. RNA virus mutation rate contributes to viral adaptation creating a balance between the integrity of

genetic information and genome variability[4-6]. Biological characterization of viral mutations can provide precious insights for assessing viral drug resistance, immune escape and pathogenesis related mechanisms. Additionally, viral mutation studies can be crucial for designing new vaccines, antiviral drugs and diagnostics assays. The viral genome mutagenic process depends on the viral enzymes that replicate the nucleic acids, influenced by few or no proofreading capability and/or post-replicative nucleic acid repair. Other mutation-generating processes include: host enzymes, spontaneous nucleic acid damages due to physical and chemical mutagens, recombination events and also particular genetic elements responsible for production of new variants. Mutation rates are modulated by other factors such as determinants of the template sequence and structure involved in viral replication.

RNA-dependent RNA polymerases (RdRps) are multi-domain proteins able to catalyze RNA-template dependent formation of phosphodiester bonds between ribonucleotides in the presence of divalent metal ion[7-9]. In most viruses, RNA polymerase lacks proofreading capability, with some exceptions such as Nidovirales order (to which the Coronavirus genus belongs), that stands out for having the largest RNA genomes. Nidoviruses are characterized by a complex machinery dedicated to RNA synthesis, that is operated by non-structural proteins (nsps), being produced as cleavage products of the ORF1a and ORF1b viral polyproteins[10] to facilitate virus replication and transcription.

The SARS-CoV-2 RdRp (also named nsp12) is a key component of the replication/transcription machinery. SARS-CoV-2 shares a high homology for nsp12 compared to SARS-CoV, suggesting that its function and mechanism of action might be well conserved[11]. This has been confirmed by a recent cryo-EM structural study obtained for SARS-CoV-2 nsp12[12]. In SARS-CoV, an exonuclease activity with proofreading function has been reported for the nsp14 (ExoN), and a homologue nsp14 protein is found in the SARS-CoV-2 as well[13]. ExoN increases the fidelity of RNA synthesis by correcting nucleotide incorporation errors made by RdRp[14]. Genetic inactivation of the coronavirus ExoN results in a 21-fold decrease in replication fidelity compared to wild type SARS-CoV[15]. Moreover, Kirchdoerfer

and colleagues showed the involvement of nsp7 and nsp8 in the formation of a supercomplex with RdRp in SARS-CoV[16], and this has been confirmed also for SARS-CoV-2 in a recent study unveiling the structure of SARS-CoV-2 RdRp/nsp7/nsp8 complex[12]. This complex ensures RdRp processivity, becoming fundamental in the transcription fidelity. Nevertheless, the critical SARS-CoV RdRp residues involved in ExoN, nsp7 and nsp8 interaction have still to be identified.

RdRps are considered among primary targets for antiviral drug development, against a wide variety of viruses. Some RdRp inhibitors have been considered to target SARS-CoV-2: Favipiravir[17], Galidesivir[18], Remdesivir[19] and Ribavirin[20]. Interestingly, the docking site is not located in proximity to the catalytic domain of the RdRp[21]. In addition, other possible drugs such as Filibuvir, Cepharanthine, Simeprevir and Tegobuvir, are predicted to be potential inhibitors of RdRp[22]. Naturally occurring mutations in critical residues for drug efficacy can lead to drug resistance phenomena, with a significant loss in the binding affinity of these molecules to the RdRp.

We focused our study on SARS-CoV-2 mutations in order to assess if new viral variants were spreading across the Countries. This characterization of SARS-CoV-2 variants could lead to better therapeutics treatments, vaccines designs and diagnostics approaches.

Methods

SARS-CoV-2 virus reference sequence used for the analysis was deposited in January 2020 by Wu and coworkers[11] formerly called “Wuhan seafood market pneumonia virus” (WSM, NC_045512) (https://www.ncbi.nlm.nih.gov/nuccore/NC_045512). GISAID database (https://www.gisaid.org/) filtered from December 2019 up to March 13th, 2020 was used to collect 220 SARS-CoV-2 complete genomes of different patients all around the world (i.e. China, USA, Canada, Australia, United Kingdom, Germany, France, Japan, Italy, Switzerland, Singapore, Luxembourg, Netherlands, Spain, Portugal, Sweden, Czech Republic, Thailand, India, Cambodia, Hong Kong, Finland, Singapore, and Ireland) taking into particular consideration those deposited during the development of European outbreaks. Only complete genomes (28000–30000 bps) were analyzed.

Clustal Omega, Serial Cloner and Blast tools were used to conduct multiple sequence alignment,

comparing WSM sequence to sequences isolated from patients, whereas Swiss Model and Ez-mol were used for protein modeling.

The statistical analysis was performed by R software. We first checked the normality of data distribution with the Shapiro-Wilk test, expressing the continuous variables as median and range (min-max). Categorical variables were expressed as absolute frequency and percentages. Nonparametric Mann-Whitney and Fisher-Exact tests were used to compare the number of mutations per genome with at least one of the selected mutations with respect to the group of genomes that do not present the specific mutation analyzed. All p-values were calculated from 2-sided tests using 0.05 as the significance level.

