Title: Large-scale population analysis of SARS-CoV-2 whole genome sequences reveals host-mediated viral evolution with emergence of mutations in the viral Spike protein associated with elevated mortality rates.

Authors

Carlos Farkas[1,2*], Andy Mella[3] and Jody J. Haigh[1,2*]

Running Title

Host-mediated viral evolution in SARS-CoV-2 and mutations in the Spike protein.

Affiliations

[1] Research Institute in Oncology and Hematology (RIOH), CancerCare Manitoba, Winnipeg, MB, Canada

[2] Department of Pharmacology and Therapeutics, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, Canada

[3] Departamento de Física, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Blanco Encalada 2008, Santiago, Casilla 487-3, Chile.

***corresponding author(s) e-mail address(es)**

Carlos.FarkasPool@umanitoba.ca, jody.haigh@umanitoba.ca

Abstract

Background

We aimed to further characterize and analyze in depth intra-host variation and founder variants of SARS-CoV-2 worldwide up until August 2020, by examining in excess of 94,000 SARS-CoV-2 viral sequences in order to understand SARS-CoV-2 variant evolution, how these variants arose and identify any increased mortality associated with these variants.

Methods and Findings

We combined worldwide sequencing data from GISAID and Sequence Read Archive (SRA) repositories and discovered SARS-CoV-2 hypermutation occurring in less than 2% of COVID19 patients, likely caused by host mechanisms involved APOBEC3G complexes and intra-host microdiversity. Most of this intra-host variation occurring in SARS-CoV-2 are predicted to change viral proteins with defined variant signatures, demonstrating that SARS-CoV-2 can be actively shaped by the host immune system to varying degrees. At the global population level, several SARS-CoV-2 proteins such as Nsp2, 3C-like proteinase, ORF3a and ORF8 are under active evolution, as evidenced by their increased $\pi N/\pi S$ ratios per geographical region. Importantly, two emergent variants: V1176F in co-occurrence with D614G mutation in the viral Spike protein, and S477N, located in the Receptor Binding Domain (RBD) of the Spike protein, are associated with high fatality rates and are increasingly spreading throughout the world. The S477N variant arose quickly in Australia and experimental data support that this variant increases Spike protein fitness and its binding to ACE2.

Conclusions

SARS-CoV-2 is evolving non-randomly, and human hosts shape emergent variants with positive fitness that can easily spread into the population. We propose that V1776F and S477N variants occurring in the Spike protein are two novel mutations occurring in SARS-CoV-2 and may pose significant public health concerns in the future.

Author Summary

We have developed an efficient bioinformatics pipeline that has allowed us obtain the most complete picture to date of how the SARS-CoV-2 virus has changed during the last eight month global pandemic and will continue to change in the near future. We characterized the importance of the host immune response in shaping viral variants at different degrees, evidenced by hypermutation responses on SARS-CoV-2 in less than 2% of infections and positive selection of several viral proteins by geographical region. We underscore how human hosts are shaping emergent variants with positive fitness that can easily spread into the population, evidenced by variants V1176F and S477N, located in the stalk and receptor binding domains of the Spike protein, respectively. Variant V1176 is associated with increased mortality rates in Brazil and variant S477N is associated with increased mortality rates over the world. In addition, it has been experimentally demonstrated that S477N variant increase fitness of Spike protein and its binding with ACE2, thus predicting to increase virulence of SARS-CoV-2. This limits the concept of 'herd immunity' proposals and re-emphasize the need to limit the spread of the virus to avoid emergence of more virulent forms of SARS-CoV-2 that can spread worldwide.

Introduction

The novel SARS-CoV-2 coronavirus that causes COVID19 has surpassed 34 million infections worldwide within nine months of pandemic, resulting in more than one million deaths until September 2020 (https://coronavirus.jhu.edu/map.html) (1). In-depth characterization of this virus is urgently needed to improve outbreak surveillance, vaccine development and for effective treatments now and in the immediate future. SARS-CoV-2 is a positive single-stranded RNA virus (+ssRNA) with a crown-like appearance observed by electron microscopy that is due the presence of the of spike glycoproteins on the lipid bilayer envelope (2, 3). Another three transmembrane proteins are incorporated into the envelope: small envelope protein (E), matrix protein (M), and nucleocapsid protein (N) (4). As seen with SARS-CoV-1, SARS-CoV-2 binds through its Spike glycoprotein to cell membrane-bound angiotensin-converting enzyme 2 (ACE2) for entry into host cells (5-8). Advancements in COVID19 treatments have been recently developed including Remdesivir, a nucleoside analog that inhibits viral RNA-dependent RNA polymerase and is an effective treatment to reduce viral titers in rhesus macaques that is clinically approved for COVID19 treatment (9). As well, either wild-type or catalytically inactive ACE2 has been demonstrated to block viral entry *in vitro*, and are proposed as promising treatments (10, 11). A remaining question is how the human humoral immune response develops after SARS-CoV-2 infection. Studies in Iceland have shown that around 90% of infected patients develop antiviral antibodies that last up to four months (12), but it has also been suggested that around one third of the seropositive infections are asymptomatic and become antibody-negative early in the convalescence period (13). Also, the unexpectedly low secondary infection risk reported for SARS-CoV-2 infection suggests innate immune responses are active in humans (14, 15). To explore host– SARS-CoV-2 interactions at the genetic level it is useful to analyze viral sequencing results per individual and at the population level. Initiatives such as

GISAID (https://www.gisaid.org/) (16, 17) and the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) have been storing SARS-CoV-2 sequencing datasets worldwide from the beginning of the pandemic starting in January 2020, allowing researchers to track fixed variants and follow viral evolution by geographical region. The unprecedented amount of SARS-CoV-2 whole genome sequencing data can help to 1) characterize viral variants that occur within a given host, 2) understand variant fixation in a given population and 3) understand how the virus changes over time. In fact, the Spike protein mutation D614G global transition that occurred very recently was discovered in this way and is associated with higher viral titers and higher fatality rates (18, 19). Thus, it is probable that more mutations are to be discovered by tracking SARS-CoV-2 genomic changes globally. In this study we aimed to characterize in depth intra-host variation and population-fixed variants worldwide up until the beginning of August 2020 by using over 76,000 SARS-CoV-2 sequences and 17,500 sequencing datasets from GISAID and SRA repositories, respectively. First, we found evidence for SARS-CoV-2 hypermutation, occurring in less than 2% of COVID19 patients. This mechanism is predicted to inactivate the virus and is likely caused by host mechanisms involved APOBEC3G complexes and intra-host microdiversity, where G>T transversions and C>T transitions are frequent signatures observed both in hypermutant and non-hypermutant samples. These results suggest that SARS-CoV-2 is actively shaped by the host immune system to varying degrees. From a population context, several SARS-CoV-2 proteins such as Nsp2, 3C-like proteinase, ORF3a and ORF8 are under active evolution, evidenced by their increasing $\pi N/\pi S$ ratios. Noteworthy, most of the population-fixed variants in SARS-CoV-2 are predicted to destabilize viral proteins, as already reported for other RNA viruses. Of these variants, those occurring in the ORF3a (Q57H), Nucleocapsid (I292T, RG203KR) and Spike protein (V1176F) have a

positive association with increased mortality ratios in populations from Saudi-Arabia and Brazil, respectively. In particular, the V1176F variant co-occurs with the D614G mutation in the Spike protein mutation in Brazil and arose independently in at least in three independent SARS-CoV-2 clades. This variant is predicted to stabilize the SARS-CoV-2 Spike trimmer complex and confer flexibility to the stalk domain of the trimmer, potentially facilitating Spike binding properties to ACE2. Also, this variant is associated with increased mortality ratios in Brazil and is increasingly spreading throughout the world. Similarly, the emerging variant S477N occurring in the Receptor Binding Domain, dramatically increase its frequency and became dominant in Australia within two months. Experimental data support that S477N increase both fitness and binding to ACE2 receptor, explaining its selection among other viruses in Australia. S477N also is presently spreading across countries and is associated with higher fatalities throughout the world. We propose that these variants are novel mutations occurring in SARS-CoV-2 and their spread may pose serious concerns in public health in the future of the pandemic.