Results

Identification of recurrent mutation hotspots in different geographic areas

A database of 220 complete SARS-CoV-2 patient-isolated genome sequences randomly collected from the GISAID database were aligned and compared to the WSM SARS-CoV-2 reference genome. In particular, 5 patient-isolated genomes were submitted to the GISAID database in December 2019 (2.3%), 67 in January 2020 (30.45%), 67 in February 2020 (30.45%) and 81 (36.8%) up to the 13th of March 2020. About 33.6% of complete genomes belong to patients aged less than 44 years old, which is the average age of the patients included in the database. The majority of patients are men (55.5%). We divided our dataset into 4 geographic areas: Asia, Oceania, Europe, North America (Fig. 1). Within each area we performed alignment analysis comparing patients' genomes with the reference sequence. The Asian group comprises genomes obtained from patients located in China, Japan, South-East-Asia and India. The Oceanian group comprises genomes from Australian patients, whereas the European one includes every genome obtained from patients located in each one of the European states (Spain, Portugal, United Kingdom, Netherlands, Italy, Germany, Switzerland, France, Luxemburg, Sweden, Finland, Denmark and Belgium). Finally, the North America group contains genomes from US and Canadian patients.

We evaluated the distribution of SARS-CoV-2 mutations through different geographic areas (see Fig. 1), calculating the mutation frequency within these 4 geographic areas, by normalizing the

number of genomes carrying a given mutation per geographic area.

We confirmed the occurrence of mutations located at positions 3036, 8782, 11082, 28144 and 26143[3,23–25]. Moreover, we highlighted the presence of additional “conserved mutations” in all the geographic areas, taking into account only those occurring more than 10 times in our database. Those with a lower occurrence were not reported. These mutations were found in position 1397, 2891, 14408, 17746, 17857, 18060, 23403, 28881, belonging to ORF1ab (1397 nsp2, 2891 nsp3, 14408 RdRp, 17746 and 17857 nsp13, 18060 nsp14), S (23403, spike protein) and ORF9a (28881, nucleocapsid phosphoprotein) sequences, respectively.

We found that 3 out of the 12 most frequent mutations (positions 3036, 8782 and 18060) were silent, whereas one mutation (position 11083) was outside the ORF sequence. On the other hand, mutations 1397, 2891, 14408, 17746, 17857, 23403, 26143, 28144 and 28881 resulted in amino acid changes as follows: 1397 (V to I), 14408 (P to L), 17746 (P to L), 17857 (C to Y), 23403 (D to G), 26143 (G to V), 28144 (L to S). Mutation located at position 28881 is related to a double codon mutation, inducing the substitution of two amino acids, namely 28881 (R to K) and (G to R). The new amino acid present in 1397 (V to I), 14408 (P to L), 17746 (P to L), 17857 (C to Y), 26143 (G to V) and 28144 (L to S) had a similar isoelectric point compared to the original amino acid present in the reference protein sequences, with the exception of the mutations at positions 23403 (D to G), 28881 (R to K) and 28881 (G to R), where the mutated amino acid has a significantly different isoelectric point. Further studies are needed to determine whether these mutations have an impact on proteins’ function and structure. We noted that the number and the occurrence of each mutation increase in genomes found out of Asia, reaching a maximum in genomes found in Europe and North America. We also noted that the viral strains found in Europe and North America are derived from the L-“strain” originated in Asia[23].

Characterization of geographically distinct hotspots over time

In order to determine the appearance of each mutation, we analyzed each genome from each geographic area over time, by classifying them according to the timing of sample collection, as indicated in the GISAID database. According to this analysis, 6 time subgroups were defined, namely

December 2019 (genomes from 5 patients), 1st -15th Jan. 2020 (genomes from 15 patients), 16th -31st Jan. 2020 (genomes from 52 patients), 1st -15th Feb. 2020 (genomes from 13 patients), 16th -29th Feb- 2020 (genomes from 55 patients) and 1st -13th Mar. 2020 (genomes from 80 patients). The number of mutations (normalized by the population taken into account for each period of time) increases over time during viral spread out of Asia (see Fig. 2). No mutations were observed in the Asian genomes analyzed in December 2019. Interestingly, a different pattern of mutations was observed in Europe between January and February, when a new mutation, at position 14408, emerged (depicted in red). This mutation is located in the RdRp gene. Also starting from February 2020, the emergence of additional new mutations (i.e. 23403, 28881 and 2891 - black, electric blue, dark green, respectively) is observed. Over time, we also noted an increase in the frequency of mutation 3036 (orange), already present in mid-January (2.2%).

Moreover, a different pattern of hotspot mutations is clearly distinguishable in viral genomes detected in North American patients starting from March 2020, when the outbreak of positive cases was reported in the US and Canada. In this group, three novel mutations (17746, 17857 and 18060 - light blue, purple and light pink, respectively) were reported. Interestingly, viral genomes present in North American patients carrying RdRp mutation (14%) do not carry any of the European specific mutations.

Mutations hotspots pattern after February 9th, 2020

Given the importance of RdRp for viability and replication of RNA viruses, mutations in this gene are statistically less likely to occur. However, in some cases, such as in poliovirus, episodes of drug-resistance induced by a point mutation in RdRp have been reported[26]. In our database, the first appearance of RdRp mutation is manifested on February 9th, 2020 in the UK (England), when a dramatic increase of the number of European infected patients was reported from the WHO website[27]. We evaluated the increase/decrease of each mutation frequency before and after February 9th, 2020 across the different geographic areas (Fig. 3). In particular, we observed a strong increase (+ 60.5%) of genomes carrying the 14408 mutation (affecting RdRp) in Europe, together with an increase of genomes carrying the 3036 mutation (+ 61.7%), and the 28881 mutation (+ 29.6%) (see upper table Fig. 3).

Simultaneous occurrence of RdRp mutation with other mutations

Next, we analyzed genomes collected after February 9th 2020, when mutation in position 14408 (RdRp) was reported in the database for the first time. For the purpose of analysis, we divided the genomes into two groups: group 1 contains genomes with mutation in position 14408 (RdRp) ($n = 53$, 4 North America and 49 European), and group 2 without RdRp mutation ($n = 84$).

Genomes in group 1 showed an increased number of mutations compared to group 2. In particular, group 1 shows 6 genomes with two mutations (11.3%), 25 genomes with three mutations (47.2%), 21 genomes with four mutations (39.6%), and 1 genome with 5 mutations (1.9%). In group 1, the most reported mutations are the ones in positions 3036, 14408, 23403 and 28881. Regarding genomes in group 2, 20 do not carry any mutations (23.8%), 25 genomes have a single mutation (29.8%), 19 genomes have two mutations (22.6%), 6 genomes have three mutations (7.1%), 9 genomes have four mutations (10.7%), 2 and 3 genomes have five and six mutations respectively (2.4% and 3.6%). In group 2, the most reported mutations are located at positions 8782, 11083, 17746 and 17857.

The distribution between the two groups in terms of number of mutations is statistically relevant (Fisher-Exact test, p-value < 0.001). In particular, group 1 and 2 are significantly different in terms of the distribution of genomes having 0, 1, 3 and 4 numbers of mutations (Fisher-Exact test, $p < 0.001$) (Fig. 4). This difference, instead, is insignificant when the number of mutations is 2, 5 or 6.

We found that viral strains with RdRp mutation have a median of 3 point mutations [range:2–5], whereas viral strains with no RdRp mutation have a median of 1 mutation [range:0–3] (p value $<$ 0.001, Mann-Whitney test). The different distribution between the two groups relative to the number of mutations is statistically significant (Fig. 4).

We also analyzed the most frequent mutations detected: the ones in positions 3036, 23403 and 28881 (in Europe), and the ones in positions 17746, 17857 and 18060 (in North America). Viral genomes carrying each one of these mutations were compared with viral genomes without mutations, by using Mann-Whitney test for paired-groups comparison analysis. Genomes carrying mutations in positions 3036, 23403, 28881, 17746, 17857 and 18060 show a median of 3–4 mutations (range [2:5]), whereas genomes carrying none of them have a median of 1 or 2 mutations (range [0:3], p-

value < 0.001, Mann-Whitney test). This difference is statistically significant and implies that if one of those mutations is present, other mutations are more likely to occur.

Homology study of mutant RdRp protein

Among all mutation sites analyzed, RdRp mutant is particularly interesting given that the enzyme is directly involved in viral replication and its fidelity determines the mutagenic capabilities of SARS-CoV-2. Due to the high homology between RdRps of SARS-CoV and SARS-CoV-2, we aligned SARS-CoV-2 RdRp reference sequence with the reported catalytic site sequence of SARS-CoV RdRp.

The amino acid substitution 323 (P to L) (due to nucleotide mutation 14408) falls outside the catalytic site, in a region that in SARS-CoV is reported to be an Interface Domain, a still poorly characterized surface structure, supposedly implicated in the interaction with other proteins which may regulate the activity of RdRp[16]. To this regard, it is well-known that SARS-CoV RdRp forms a hollow cylinder-like supercomplex with nsp7 and nsp8, which confer processivity to RdRp[28]. Additionally, replication supercomplex interacts with nsp14, an exonuclease having the Nidovirales-typical proofreading capability. This activity is important in the context of the mutation rate and for controlling the fidelity in RNA replication. However, critical RdRp residues involved in this interaction are still to be identified, and for this reason further studies are needed to assess the possible role of mutation 14408 concerning RdRp fidelity.

Discussion

In the present work we have compared the SARS-CoV-2 reference genome to those exported from the GISAID database with the aim of gaining important insights into virus mutations, their occurrence over time and within different geographic areas.

We observed that after February 2020, when the first locally transmitted SARS-CoV-2 cases out of Asia were reported, viral genomes presented different point mutations, clearly distinguishable within different geographic areas. Over time, we were able to identify three recurrent mutations in Europe (in positions 3036, 14408 and 23403) and 3 other different mutations in North America (in positions 17746, 17857 and 18060). So far, these mutations have not been detected in Asia. The number and the occurrence, as well as the median value of virus point mutations registered out of Asia, increase

over time.

In our study, we found that the RdRp mutation, located at position 14408, which is present in European viral genomes starting from February 9th 2020, is associated with a higher number of point mutations compared to viral genomes from Asia. Given that RdRp works in a complex machinery that includes proofreading activities (in cooperation with other viral cofactors, like ExoN, nsp7 and nsp8), it is tempting to speculate that this mutation have contributed in impairing its proofreading capability. One possible mechanism could involve a minor change in the RdRp structure, without affecting its catalytic activity, that might nonetheless altered its binding capability with other cofactors such as ExoN, nsp7 or nsp8, thus altering the mutation rate. This could explain the increased number of mutations that we observed in Europe since February 2020. Further studies are needed to determine whether the observed cluster mutations originate from the same molecular mechanism. Further studies are also needed to determine whether the mutation in RdRp results in increased viral replication.

Some polymerase inhibitors[29,30] are currently being tested in clinical studies to target SARS-CoV-2 RdRp, including Favipiravir[17,19], Galidesivir[18], Remdesivir[19], Ribavirin[20], Penciclovir[31], Galidesivir[32] and Ponatinib[33]. Additionally, other drugs such as Simeprevir (FDA approved HCV protease inhibitor), as well as Filibuvir and Tegobuvir (both RdRp inhibitors)[22], are predicted to bind RdRp by molecular docking studies. In particular, a putative docking site was identified in a hydrophobic cleft very close to the mutated site 323 (P to L), corresponding to mutation 14408 identified in our study[22]. Naturally occurring mutations in RdRp can potentially lead to drug-resistance phenomena, as already observed previously[19,34]. Alternatively, it might induce a significant decrease in drug-RdRp complex binding affinity. This could lead to different effectiveness of antiviral treatments where mutation 14408 is present. For this reason, due to the high frequency of RdRp mutation in the infected population, it is important to characterize the impact of 14408 mutation on the activity of RdRp and its susceptibility to antiviral drugs.