Methods

Data and Code Availability

76,553 FASTA genomes and associated sequencing metadata were downloaded from GISAID database from January 1, 2019 until August 3, 2020, specifying "human" as source host (https://www.gisaid.org/). The associated sequencing metadata including major variants per sample are available at Supplementary Table 1. 974 Brazilian FASTA sequences were downloaded from GISAID database from January 1, 2019 until September 25, 2020, specifying "human" as source host and "South America / Brazil" as location. Acknowledgements to all laboratories/consortia involved in the generation of GISAID genomes used in this study are listed in Supplementary Table 2. 17,560 sequencing datasets were downloaded from Sequence Read Archive Repository (SRA, https://www.ncbi.nlm.nih.gov/sars-cov-2/) From December 1, 2019 until July 28, 2020. Associated sequencing run accessions, sequencing metadata and related BioProjects are listed in Supplementary Table 3. The code generated during this study to replicate most of the computational calculations performed in this manuscript is available at the following github repository: https://github.com/cfarkas/SARS-CoV-2-freebayes.

Next-generation sequencing and FASTA dataset processing

To process next generation sequencing datasets, we employed our pipeline (SARS-CoV-2_freebayes) consisting in a bash/UNIX script that pipes several programs in sequential order. Imputed list of SRA accessions is processed with sra-tools, https://github.com/ncbi/sra-tools), generating compressed FASTQ files per sequencing, automatically trimmed with fastp tool (20). Minimap2 splice-aware aligner in preset mode -ax sr (21) align each trimmed fastq file against a provided reference genome (Wuhan-Hu-1, GenBank Accession: MN908947.3). The resulting

BAM files were sorted and indexed by using Samtools (22). Freebayes, as frequency-based pooled caller (-F 0.49) (https://github.com/ekg/freebayes) (23) perform variant calling on every sorted BAM file, obtaining major frequency viral variants per genome in VCF format,. Then, the Jacquard program (https://jacquard.readthedocs.io/en/v0.42/index.html) in python environment (24) merges every VCF file containing variants associated to each bam file into a single VCF file, containing aggregated variants from all genomes. Viral frequencies were recalculated in the merge VCF file by using several UNIX tools (25), in combination with vcflib (https://github.com/vcflib/vcflib). Variants per genome reported in the resulting file "logfile_variants_SRA_freebayes" were used to construct Figure 1B using GraphPad Prism 8 software (https://www.graphpad.com/scientific-software/prism/). GISAID FASTA genomes were processed in a similar manner. We preprocess a single GISAID genome collection with SeqKit (26) to decompose a single FASTA file into individual FASTA files, each file containing a single genome. Then, Minimap2 aligner with preset -ax asm5 (21) align every FASTA genome against SARS-CoV-2 reference genome. Freebayes variant caller with --min-alternate-count 1 (C 1) option (https://github.com/ekg/freebayes) perform variant calling on each BAM file, outputting variants in VCF format. With these operations, major frequency viral variants in VCF format are obtained from each FASTA genome. Then, variants are aggregated into a single VCF file, as described with Jacquard. Figure 1A graph was constructed by using variants per genome, reported in the output file "logfile_variants_GISAID_freebayes", inputted into the GraphPad Prism 8 software. All these computational analyses are described here: https://github.com/cfarkas/SARS-CoV-2-freebayes (case examples I and II, respectively).

Variant Visualization

The Integrative Genomics Viewer (IGV) software (http://software.broadinstitute.org/software/igv/home) was used to visualize next generation sequencing alignments in bam format (27-29). To visualize mayor viral frequency variants, the variant frequency threshold was set at 0.49.

SnpEff annotation

Merged variants from GISAID genomes (n=76563) were annotated by using in a repurposed version of SnpEff program, available in the Galaxy server (30-32). The resulting annotated VCF file was parsed by using conventional UNIX tools. All these computational analyses are described here: https://github.com/cfarkas/SARS-CoV-2-freebayes (case example III).

πN/πS calculation

We estimated nonsynonymous and synonymous nucleotide diversity (πN and πS, respectively) in 1279, 6841, 46042, 17989, 1205 and 2924 GISAID FASTA genomes from Africa, Asia, Europe, North America, South America, and Oceania, respectively. The FASTA files were processed from the alignment to the variant calling step as described in "GISAID FASTA dataset processing section". All these computational analyses are described here: https://github.com/cfarkas/SARS-CoV-2-freebayes (case example IV).

Microdiversity and low frequency viral variants.

We estimated nucleotide diversity in 397, 448 and 308 next generation sequencing (NGS) samples from Australia, Spain, and USA populations, respectively by using aligned reads per

sample in BAM format against SARS-CoV-2 reference genome. These BAM files were inputted in loop to InStrain program (https://instrain.readthedocs.io/en/latest/) (33), obtaining several parameters such as analysis of coverage, microdiversity, SNV linkage, and sensitive SNP detection, among others. As recommended by inStrain, we analyzed only sequencing samples with sufficient breadth of coverage (>0.9), resulting in 397, 374 and 216 NGS samples from Australia, Spain and from USA, respectively. The list of the NGS samples in the three populations, including the referred calculations are detailed in the spreadsheet inStrain_results.xlsx, available here: https://github.com/cfarkas/SARS-CoV-2-freebayes. We correlated in each country the number of variants with viral frequency > 5% against the nucleotide diversity (π) by using Spearman correlation. Spearman's correlation coefficients (r) and confident p-values (P, to discard random sampling) were calculated in GraphPad Prism 8. The significance thresholds were as follows: P<0.05 *, P<0.01 **, P<0.001 ***, P<0.0001 ****, P>0.05 ns. All these computational analyses are described here: https://github.com/cfarkas/SARS-CoV-2-freebayes (case example V).

SNP-mortality associations

We downloaded 7634 genomes with associated metadata from GISAID until September 28, 2020 and we grouped the genomes from released/deceased patients per country (India, Saudi Arabia, USA, and Brazil, respectively). Then, we parsed genomes and associated metadata by country (in particular, deceased and released cases) by using a combination of standard UNIX tools, vcflib (https://github.com/vcflib/vcflib) and BEDOPS (34). After these steps, we uploaded to the Galaxy server (https://usegalaxy.org/) the resulting output per country (Deceased-Released.subset file) (31, 35) and we performed Fisher's exact test to identified variants with a significant difference in the viral frequencies between the groups (snpFreq program,

https://rdrr.io/github/lvclark/SNPfreq/). P values from Fisher's exact test were converted with to negative logarithm in base 10 by using R version 3.6.3 (https://www.r-project.org/). The latter values per variant were used to construct graph from Figures 3E and 3H, respectively by using GraphPad Prism 8 software. All computational steps required for these analyses are available here: https://github.com/cfarkas/SARS-CoV-2-freebayes (case example VI).

Phylogenetic Tree Construction

We downloaded 393 SAR2-CoV-2 GISAID genomes containing variant V1176F until September 25, 2020 from GISAID database (https://www.gisaid.org/). We filtered countries with at least two sequences per country, leading 358 sequences encompassing five country/regions (Brazil, Scottland, USA, Australia, and Gibraltar, respectively). MAFFT multiple sequence alignment program version 7.271 (36, 37) was used to align FASTA sequences against Wuhan-Wu-1 reference genome (GenBank Accession: MN908947.3) by using the --auto --thread -1 --keeplength –addfragments flags. Fasttree version 2.1 (38) was used to infer an approximately-maximum-likelihood phylogenetic tree from the aligned sequences in fasta format, by using heuristic neighbor-joining clustering method (38) and the Jukes-Cantor model of evolution (39). Visualization and editing of the phylogenetic tree was performed by using Interactive Tree of Life server (iTOL), collapsing all clades whose average branch length distance was below 0.0002 (40, 41).