Conclusions

We identified novel mutation hotspots in the SARS-CoV-2 genome sequences. Interestingly, some appeared after February 2020, only in European patients. Among these hotspots, one mutation in position 14408 is located within the RdRp protein and is associated with an overall increased mutation rate. An in silico analysis comparing annotated functional domains of SARS-CoV and SARS-CoV-2 sequences, showed that this particular mutation occurs in the so-called RdRp interface domain, a still poorly characterized surface structure, involved in protein-protein interactions[16]. The role for the RdRp interface domain requires further investigations, and in particular the effect of mutation in position 14408, its interaction with other cofactors (such as ExoN, nsp7 and nsp8), possibly affecting its proofreading activity and potentially altering its mutation rate. It is also essential to understand if the described mutations could result in the emergence of drug-resistance viral phenotypes. Our data may help the development of diagnostic and therapeutic strategies and to study potential drug resistance mechanisms.

Abbreviations

RdRp: RNA-dependent RNA polymerase;

nsp: non-structural protein

SARS-CoV: Severe Acute Respiratory Syndrome Coronavirus

SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2 (COVID-19)

MERS-CoV: Middle East Respiratory Syndrome Coronavirus

ExoN: exonuclease

WHO: World Health Organization

WSM: Wuhan Seafood Market

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets analysed during the current study are available in the GISAID (https://www.gisaid.org/) and GenBank (https://www.ncbi.nlm.nih.gov/nuccore/NC_045512) repositories.

Competing interests

The authors declare that they have no competing interests

Funding

No funding was used to conduct this research.

Author's contributions

MP, BM and FG conducted the analysis. MP, BM, FB, DZ and RI wrote the paper. CM, MC, SA, PS, EM, RCG and RI revised the paper and participate in the scientific discussion of these results. RI and DZ supervised the project. All authors contributed to the final revision of the project.

Acknowledgements

We would like to thank Joe R. Kates for its contribution in the stimulating scientific discussion.

References

1. Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G. *Real estimates of mortality following COVID-19 infection.* Lancet, 2020. DOI:10.1016/S1473-3099(20)30195-X.
2. Duffy et al. *Why are RNA virus mutation rates so damn high?* Plos Biology, 2018.DOI:10.1371/journal.pbio.3000003.
3. Wang C, Liu Z, Chen Z, Huang X, Xu M, He T. Zhang Z.. *The establishment of reference sequence for SARS-CoV-2 and variation analysis.* Journal of Medical Virology, 2020. DOI: 10.1002/jmv.25762.
4. Domingo E. *Viruses at the edge of adaptation*. Virology 2000. DOI: 10.1006/viro.2000.0320.
5. Domingo E. *Quasispecies theory in virology.* J. Virol. 2002. DOI: 10.1128/JVI.76.1.463-465.2002.
6. Domingo E, Holland JJ. *RNA virus mutations and fitness for survival.* Annu. Rev.

Microbiol. 1997. DOI: 10.1146/annurev.micro.51.1.151.

7. Bruenn, J.A. *A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases.* Nucleic Acids Res. 2003. DOI: 10.1093/nar/gkg277.
8. Venkataraman S, Prasad BVLS, Selvarajan R. *RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution.* Viruses 2018. DOI: 10.3390/v10020076
9. Jia H, Gong P. *A Structure-Function Diversity Survey of the RNA-Dependent RNA Polymerases From the Positive-Strand RNA Viruses.* Front. Microbiol., 2019. DOI: 10.3389/fmicb.2019.01945.
10. Ziebuhr J *The coronavirus replicase.* Curr Top Microbiol Immunol 2005. DOI: 10.1007/3-540-26765-4_3
11. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Zu L, Holmes, Zhang YZ. *A new coronavirus associated with human respiratory disease in China*. Nature 2020. DOI: 10.1038/s41586-020-2008-3
12. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, Wang T, Sun Q, Ming Z, Zhang L, Ge J, Zheng L, Zhang Y, Wang H, Zhu Y, Zhu C, Hu T, Hua T, Zhang B, Yang X, Li J, Yang H, Liu Z, Xu W, Guddat LW, Wang Q, Lou Z, Rao Z. *Structure of RNA-dependent RNA polymerase from 2019-nCoV, a major antiviral drug target*. bioRxiv 2020. DOI:/10.1101/2020.03.16.993386
13. Mirza MU, Froeyen M. *Structural elucidation of SARS-CoV-2 vital proteins: computational methods reveal potential drug candidates against main protease, Nsp12 RNA-dependent RNA polymerase and NSP13 helicase.* Preprints 2020. DOI: 10.20944/preprints202003.0085.v1
14. Ogando NS, Ferron F, Decroly E, Canard B, Posthuma CC, Snijder EJ. *The Curious Case*

of the Nidovirus Exoribonuclease: Its Role in RNA Synthesis and Replication Fidelity - Front. Microbiol., 2019. DOI: 10.3389/fmicb.2019.01813

15. Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, Scherbakova S, Graham RL, Baric RS, Stockwell TB, Spiro DJ, Denison MR. *Infidelity of SARS-CoV Nsp14-Exonuclease Mutant Virus Replication Is Revealed by Complete Genome Sequencing.* Plos pathogens, 2010. DOI: 10.1371/journal.ppat.1000896
16. Kirchdoerfer RN, Ward AB *Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors*, Nature Comm 2019. DOI: 10.1038/s41467-019-10280-3.
17. Furuta Y, Gowen BB, Takahashi K, Shiraki K, Smee DF, Barnard DL. *Favipiravir (T-705), a novel viral RNA polymerase inhibitor.* Antiviral Res. 2013. DOI: 10.1016/j.antiviral.2013.09.015.
18. Lim, SY, Osuna C, Lakritz J, Chen E, Yoon G, Taylor R, MacLennan S, Leonard M, Giuliano P, Mathis A, Berger E, Babu Y, Sheridan W, Whitney J. G*alidesivir, a Direct-Acting Antiviral Drug, Abrogates Viremia in Rhesus Macaques Challenged with Zika Virus*. Open Forum Infect Dis 2017. DOI: 10.1093/ofid/ofx162.129
19. Agostini, MK, Andre EL, Sims AC, Graham RL, Sheahan TP, Lu X, Smith EC, Case JB, Feng JY, Jordan R, Ray AS, Cihlar T, Siegel D, Mackman RL, Clarke MO, Baric RS, Denison MR. *Coronavirus Susceptibility to the Antiviral Remdesivir (GS-5734) Is Mediated by the Viral Polymerase and the Proofreading Exoribonuclease.* MBio 2018. DOI: 10.1128/mBio.00221-18
20. Morgenstern B, Michaelis M, Baer PC, Doerr HW, Cinatl JJr. *Ribavirin and interferon-β synergistically inhibit SARS-associated coronavirus replication in animal and human cell lines.* Biochem. Biophys. Res. Commun. 2005. DOI: 10.1016/j.bbrc.2004.11.128.
21. Chang Y, Tung Y, Lee K, Chen T, Hsiao Y, Chang H, Hsieh T, Su C, Wang S, Yu J, Shih S, Lin Y, Lin Y, Tu Y.E, Tung C, Chen C. *Potential Therapeutic Agents for COVID-19*

Based on the Analysis of Protease and RNA Polymerase Docking. Preprints 2020. DOI:10.20944/preprints202002.0242.v1.

22. Ruan Z, Liu C, Guo Y, He Z, Huang X, Jia X, Yang T. *Potential Inhibitors Targeting RNA-Dependent RNA Polymerase Activity (NSP12) of SARS-CoV-2.* Preprints 2020 DOI: 10.20944/preprints202003.0024.v1.
23. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu J. On the origin and continuing evolution of SARS-CoV-2, *National Science Review*, 2020. DOI:10.1093/nsr/nwaa036.
24. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, Zhou Z, Yang J, Zhong J, Yang D, Guo L, Zhang G, Li H, Xu Y, Chen M, Gao Z, Wang J, Ren L, Li M. *Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients*. Clin Infect Dis 2020. DOI: 10.1093/cid/ciaa203.
25. Phan T. *Genetic diversity and evolution of SARS-CoV-2.* Infection, genetics and evolution, 2020. DOI: 10.1016/j.meegid.2020.104260.
26. Pfeiffer J, Kirkegaard K. *A single mutation in poliovirus RNA-dependent RNA polymerease confers resistance to mutagenic nucleotide analogs via increased fidelity.* PNAS 2003. DOI:10.1073/pnas.1232294100.
27. https://experience.arcgis.com/experience/685d0ace521648f8a5beeeee1b9125cd
28. Zhai Y, Sun F, Li X, Pang H, Xu X, Bartlam M, Rao Z. *Insights into SARS-CoV transcription and replication from the structure of the nsp7-nsp8 hexadecamer.* Nature Structural and Molecular biology, 2005. DOI: 10.1038/nsmb999.
29. Guangdi L, De Clercq E. *Therapeutic options for the 2019 novel coronavirus (2019-nCoV)* Nature Reviews Drug Discovery 2020. DOI: 10.1038/d41573-020-00016-0
30. Gordon DE et al. *A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing.* bioRxiv 2020. DOI:

10.1101/2020.03.22.002386

31. Manli W, Ruiyuan C, Leike Z, Yang X, Liu J, Xu M, Shi Z, Hu Z, Zhong W, Xiao G.. *Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro.* Cell Res. 2020. DOI: 10.1038/s41422-020-0282-0.

32. Warren TK, Wells J, Panchal RG, Stuthman KS, Garza NL, Van Tongeren SA, Dong L, Retterer CJ, Eaton BP, Pegoraro G, Honnold S, Bantia S, Kotian P, Chen X, Taubenheim BR, Welch LS, Minning DM, Babu YS, Sheridan WP, Bavari S. *Protection against filovirus diseases by a novel broad-spectrum nucleoside analogue BCX4430.* Nature 2014. DOI: 10.1038/nature13027.

33. Najjar M, Suebsuwong C, Ray SS, Thapa RJ, Maki JL, Nogusa S, Shah S, Saleh D, Gough PJ, Bertin J, Yuan J, Balachandran S, Cuny GD, Degterev A. *Structure guided design of potent and selective ponatinib-based hybrid inhibitors for RIPK1.* Cell Rep. 2015. DOI: 10.1016/j.celrep.2015.02.052

34. Young KC1, Lindsay KL, Lee KJ, Liu WC, He JW, Milstein SL, Lai MM. *Identification of a ribavirin-resistant NS5B mutation of hepatitis C virus during ribavirin monotherapy.* Hepatology 2003.DOI: 10.1053/jhep.2003.50445

35. Goldhill DH, te Velthuis AJW, Fletcher RA, Langat P, Zambon M, Lackenby A, and Barclay WS. The mechanism of resistance to favipiravir in influenza. PNAS 2018. DOI:10.1073/pnas.1811345115.

Figures

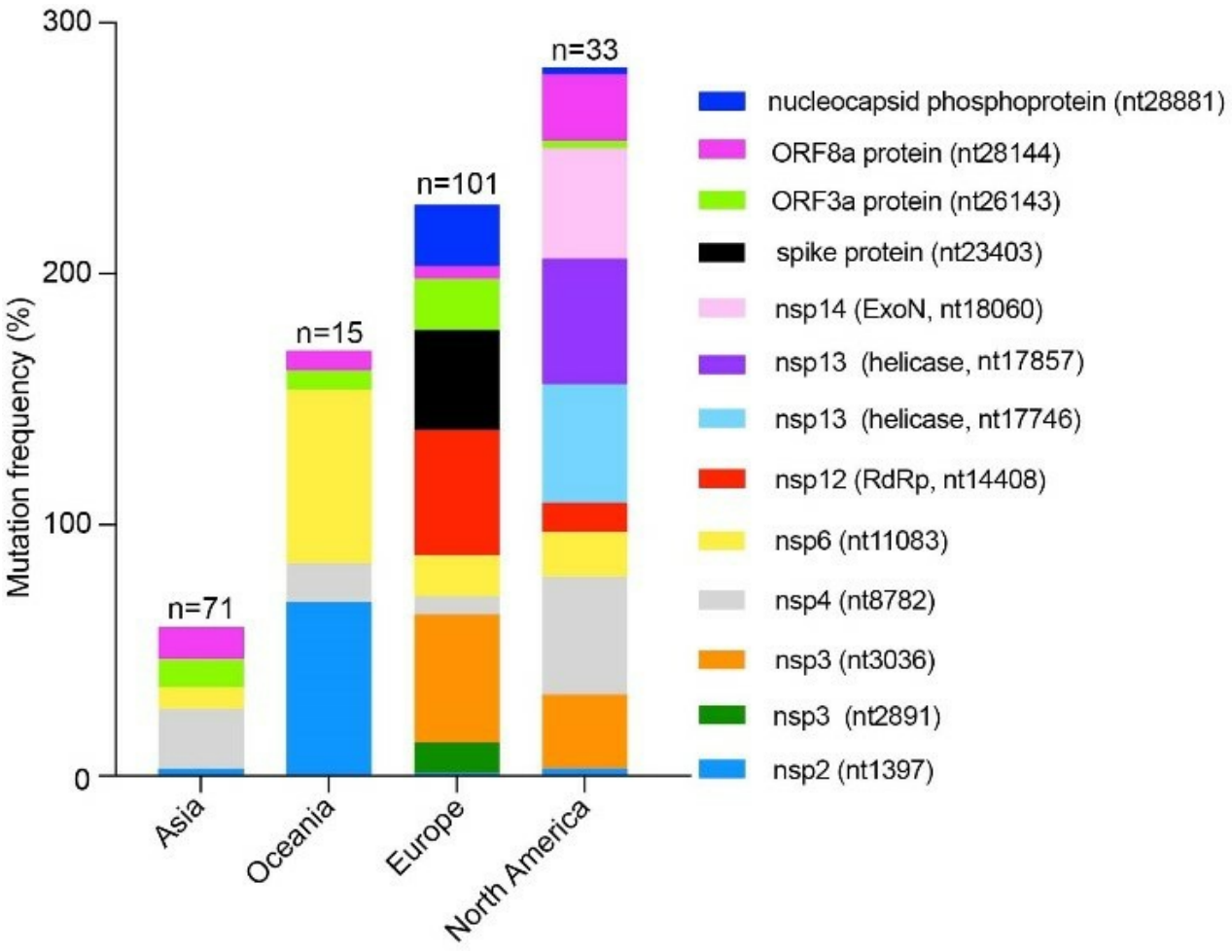

Figure 1

SARS-CoV-2 mutation frequency in different geographic areas. Eight novel recurrent hotspots mutations (namely 1397, 2891, 14408, 17746, 17857, 18060, 23403 and 28881) and 5 hotspots already reported in literature (namely 3036, 8782, 11082, 28144 and 26143) were subdivided into 4 geographic areas: Asia (n=71), Oceania (n=15), Europe (n=101), North America (n=33). The mutation frequency was estimated for each of them, by normalizing the number of genomes carrying a given mutation in a geographic area, by the overall number of retrieved genomes per geographic area; the graph shows the cumulative mutation frequency of all given mutations present in each geographic area. Mutation locations in viral genes are reported in the legend as well as the proteins (i.e. non-structural protein, nsp) presenting this mutations. The figure shows that genomes from European and North American patients present an increase in mutation frequency compared to Asia. It is

also possible to observe that Europe and North America show a differential pattern of mutations: mutation 14408 (red), 23404 (black), 28881 (electric blue) and 26143 (light green) are present mostly in Europe, whereas 18060 (pink), 17875 (purple) and 17746 (light blue) are present mostly in North America.