Free energy estimation calculations

SARS-CoV-2 protein models for nsp2, nsp3, nsp4, 3C-Like proteinase, nsp6, nsp7, nsp8, RNA-dependent RNA polymerase, Helicase, Spike protein trimmer, ORF3a, ORF6, ORF8 and Nucleocapsid, respectively were accessed and downloaded from I-TASSER server

(https://zhanglab.ccmb.med.umich.edu/COVID-19/) on June 20, 2020, in Protein Data Bank (PDB) format. These models were generated by the C-I-TASSER pipeline (42-45). We calculated the free energy of Gibbs upon variant changes (ΔΔG; $\Delta G_{wild\text{-}type} - \Delta G_{variant}$, in kcal/mol) from all missense variants listed in Figure 3A using SARS-CoV-2 protein structures as inputs for the Foldx5 program (46, 47). We repaired every PDB by using the following command: foldx --command=RepairPDB --pdb=name-of-protein.pdb --ionStrength=0.05 --pH=7 --water=CRYSTAL --vdwDesign=2 --out-pdb=true --pdbHydrogens=false. Then, we modelled the variant and calculated the free energy upon aminoacid changes as follows: foldx --command=BuildModel --pdb=name-of-protein.pdb --mutant-file=individual_list.txt --ionStrength=0.05 --pH=7 --water=CRYSTAL --vdwDesign=2 --out-pdb=true --pdbHydrogens=false --numberOfRuns=30, where individual_list.txt contain the aminoacid change (as example, for a serine/asparagine change occurring in the three chains of the spike protein trimmer: SA477N,SB477N,SC477N;). We classified the energetic effects as follows (in kcal/mol): highly stabilizing (< −1.84), stabilizing (−1.84 to −0.92), slightly stabilizing (−0.92 to −0.46), neutral (−0.46 to +0.46), slightly destabilizing (+0.46 to +0.92), destabilizing (+0.92 to +1.84), and highly destabilizing (> +1.84) (48). Additionally, the effect of D614G/V1176F variants on protein–protein interaction energy in the full Spike protein trimmer and the effect of S477N variant in the Receptor Binding Domain (RBD) complexed with ACE2 dimer were assessed by submitting the referred structures on mCSM-PPI2 server (http://biosig.unimelb.edu.au/mcsm_ppi2/) (49).

Molecular dynamics simulations

We computed molecular dynamics simulations of the wild-type (Valine at position 1176) and 1176F (Phenylalanine at position 1176) stalk domain trimmers from the spike protein

(aminoacids 1130-1273). The full Spike protein trimmer was obtained from I-TASSER and the variant V1176F was modelled by using Foldx5, as previously described in the Free energy estimation calculations section (--command=BuildModel, first outputted model). The wild-type and F1176 variant trimmers were subjected to molecular dynamics by using GROMACS/2020.3 version, in gpu mode (http://manual.gromacs.org/documentation/) (50, 51) in the supercomputer infrastructure LEFTRARU NLHPC (ECM-02), allocating one node with total 44 cores (logical) and one compatible GPU (NVIDIA Tesla V100-PCIE-16GB). The trajectories were visualized by using VMD 1.9.3 (52). As example, for a given pdb (molecula_1.pdb), the commands used to perform the molecular dynamics are the following:

```
srun -p general gmx pdb2gmx -f molecula_1.pdb -o molecula_2.gro -water spce

srun -p general gmx editconf -f molecula_2.gro -o molecula_3.gro -c -d 1.0 -bt cubic

srun -p general gmx solvate -cp molecula_3.gro -cs spc216.gro -o molecula_4.gro -p topol.top

srun -p general gmx grompp -f ions.mdp -c molecula_4.gro -p topol.top -o ions.tpr

srun -p general gmx genion -s ions.tpr -o molecula_5.gro -p topol.top -pname NA -nname CL -
neutral

gmx grompp -f 1.mdp -c molecula_5.gro -p topol.top -o em.tpr

gmx mdrun -nt 20 -nb gpu -deffnm em          #EM

gmx grompp -f 2.mdp -c em.gro -r em.gro -p topol.top -o nvt.tpr

gmx mdrun -nt 20 -nb gpu -deffnm nvt         # NPT

gmx grompp -f 3.mdp -c nvt.gro -r nvt.gro -t nvt.cpt -p topol.top -o npt.tpr
```

```
gmx mdrun -nt 20 -nb gpu -deffnm npt        # NVT
gmx grompp -f 4.mdp -c npt.gro -t npt.cpt -p topol.top -o md_0_1.tpr
gmx mdrun -nt 20 -nb gpu -deffnm md_0_1  # MD
```

Where commands starting with "srun" were executed directly in the cluster and the remaining steps were submitted via SLURM workload manager (https://slurm.schedmd.com/documentation.html). We choose forcefield OPLS-AA/L all-atom force field (2001 aminoacid dihedrals) for step one, and SOLVENT for step five (choice 13, SOL). Atom clashed in the system were minimized by the steepest descent method (53), until potential energy were below 1000 kJ/(mol*nm). We considered a cutoff of 1.0 nm for non-bonded interactions under periodic boundary conditions (PBC). NVT ensemble (constant Number of particles, Volume, and Temperature) was performed setting no pressure coupling and modified Berendsen thermostat at 300K, respectively. The NPT ensemble was used to keep the constant pressure at 1 bar, using the Parrinello-Rahman barostat and temperature at 300 K, using the modified Berendsen thermostat, respectively. Long-range electrostatic forces were considered using the Particle Mesh Ewald for long-range electrostatics method (54). Both equilibrations were performed for 5000 picoseconds (5 nanoseconds). The total energy, temperature, pressure and the of the stalk domain trimmers were used to corroborate both system equilibrations. After these steps, production dynamics were carried out for 20 nanoseconds, by using the leap-frog algorithm with an integration step of 2 femtoseconds, as motion setting. Bonds were fixed using the P-LINCS method, with constrained H-bonds (55, 56). Root mean square deviation (RMSD) and radius of gyration (Rg) were obtained with the following commands, respectively:

```
gmx rms -s md.tpr -f md_0_1.xtc -o rmsd.xvg -tu ns    # Choose twice 4 ("Backbone")

gmx gyrate -s md.tpr -f md_0_1.xtc -o gyrate.xvg       # Group 1 (Protein)
```

The xvg file records per picosecond were used to plot graphs from Figure 3C, on GraphPad Prism 8 software. PDB, solvated molecules (.gro) and correspondent compressed gromacs trajectories (with or without periodic border conditions) are available here : https://usegalaxy.org/u/carlosfarkas/h/sars-cov-2-proteins-and-trayectories.

Protein Visualization and Rendering

The Spike protein trimmer image related to Figure 3C was rendered online with the EzMol server (http://www.sbg.bio.ic.ac.uk/ezmol/) (57).

Statistical analysis

All statistical analyses were carried out by using GraphPad Prism 8 software (https://www.graphpad.com/scientific-software/prism/). Mann-Whitney test was employed to test data following non normal distribution. The significance thresholds were the following: P<0.05 *, P<0.01 **, P<0.001 ***, P>0.05 ns. We interpreted Spearman nonparametric correlations analyses as follows: perfect correlation (1), the two variables tend to increase or decrease together (0 to 1), The two variables do not vary together at all (0), One variable increases as the other decreases (-1 to 0), and perfect inverse correlations (-1). We computed an approximate P value because in all correlations we employed more than 17 pair values. The significance thresholds were the following: P<0.05 *, P<0.01 **, P<0.001 ***, P<0.0001 ****, P>0.05 ns. We employed robust regression and outlier removal (ROUT) method (58) to remove outliers from stacks of data, with a strict false discovery ratio (Q=1%). SNP-mortality correlations were assessed by using the snpFreq program

(https://rdrr.io/github/lvclark/SNPfreq/), implemented in the R language (https://www.r-project.org/). We employed the Fisher's exact test to identified variants with a significant difference in the viral frequencies between deceased-released groups. We used as a significance threshold a false discovery rate (q-value) of 0.005.

Results

Worldwide Intra-host variation in SARS-CoV-2 genomes reveal microdiversity and hypermutation likely elicited by APOBEC3G/ADAR complexes.

To trace intra-host viral variation worldwide, we downloaded and analyzed 76,553 SARS-CoV-2 genome sequences available in the GISAID database up until August 3, 2020 (**Supplementary Table 1,** see Acknowledgements in **Supplementary Table 2**). We also downloaded and analyzed 17,560 next-generation sequencing datasets from Sequence Read Archive (SRA) available until July 28, 2020 (**Supplementary Table 3**). SARS-CoV-2 genomes from GISAID accounted for the presence of major viral frequency variants (via a consensus calling approach) and the next-generation sequencing datasets (NGS) also allowed us to analyze intra-host microdiversity given the depth of sequencing. By analyzing the occurrence of major viral alleles per SARS-CoV-2 genome, both sources consistently demonstrate on average 7-8 viral variants with major alleles per genome (viral frequency > 0.5) (see "mean" in **Figure 1A** and **1B**, for SRA and GISAID datasets, respectively), demonstrating that our variant calling pipeline is reliable to call major viral alleles from FASTA and NGS datasets, respectively (see https://github.com/cfarkas/SARS-CoV-2-freebayes). The distribution from both sources also identified outliers with more than 16 viral sequence variants per genome including some samples harboring more than 100 variants per genome, greatly surpassing the average (2% and 0.3% in SRA and GIDAID sequencing datasets, see "outliers" in **Figure 1A and 1B**, respectively, Q=1% Grubbs's test). Integrative genomics viewer (IGV) snapshots of outlier samples from Spain, USA and Australian sequencing datasets clearly show hypermutability to varying degrees (see samples with black arrows, **Figure 1C**). Australian outlier samples represent an extreme case of

hypermutability (see **Figure 1C**, bottom). 16,307 aggregated variants from SRA datasets reflect most recurrent single nucleotide substitutions occurring in all genomes from SRA repository are enriched by the C>U (C>T) transitions and G>T (G>U) transversions, changes already reported for SARS-CoV-2 and MERS-CoV genomes (59) and is likely elicited by APOBEC deaminases, as already reported (60). (**Figure 1D**, left). The latter observation is also consistent for genomes containing outlier number of variants from SRA (**Figure 1D**, right). Most of the nucleotide substitutions harboring outlier samples from Figure 1C correspond to missense/nonsense variants rather than silent variants (**Figure 1E**) and are enriched in C>T (C>U) changes as well, consistent with latter observations (**Figure 1F**). Also, G>T (G>U) transversions are significantly present, implying a different mechanism than APOBEC3G editing and likely exerted by ADAR deaminase (see discussion). We estimated intra-host nucleotide diversity occurring in 397, 374 and 215 next generation sequencing samples from Australia, Spain and USA populations, respectively by using aligned reads per sample against the SARS-CoV-2 reference genome (Wuhan-Hu-1, GenBank Accession: MN908947.3). This calculation has been already validated to capture intra-host viral microdiversity, overcoming sequencing errors (61). In the three populations, average nucleotide diversity positively correlates with the number of Single Nucleotide Variants (SNVs) with viral frequencies over 5% (Spearman correlation, r values from 0.32-0.66, $P<0.0001$). The latter supports the existence of intra-host minor variants and therefore SARS-CoV-2 quasi-species, coexisting within the same host (62-65) (**Figure 1G**). We hypothesized the existence of hypermutants can be explained by APOBEC3G/ADAR-mediating RNA editing at early stages of SARS-CoV-2 infection (C>U and A>G editing) accompanied by intra-host diversity, i.e. low frequency variance and homoplasy (different viral lineages emerged after the infection) that probably are maintained at low frequency due the virus error correction

machinery (**Figure 1H**). Overall, we propose that human hosts substantially contribute to shape SARS-CoV-2 genetic diversity. Although microdiversity is probably one of the main sources of SARS-CoV-2 evolution, this is accompanied by RNA-editing at different levels, with SARS-CoV-2 RNA hypermutation as an extreme case of the latter.

Non-neutral codon changes actively shape evolution of several SARS-CoV-2 proteins

We next analyzed all intra-host major viral alleles occurring in SARS-CoV-2 genomes worldwide, by using the GISAID consensus called variants. 23,269 aggregated variants from GISAID demonstrate that overall, C>U and A>G edition are the predominant nucleotide changes (**Figure 2A**), also present in GISAID samples with outlier number of variants per genome (**Figure 2B**). These changes are consistent with nucleotide changes occurring in SRA sequencing datasets (see **Figure 1D**) and with the APOBEC3G-mediating RNA editing mechanisms. To deduce aminoacid changes as consequences of these nucleotide changes, we analyzed codon changes occurring in the aggregated GISAID variants and we predicted its consequences by using the SnpEff program (30). Occurrences per variant type demonstrate that missense and synonymous variant occurrences are more frequent compared to frameshift/nonsense variant occurrences per genome (**Figure 2C**). Overall, missense/silent ratio of GISAID aggregated variants is 1.82, revealing a greater diversity in missense variants as well (**Supplementary Figure 1**). Consequently, codon change analysis demonstrates frequent non-neutral changes in the second position of the codons ACA>ATA, ACT>ATT and GCT>GTT and leads to missense variants Thr>Ile and Ala>Val, respectively. Also, non-neutral changes in the first position of codons CTT>TTT and GTT>TTT are frequent, leading to Leu>Phe and Val>Phe changes, respectively (**Figure 2D**). Almost all these changes can be explained by C>T (C>U) transition, already reported for SARS-CoV-2 protein changes. To understand these changes at the global

population level, we calculated nucleotide diversity in nonsynonymous (πN) and synonymous (πS) sites of every SARS-CoV-2 protein across six different populations (Asia, Oceania, Europe, Africa, North America, and South America) by using GISAID genomes per geographical region. A given protein is evolving by natural selection if the πN/πS ratio is over one. Conversely, if πN/πS ratio is less than one, a given protein is considered to be undergoing purifying selection, as previously described (66-68). An excess of shared nonsynonymous changes (implying positive natural selection) is present in non-structural proteins nsp2, nsp7, 3C-like proteinase, ORF3a and ORF8, respectively (**Figure 2E**). Conversely, purifying selection is present in the RNA-dependent RNA polymerase, Membrane protein (M), 3'-to 5'-exonuclease, non-structural proteins nsp3 and nsp4, respectively. Remarkably, ORF6 and to a lesser extent the Spike protein (S) of SARS-CoV-2 are actively evolving in South America.