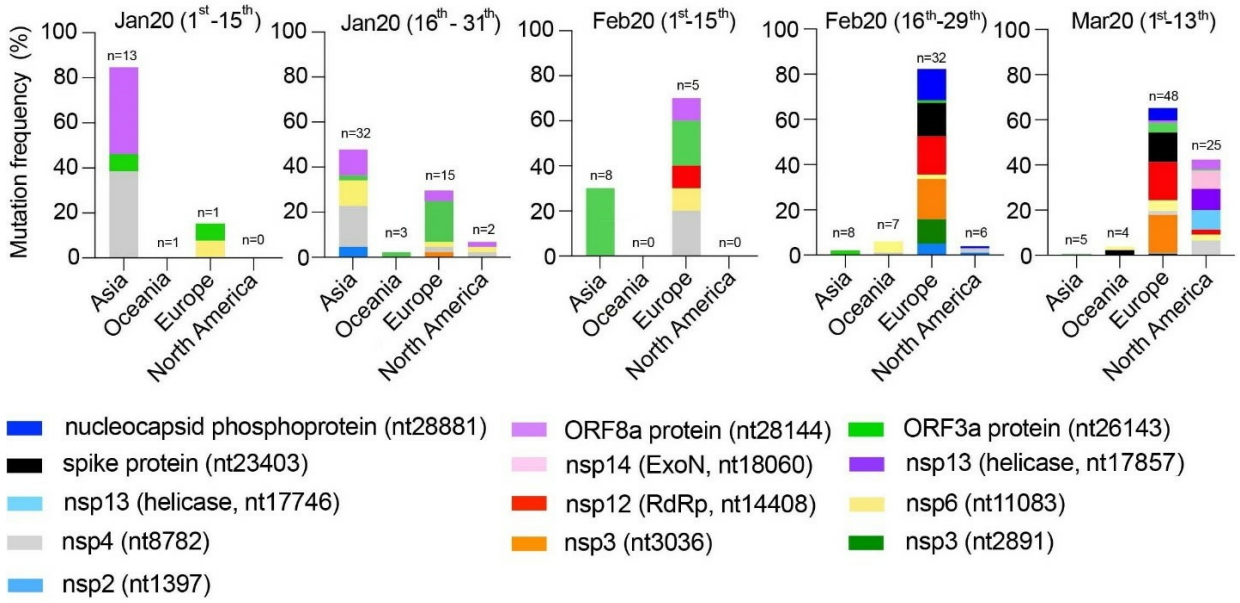

Figure 2

SARS-CoV-2 Mutation occurrence over time divided per geographic area. Eight novel recurrent hotspots mutations (namely 1397, 2891, 14408, 17746, 17857, 18060, 23403 and 28881) and 5 hotspots already reported in literature (namely 3036, 8782, 11082, 28144 and 26143) were subdivided first into 5 period subgroups: December 2019 (n=5), 1st-15th Jan. 2020 (n=15), 16th-31st Jan. 2020 (n=52), 1st-15th Feb. 2020 (n=13), 16th-29th Feb- 2020 (n=55) and 1st-13th Mar. 2020 (n=80). Next, for each time group, a further subclassification per geographic area (Asia, Oceania, Europe and North America) was performed (number of genomes in each area are reported in the figure inset). The number of mutations in each area was normalized by the number of genomes analyzed for each period of time. This figure shows that mutation frequency increases over time during viral spread out of Asia. No mutations were observed in the Asian genomes analyzed in December 2019. In the time group of February 16th-29th, a defined cluster of mutations emerged in Europe; in March 1st-13th, a different cluster of mutations emerged in North America.

	n pts before	n pts after	1397	2891	3036	8782	11083	14408	17746	17857	18060	23403	26143	28144	28881
Europe	17	81	+2.5%	+14.8%	+61.7%	-12.7%	-22.9%	+60.5%	0%	0%	0%	48.1%	-46.5%	-15.2%	29.6%
Asia	59	12	-3.4%	0%	0%	11.3%	-0.14%	0%	0%	0%	0%	0%	+26.6.%	-15.2%	0%
North America	0	28	0%	0%	+35.7%	+42.9%	+17.9%	+14.3%	+57.2%	+60.7%	+53.4%	0%	+3.6%	+21.4%	+3.6%