V1176F variant occurring in the Spike protein is predicted to improve fitness of Spike protein complex and is likely a novel SARS-CoV-2 mutation associated with increased mortality

We analyzed SARS-CoV-2 fixed alleles shared with at least a 3% variant frequency (VF) in one of the five referred to continental populations until August 3, 2020. The merge encompassed 51 variants. Consistent with intra-host variation, more than half of these changes lead to missense variants. Non-structural proteins nsp2, nsp3, nsp6 and nsp7 harbors Leu>Phe and Thr>Ile as frequent changes as seen in SARS-CoV-2 codon changes frequencies (**Figure 3A**). Also, consistent with the increased observed πN/πS ratios per population, ORF3a and ORF8 display missense variants with high viral frequencies. Some variants are shared among populations with high viral frequencies, such as D614G variant in Spike protein, P323L variant in RNA-dependent RNA polymerase and RG203KR variant in Nucleocapsid. Of these, D614G

variant in the Spike protein fulfill the category of mutation since this variant is positively associated with mortality and increases SARS-CoV-2 infectivity, as previously reported (18, 19, 69). As D614G was a rapidly emergent variant, this suggests novel emergent variants can also be mutations. Of importance, two more novel missense variants occurring in the Spike protein (S477N and V1176F) are exclusively from Oceania and South America (**Figure 3A,** red variants). To gain understanding of the effects of these variants, we calculated the Gibbs' free energy associated with the variant changes ($\Delta\Delta G$; $\Delta G_{wild\text{-}type} - \Delta G_{variant}$) from all missense variants listed in Figure 3A using SARS-CoV-2 protein structures generated by the C-I-TASSER pipeline as inputs for the Foldx5 program (42-47). While most of the changes are predicted to be neutral or slightly stabilizing/destabilizing, there are more changes predicted to be energetically unfavorable rather than favorable, such as occurring in the RNA-polymerase (A97V), ORF3a (G251V) and Nucleocapsid (I292T). Conversely, other changes occurring in the non-structural proteins nsp3 (T1198K), nsp4 (F308Y) and Spike protein (D614G, V1176F) are predicted to stabilize these proteins (**Figure 3B**). The emergent V1176F variant, containing the recurrent signature Val>Phe (**Figure 2D**) is located at the stalk domain of Spike protein, specifically at the beginning of the heptad repeat 2 (HR2) domain (70). In agreement with a recent report, the D614G variant has a mildly stabilizing effect on protein stability but also alters protein dynamics according to mCSM-PPI2 analysis, since the predicted affinity change of the spike protein trimmer decreases (71). Conversely, the V1176F variant is predicted to increase affinity of the Spike protein trimmer (**Figure 3C, upper right**). Molecular dynamics simulations of the stalk domain trimmer demonstrate larger amplitude motions, since Root-mean-square deviations (RMSD) from wild-type stalk domain trimmer fluctuates in ~1 nm (10 Å) over 20 nanoseconds of simulation. The V1176F variant increases this value after the same settings, increasing

motility of the domain trimmer (~1.4 nm, **Figure 3C middle right**). Also, the V1176F induces compactness of Stalk domain trimmer in ~0.52 nm (5 Å) (**Figure 3C low right**) (72). Thus, the V1176F variant confers more flexibility to the Stalk trimmer domain. These observations agree with a recent report demonstrating extensive flexibility of this domain composed by three hinges in the pre-fusion model of the Spike protein, potentially necessary for enhanced binding mechanics with ligands such ACE2 (73). Of note, up until September 25, 2020 Brazil was the country with the highest frequency of genomes containing the V1176F variant, that was completely linked to the D614G mutation (**Figure 3D**). We also found four additional regions that present increasing cumulative distributions of genomes containing V1176F, including Scotland, USA, Australia, and Gibraltar that have occurred during global travel bans/restrictions (**Figure 3E**). Indeed, phylogenetic tree analysis containing genomes from Figure 3E demonstrate V1176F variant arose independently in clades G (containing S: D614G mutation), clade GH (containing ORF3a: Q57H variant) and clade GR (containing N: RG203KR variant) (**Figure 3F**). The predominant clade GR contain genomes from Brazil, Scotland, and Gibraltar while clade GH is represented exclusively of genomes from USA. Clade G is also present in Brazil and Scotland. Overall, this analysis suggests community transmission spread of V1176F occurred from independent sources rather than from a single source arising from travel. To correlate if this variant has an associated phenotype, we downloaded 7634 genomes with associated metadata from GISAID until September 28, 2020 and we grouped the genomes from released/deceased patients per country/region. Then, we performed Fisher's exact test and identified variants with a significant difference in the viral frequencies between the groups (snpFreq program, available in the Galaxy server) (31, 35). Interestingly, a novel variant in the spike protein (QD613HG), leading to D614G mutation and variants occurring in the ORF3a (Q57H) and Nucleocapsid

(RG203KR), are positively correlated with increased mortality ratios in Saudi-Arabia (**Figure 3H**). In Brazil, the D614G mutation along with the V1176F emergent variant in the Spike protein are also positively correlated with increased mortality ratios, including variants occurring in the nsp7, RNA-dependent RNA polymerase, and two emergent variants in South America occurring in the ORF6 and the nucleocapsid proteins (I33T and I292T variants, respectively, **Figure 3A, Figure 3G**). The RG203KR variants occurring in the Nucleocapsid was also found in Brazil. Four out seven of these variants occurring in Brazil (nsp7: L71F, S: V1176F, ORF6: I33T and N: I292T) are also positively correlated with increased mortality when comparing deceased patients versus Released + Hospitalized patients, suggesting these variants are robustly correlated with increased mortality in Brazil (**Figure 3H**). Of note, the V1176F variant arose independently and is not ligated to the I292T variant occurring in the Nucleocapsid **(Supplementary Table 4**). Thus, among emergent SARS-CoV-2 variants in Brazil, we prioritize the V1176F variant for further experimental study since it likely arose independently across SARS-CoV-2 clades in different countries, is predicted to improve fitness of the Spike protein and correlates with increased mortality ratios in Brazil.

Variant S447N occurring in the Spike protein is a novel mutation that increases Spike-ACE2 binding and is associated with higher worldwide fatality rates.

We next addressed whether the referred variants are associated with higher mortality ratios in Brazil and Saudi-Arabia, as well as other emergent variants in the Spike protein, are correlated with higher fatality ratios worldwide until September 28, 2020. Among the refereed variants, the I292T variant occurring in the nucleocapsid is associated with higher fatality rates across several countries ($p<0.033$, Spearman correlation, **Figure 4A, Supplementary Table 5**). Eleven variants occurring in the Spike protein were present in more than four countries and out

of them, variants A222V, S477N and E780Q are positively correlated with increased fatality ratios. These findings are consistent with previous reports for the D614G mutation that is also positively correlated with higher fatality ratios (**Figure 4B**) (19). The V1176F variant was not found to be correlated with increased fatality ratios across the world, probably due to highly unbiased viral frequencies across the world. Up until August 3, 2020, the S477N spike variant that emerged in Oceania (Australia) had a low viral population frequency (~ 3.5%, **Figure 3A**) along with the A222V variant being absent or maintained at a very low frequency. However, a dramatic increase of cumulative genomes was observed between the months June and July 2020 for both variants, not seen for the I292T variant occurring in the nucleocapsid (**Figure 4C**). The rapid emergence of the S477N and A222V variants correlates with the recent second wave of COVID19 that has occurred in Australia and the United Kingdom (**Figure 4D,** left and right, respectively). Nevertheless, this rapid increase in the population frequencies of these variants could be due to the founder effect usually seen in outbreaks, and not due increased fitness of SARS-CoV-2 (see countries in Figure 4B, A222V and S477N, respectively) (69). To discard the former founder effect, we examined GISAID clade frequency of both variants, demonstrating variant A222V arose independently in all major SARS-CoV-2 clades (74) and variant S477N arose in clades G, GH and GR, respectively (**Figure 4E**, upper and lower, respectively). Variant A222V occurred in the N-terminal domain of the Spike protein (light brown domain, **Figure 4F**, left) and variant S477N occur in the Receptor Binding Domain (RBD) of the Spike protein (purple domain, **Figure 4F**, left). This variant is located near the interface between ACE2 and the RBD, the latter is expected to cause enhanced binding of the RBD to the ACE2 human receptor (**Figure 4F**, right). We replaced serine for asparagine in position 477 in the RBD (complexed with ACE2 dimer) with foldx (46), and we calculated the predicted binding energy

upon this change by using mCSM-PPI2 server (49). The change is predicted to add one more polar interaction when asparagine is present in the RBD, increasing the affinity between RBD and ACE2 (**Figure 4G**). We examined the deep mutational scanning of amino acid changes in the RBD performed by Starr et al (75) in a high throughput yeast-surface-display system for measuring expression of folded RBD protein and its binding to ACE2 (76). Among fourteen variants occurring in the RBD, only variant S477N increased both expression of the RBD, a parameter positively correlated with folding (**Figure 4H**) and its binding to ACE2, respectively **(Figure 4I**). The combination of these two properties can lead to the generation of a more infectious viruse, explaining to a large extent the dramatic increase of the S477N variant in Australia.

In conclusion, we have obtained evidence that the S477N variant is a gain of function mutation occurring in the Spike protein, that is positively correlated with increased fatality rates and is becoming dominant as its increases, as was also recently observed for the D614G mutation that became dominant throughout the world (18).

Discussion

In this study we aimed to analyze over 94,000 SARS-CoV-2 sequences deposited between GISAID and SRA databases within the first eight months of this pandemic (up until the beginning of August 2020). We characterized the existence of intra-host viral hypermutation that results in an excessive number of variants per genome in less than 2% of SARS-CoV-2 sequences (**Figure 1A** and **1B**, respectively). This phenomenon was already described for HIV-1 virus *in vivo*, demonstrating that HIV-1 reverse transcriptase contributed only to 2% of mutations, and the majority was caused by host cytidine deaminases of the A3 family mediated editing (77). Here, we present evidence that enzymatic RNA editing in combination with microdiversity contributes to SARS-CoV-2 diversity at a global level, leading to more than 23,000 major viral frequency variants within 76,000 GISAID genomes. In SARS-CoV-2 genomes, the C>T (C>U) transversion is substantially present both in hypermutants and non-hypermutant samples, suggesting APOBEC3G mediated RNA editing involvement, as previously reported in smaller sample sizes (59, 60). Also, the A>G transversion is also present overall but not in hypermutant genomes (Figures **1D**, **2A** and **2B**, respectively), implying an active role of ADAR deaminases during SARS-CoV-2 infection (78). We argue ADAR-mediating RNA editing is not the main enzyme involved in the hypermutation mechanism, but rather APOBEC3G deaminase complexes. Also, we have observed substantial G>T transversions in SARS-CoV-2 genomes. This transversion has been already reported for other RNA viruses such as Maize streak virus (79) and is has been linked with the formation of 8-oxoguanine, known to be the most common cause of spontaneous G>T (G>U) transversions in RNA (80). Recently, it has been reported that tissue damage from neutrophils induces oxidative stress upon SARS-CoV-2 infection (81), implying that reactive oxidative species (ROS) mediated mutagenesis is likely the mechanism that cause this transversion. The hypermutated SARS-CoV-

2 variant signature often contains nonsense variants that are predicted to inactivate several SARS-CoV-2 proteins, probably leading to an efficient mechanism of lethal mutagenesis to control viral spread (**Figure 1E**). In addition, other signatures are present at the intra-host level, implying microdiversity as another potential source of this variation (**Figure 2A** and **2B**, respectively). It is possible that these combined forces produce quasi-species of viruses with enough sequence diversity that may influence viral pathogenesis and drive SARS-CoV-2 evolution, sometimes also leading to viral extinction (82, 83). Overall, we propose that human host are major drivers of SARS-CoV-2 diversity rather than the virus itself, evidenced by the levels of intra-host variation and hypermutation at different degrees, both fueled by enzymatic RNA-editing mechanisms. We support the latter with the observed non-random signatures of nucleotide changes in these mechanisms, and the presence of SARS-CoV-2 error-correction machinery, not seen in other RNA viruses. Although we found a significant amount of intra-host variation in SARS-CoV-2, neutral evolutionary theory predicts most of these variants as having no or neutral effects (84). Nevertheless, positive, or negative selection can occur within-populations on viral variants, by conferring advantageous properties to viruses that ultimate lead to mutations. Here we have demonstrated positive natural selection of several SARS-CoV-2 proteins per population, occurring in nsp2, nsp7, 3C-like proteinase, ORF3a and ORF8 proteins and we highlighted mutations in the Spike protein, evolving in South America (Brazil, V1176F) and Oceania (Australia, S477N). To begin to understand how these variations may affect their encoded proteins, we assessed their structural consequences in protein models that demonstrate that these variants tend to cause more unfavorable than favorable energetic changes. This phenomenon has been observed in the spectrum of variants occurring in the RBD of the Spike protein, proving that most of the variants occurring in the RBD constraints its folding and

binding to ACE2 (75). In the case of SARS-CoV-2, immunological pressure from host could lead to this type of phenomena as well and is consistent with the observed number of intra-host missense over synonymous variants, respectively. Conversely, we found missense variant V1176F, occurring in the Spike protein, is predicted to be energetically favorable and confers flexibility to the Stalk domain of the viral Spike protein trimmer, previously described to be important for Spike protein flexibility and binding to ACE2 (73). Phylogenetic analysis demonstrated V1176F variant likely emerged in South America and arose independently in several countries associated with the D614G mutation, suggesting that this variant is being positively selected among others occurring in the Spike protein (**Figure 3F**). This variant is also correlated with higher mortality ratios (**Figure 3G** and **3H**, respectively) and it is possible that it increases the fitness of SARS-CoV-2 infection by conferring flexibility to the stalk domain of the spike protein. The same conclusions applied to variant S477N, occurring in the RBD of the Spike protein: is energetically favorable in RBD-ACE2 binding and favor expression of the RBD, with the latter conclusions being experimentally supported. This mutation also arose independently in a noticeably short period of time and become dominant in Australia within two months (**Figure 4D**). In addition, the S477N variant is constantly spreading across European countries (**Figure 4B**) and correlates with higher mortality ratios. These observations provide strong evidence that the S477N variant is a novel gain of function Spike protein mutation, as has recently been demonstrated for the D614G mutation (18). We argue that the constant spread of V1176F and S477N variants over the world ultimately may lead to a further significant concern in public health, due to their association with higher mortality rates.

A remaining question is the association of higher fatality rates of the A222V variant occurring in the N-terminal domain of the Spike protein, and the I292T variant occurring in the

Nucleocapsid, among others. It is possible that these variants can confer antigenic escape, since recently, it has been registered that a reinfection case containing A222V and D614G mutations has occurred (85). Nucleocapsid variation has also been documented in the nucleoprotein of the influenza virus (86) (87) and nucleocapsid of the hepatitis virus (88). Both RNA viruses escape cellular immunity by these mechanisms and could also be the case for SARS-CoV-2.

In summary, we have presented potential molecular mechanisms that help researchers to understand variation diversity fueled natural selection in SARS-CoV-2. It is important to continue to track emergent viral variants with the bioinformatics tools developed and highlighted in this manuscript since the evidence presented here lead us to propose V1776F and S477N variants are novel SARS-CoV-2 mutations, due to their positive correlations with increased fatality ratios, as previously evidenced with D614G mutation occurring in the Spike protein. Further conclusions concerning the effects of these variants on viral fitness and host mortality will be accomplished by future structure-function based studies using viral Spike protein mutants and studying effects on viral entry and *in vivo* rodent models expressing the human ACE2 receptor.

Acknowledgments

Powered@NLHPC: This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02). This research was partially funded by research funding from the CIHR, Research Manitoba and the CancerCare MB Research Foundation.

Author Contributions

CF conceived of this study and performed all bioinformatics analysis and wrote the manuscript. AM performed mutant Spike protein analysis and assist in biophysical studies. JH assisted study design, data interpretation and manuscript writing.

Declaration of Interests

None declared

Figure titles and legends

Figure 1: Intra-host variation in SARS-CoV-2 genomes worldwide reveal microdiversity and hypermutation likely elicited by APOBEC3G complexes.