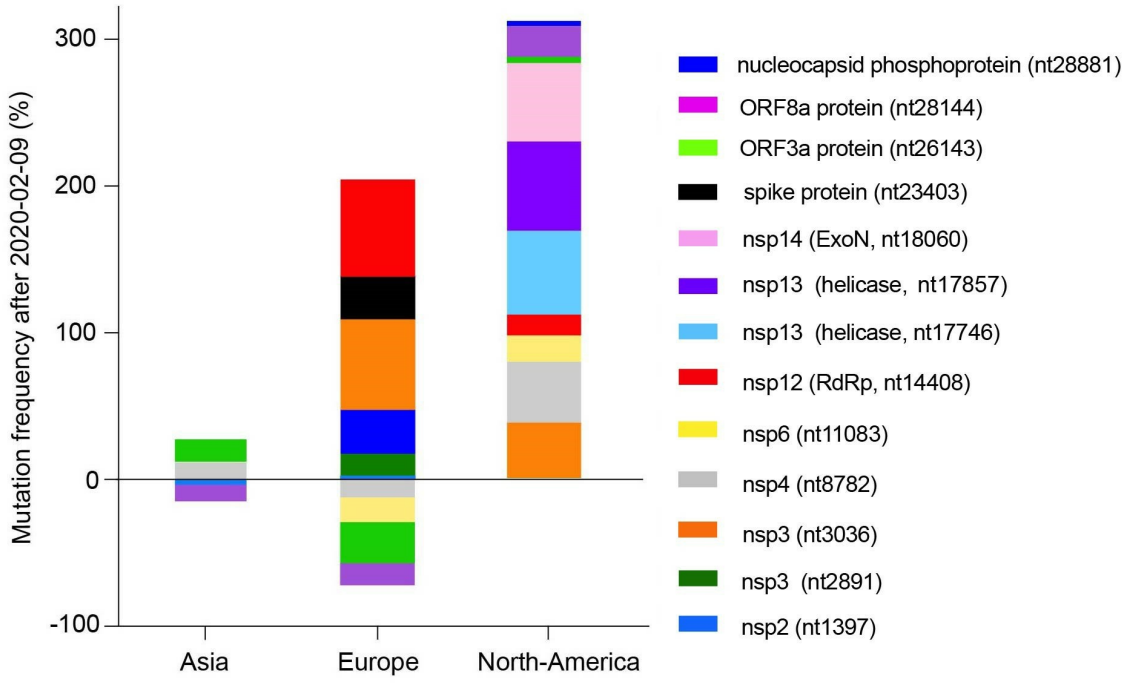

Figure 3

Increment of SARS-CoV-2 mutation frequency after RdRp mutation appearance per geographic area. The increment of mutation frequency before and after February 9th, 2020 across the different geographic areas (Asia, Europe, North America) is shown. The figure shows a diminishment of Asian mutations (i.e. 1397, 8782, 11083, 26143 and 28144) is simultaneous with the appearance of new mutations such as 2891, 3036, 23403 and 28881, when RdRp novel European mutation located at 14408 (in red) occurred. The upper table shows the increment or decrement for each single mutation, per geographic area.

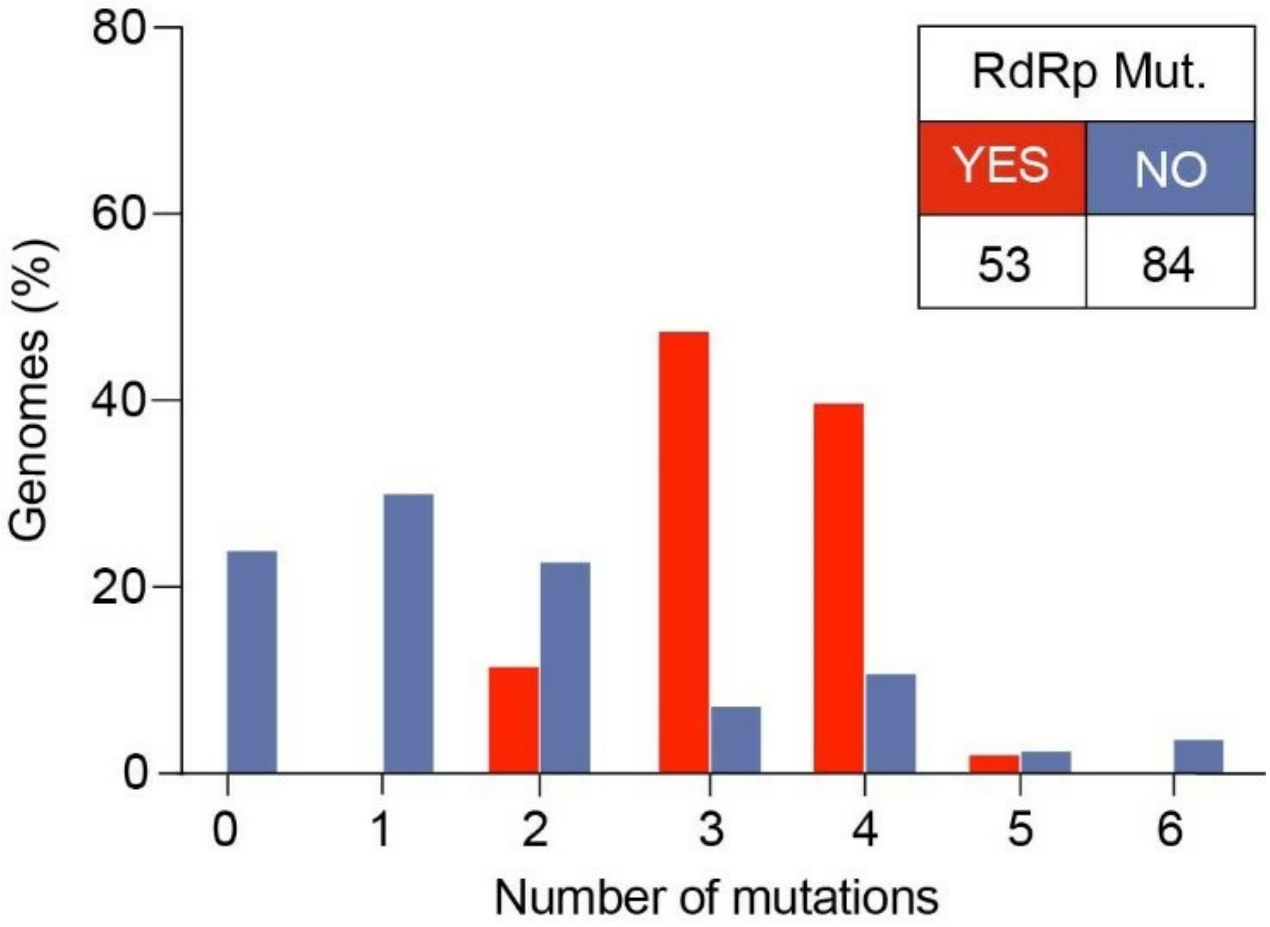

Figure 4

Number of SARS-CoV-2 mutations associated with the RdRp mutation. Genomes were subdivided into two groups: group 1 contains genomes with mutation in position 14408 (RdRp) (n=53, 4 North America and 49 European), and group 2 without RdRp mutation (n=84). We further subdivided group 1 and 2 by the number of mutations present in the genome. Genomes in group 1 (red bars) showed an increased number of mutations compared to group 2 (grey bars). Most genomes of groups 1 (86.8%) have at least 3 or 4 mutations, whereas 76.2% of genomes of group 2 have less than 2 mutations. We found that viral strains with RdRp mutation have a median of 3 point mutations [range:2-5], whereas viral strains with no RdRp mutation have a median of 1 mutation [range:0-3] (p value < 0.001, Mann-Whitney test).