A) Mayor viral frequency variants (via a consensus calling approach) for 75,563 SARS-CoV-2 GISAID genomes, separated by non-outliers (n=76,310) and outliers (n=243, Q=1%, Grubbs's test). Mean and outlier number of variants are depicted at left. **B**) Same as A for next generation sequencing (NGS) datasets downloaded from SRA (n=17,500). **C**) IGV snapshots of outliers and non-outlier NGS samples from B. Outliers samples are depicted with black arrows, exceeding number of variants from non-outliers. Single nucleotide polymorphisms are depicted in red if nucleotide differs from the reference sequence in greater than 50% of quality weighted reads. **D**) Nucleotide change frequencies from 17,560 SRA NGS aggregated variants (left) and from 360 aggregated outlier variants (right), both annotated with SnpEff program. Frequency boxes are colored from white to dark red as number of changes increases. **E**) Nucleotide substitution frequencies in each of the eight outlier samples indicated with black arrows in C) grouped by silent (green), missense (dark red) and nonsense (red). **F**) Transitions and transversions occurring in the latter samples, denoted with blue and red, respectively. Significance of comparisons were assessed with Mann-Whitney test (P<0.05 *, P<0.01 **, P<0.001 ***, P>0.05 ns). **G**) Correlation between Average nucleotide diversity (π) provided by inStrain program and SNV counts (VF>5%) for Spain (n=374, left), USA (n=215, middle) and Australian NGS samples (n=397, right). In the three countries, the two variables tend to increase together (see r values). The significance thresholds were the following: P<0.05 *, P<0.01 **, P<0.001 ***, P<0.0001

****, P>0.05 ns. **H**) Proposed model of how APOBEC3G/ADAR complex can lead to hypermutation of SARS-CoV-2 (C>U and A>G editing). The RNA editing can be accompanied by intra-host diversity (low frequency variants) and homoplasy (different viral lineages emerged after the infection), maintained at low frequency due the virus error correction machinery.

Figure 2: Non-neutral codon changes actively shape evolution of several SARS-CoV-2 proteins.

A) Nucleotide change frequencies from 76,553 aggregated GISAD genome variants annotated with SnpEff program. Frequency boxes are colored from white to dark red as number of changes increases. **B**) Same as A for 243 aggregated GISAID genomes, corresponding to GISAID outlier samples. Frequencies boxes are colored from white to dark red, as the number of changes increases. **C**) Missense, nonsense, frameshift, and synonymous number of occurrences in 76553 GISAID genomes. Significance of comparisons were assessed with Mann-Whitney test ($P<0.05$ *, $P<0.01$ **, $P<0.001$ ***, $P>0.05$ ns). **D**) Plot of change frequencies across 32 codons in SARS-CoV-2. Changes are grouped in six categories and colored from light to dark red, according to the number of changes. **E**) Nonsynonymous/synonymous nucleotide diversity calculated by SNPGenie program for each SARS-CoV-2 protein across six geographical regions (Asia, Oceania, Europe, Africa, North America, and South America, respectively). Ratios are grouped in seven categories and colored from white to dark red, according to the ratio numbers.

Figure 3: V1176F variant occurring in the Spike protein is predicted to improve fitness of Spike protein complex and is likely a novel SARS-CoV-2 mutation.

A) Heatmap of viral frequencies from 51 variants shared at least 5% frequency within populations. Variants position in SARS-CoV-2 reference genome (Wuhan-Hu-1, GenBank Accession: MN908947.3) are depicted at left. At right, amino acid changes are colored by protein. **B**) Gibbs Free Energy Calculation (ΔΔG) of missense variants depicted in A, obtained by Foldx5 program. ΔΔG (kcal/mol) was colored according to the energetic effect in SARS-CoV-2 structures depicted in the Y axis, as follows (in kcal/mol): highly stabilizing (< −1.84), stabilizing (−1.84 to −0.92), slightly stabilizing (−0.92 to −0.46), neutral (−0.46 to +0.46), slightly destabilizing (+0.46 to +0.92), destabilizing (+0.92 to +1.84), and highly destabilizing (> +1.84). **C**) (left) Rendered Spike protein trimmer structure with EzMol server using rainbow palette. D614G mutation and V1176F variant are depicted in red, including their sidechains. Volume is depicted in grey. The stalk domain of the Spike protein trimmer is depicted in dark blue. (Upper right) Predicted affinity change (kcal/mol) of Spike trimmer by mCSM-PPI2 server upon D614G and V1176F amino acid Spike protein changes. (middle right) RMSD values (in nanometers, nm) of 20 nanoseconds of simulation of the wild-type stalk domain trimmer (red) or the domain containing Phenylalanine in position 1176 (blue). (lower right) Radius of gyration values (in nanometers, nm) of the latter simulation. **D**) Occurrences per genome of the D614G mutation and V1176F variant alone or in combination in 974 Brazilian GISAID genomes until September 25, 2020. Complete linkage of V1176F variant with D614G mutation was found since non a single genome contains V1176F variant alone. **E**) Cumulative distributions of 358 genomes containing V1176F variant (G25088T) from January 2020 until September 28, 2020 in five geographical regions, respectively. **F**) Phylogenetic tree of the 358 genomes from E,

constructed by Neighbor-Joining method, and visualized with the iTOL server. The genomes were colored by SARS-CoV-2 clades as follows: G (containing S: D614G mutation, in green), GH (containing ORF3a: Q57H variant, in red) and GR (containing N: RG203KR variant, in blue). Brazilian genomes belonging clade GR were depicted with small blue font and the other genomes were highlighted with higher font, for visualization purposes. **G**) Mortality correlations associated with SARS-CoV-2 variants in Released vs Deceased patients occurring in India, USA, Saudi Arabia, and Brazil until September 25, 2020, respectively. The corrected p-values from fisher exact test (q-values) were obtained from the snpFreq program, available in the Galaxy server and plotted as the negative logarithm in base 10 of each q-value (significance: q-value>0.005). Significant variants were depicted at the right of each bar. **H**) Same as G, but with the comparison Released + Hospitalized vs Deceased patients in Brazil.

Figure 4: Variant S447N occurring in the Spike protein is a novel mutation that increase Spike-ACE2 binding and is associated with worldwide higher fatality rates.

A) Correlation analysis of COVID-19 fatality rates among 49 countries against variant frequencies occurring in various SARS-CoV-2 proteins: ORF3a (Q57H), , RNA-dependent RNA polymerase (P323L), nsp7 (L71F), N (RG203KR) and N (I292T), respectively. Spearman's correlation coefficients (r) were calculated, including confident p-values (P) to discard random sampling. The significance thresholds were the following: P<0.05 *, P<0.01 **, P<0.001 ***, P<0.0001 ****, P>0.05 ns. China were discarded from these analyses due its large population size and consistently low number of COVID19 cases since early April 2020. Fatality rate data was obtained from John Hopkins coronavirus resource center (https://coronavirus.jhu.edu/map.html), accessed on September 28, 2020. **B**) Same analysis as A, for five emergent variants occurring in the Receptor Binding Domain (RBD) portion of the Spike protein (N439K-L455F), including variants L18F, A222V, D614G, E780Q and V1176F occurring in the Spike protein. **C**) Cumulative distributions over time of genomes containing variants I292T in the nucleocapsid and emergent variants A222V, S477N and E780Q in the spike protein, respectively. Cumulated genome numbers were plotted at the end of the indicated months. **D**) Frequency plots over time of cumulated genomes containing spike protein emerging variants A222V, S477N occurring in Australia (left) and United Kingdom (right), respectively. Cumulated frequencies were plotted at the end of the indicated months. **E**) Graph chart indicating the clade frequencies of variants A222V (Upper) and S477N (lower), reported from GIDAID database, until September 28, 2020. **F**) (Left) Spike protein Trimmer indicating positions of variants N477 and V222 (red letters). Variant N477 locates in the Receptor Binding Domain (RBD, residues 331-530, depicted in purple) and Variant V222 locates in the N-terminal

domain (residues 1-316, depicted in light brown). The Spike structure was shown as cartoon with grey color, obtained with EzMol server. (Right) Magnification of the interaction between the RBD (purple) with ACE2 dimer (blue and red structures, respectively) showing the contact surfaces. Variant N477 and its surface is depicted in red. The image was rendered in the EzMol server. **G**) Polar interactions of S477 (left, wild type) and N477 (right, mutant) depicted by discontinued lines in the RBD. The polar amino acid interactions of S477 and N477, including the predicted ΔΔG (affinity, kcal/mol) between the wild-type or mutant RBD with ACE2 dimer were obtained by using the mCSM-PPI2 server. Positive values of ΔΔG indicate increasing affinity between the RBD and ACE2 dimer. **H**) Expression fitness of fourteen emergent variants occurring in the RBD domain of the spike protein (see Supplementary Figure 2), measured by high throughput yeast-surface-display system. Expression measurements were plotted as the difference in log-mean fluorescence intensity (MFI) relative to wild-type (ΔlogMFI = $\log MFI_{variant}$ - $\log MFI_{wild\text{-}type}$). Positive values (blue) indicate higher RBD expression, therefore higher folding fitness. Expression data was obtained from here: https://jbloomlab.github.io/SARS-CoV-2-RBD_DMS/. I) ACE2 binding fitness of the referred RBD variants, measured by high throughput yeast-surface-display system. Binding measurements were plotted as the difference in log10(KD, apparent) relative to wild-type (Δlog10(KD, apparent) = $\log10(KD, apparent)_{wild\text{-}type} - \log10(KD,app)_{variant}$). Positive values (blue) indicate higher affinity between RBD and ACE2. Binding data was obtained from here: https://jbloomlab.github.io/SARS-CoV-2-RBD_DMS/.

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time (vol 20, pg 533, 2020). Lancet Infectious Diseases. 2020;20(9):E215-E.
2. Ke Z, Oton J, Qu K, Cortese M, Zila V, McKeane L, et al. Structures and distributions of SARS-CoV-2 spike proteins on intact virions. Nature. 2020.
3. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019;17(3):181-92.
4. Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, et al. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell Host Microbe. 2020;27(3):325-8.
5. Hoffmann M, Kleine-Weber H, Schroeder S, Kruger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020;181(2):271-80 e8.
6. Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, Berne MA, et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature. 2003;426(6965):450-4.
7. Crackower MA, Sarao R, Oudit GY, Yagil C, Kozieradzki I, Scanga SE, et al. Angiotensin-converting enzyme 2 is an essential regulator of heart function. Nature. 2002;417(6891):822-8.
8. Ge XY, Li JL, Yang XL, Chmura AA, Zhu G, Epstein JH, et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature. 2013;503(7477):535-8.
9. Williamson BN, Feldmann F, Schwarz B, Meade-White K, Porter DP, Schulz J, et al. Clinical benefit of remdesivir in rhesus macaques infected with SARS-CoV-2. Nature. 2020;585(7824):273-6.
10. Lei C, Qian K, Li T, Zhang S, Fu W, Ding M, et al. Neutralization of SARS-CoV-2 spike pseudotyped virus by recombinant ACE2-Ig. Nat Commun. 2020;11(1):2070.
11. Monteil V, Kwon H, Prado P, Hagelkruys A, Wimmer RA, Stahl M, et al. Inhibition of SARS-CoV-2 Infections in Engineered Human Tissues Using Clinical-Grade Soluble Human ACE2. Cell. 2020;181(4):905-13 e7.
12. Gudbjartsson DF, Norddahl GL, Melsted P, Gunnarsdottir K, Holm H, Eythorsson E, et al. Humoral Immune Response to SARS-CoV-2 in Iceland. N Engl J Med. 2020.
13. Pollan M, Perez-Gomez B, Pastor-Barriuso R, Oteo J, Hernan MA, Perez-Olmeda M, et al. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. Lancet. 2020;396(10250):535-44.
14. Li W, Zhang B, Lu J, Liu S, Chang Z, Cao P, et al. The characteristics of household transmission of COVID-19. Clin Infect Dis. 2020.
15. Covid-19 National Emergency Response Center E, Case Management Team KCfDC, Prevention. Coronavirus Disease-19: Summary of 2,370 Contact Investigations of the First 30 Cases in the Republic of Korea. Osong Public Health Res Perspect. 2020;11(2):81-4.
16. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall. 2017;1(1):33-46.
17. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017;22(13).
18. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020;182(4):812-27 e19.

810 19. Becerra-Flores M, Cardozo T. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality
811 rate. Int J Clin Pract. 2020:e13525.
812 20. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics.
813 2018;34(17):i884-i90.
814 21. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-
815 100.
816 22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map
817 format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
818 23. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing2012 July
819 01, 2012:[arXiv:1207.3907 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2012arXiv1207.3907G.
820 24. Sanner MF. Python: a programming language for software integration and development. J Mol
821 Graph Model. 1999;17(1):57-61.
822 25. Kernighan BW, Morgan SP. The UNIX Operating System: A Model for Software Design. Science.
823 1982;215(4534):779-83.
824 26. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File
825 Manipulation. PLoS One. 2016;11(10):e0163962.
826 27. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative
827 genomics viewer. Nature Biotechnology. 2011;29(1):24-6.
828 28. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-
829 performance genomics data visualization and exploration. Briefings in Bioinformatics. 2013;14(2):178-
830 92.
831 29. Robinson JT, Thorvaldsdottir H, Wenger AM, Zehir A, Mesirov JP. Variant Review with the
832 Integrative Genomics Viewer. Cancer Research. 2017;77(21):E31-E4.
833 30. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and
834 predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila
835 melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80-92.
836 31. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for
837 interactive large-scale genome analysis. Genome Res. 2005;15(10):1451-5.
838 32. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for
839 accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res.
840 2018;46(W1):W537-W44.
841 33. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek B, Morowitz MJ, Banfield JF. InStrain
842 enables population genomic analysis from metagenomic data and rigorous detection of identical
843 microbial strains. bioRxiv. 2020:2020.01.22.915579.
844 34. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-
845 performance genomic feature operations. Bioinformatics. 2012;28(14):1919-20.
846 35. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Cech M, et al. The Galaxy platform
847 for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res.
848 2016;44(W1):W3-W10.
849 36. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence
850 alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059-66.
851 37. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment,
852 interactive sequence choice and visualization. Brief Bioinform. 2019;20(4):1160-6.
853 38. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large
854 alignments. PLoS One. 2010;5(3):e9490.
855 39. Steel MA, Fu YX. Classifying and counting linear phylogenetic invariants for the Jukes-Cantor
856 model. J Comput Biol. 1995;2(1):39-47.

857 40. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and
858 annotation. Bioinformatics. 2007;23(1):127-8.
859 41. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments.
860 Nucleic Acids Res. 2019;47(W1):W256-W9.
861 42. Zhang C, Mortuza SM, He B, Wang Y, Zhang Y. Template-based and free modeling of I-TASSER
862 and QUARK pipelines using predicted contact maps in CASP12. Proteins. 2018;86 Suppl 1:136-51.
863 43. He B, Mortuza SM, Wang Y, Shen HB, Zhang Y. NeBcon: protein contact map prediction using
864 neural network training coupled with naive Bayes classifiers. Bioinformatics. 2017;33(15):2296-306.
865 44. Li Y, Hu J, Zhang C, Yu DJ, Zhang Y. ResPRE: high-accuracy protein contact prediction by coupling
866 precision matrix with deep residual neural networks. Bioinformatics. 2019;35(22):4647-55.
867 45. Zheng W, Li Y, Zhang C, Pearce R, Mortuza SM, Zhang Y. Deep-learning contact-map guided
868 protein structure prediction in CASP13. Proteins. 2019;87(12):1149-64.
869 46. Delgado J, Radusky LG, Cianferoni D, Serrano L. FoldX 5.0: working with RNA, small molecules
870 and a new graphical interface. Bioinformatics. 2019;35(20):4168-9.
871 47. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online
872 force field. Nucleic Acids Res. 2005;33(Web Server issue):W382-8.
873 48. Studer RA, Christin PA, Williams MA, Orengo CA. Stability-activity tradeoffs constrain the
874 adaptive evolution of RubisCO. Proc Natl Acad Sci U S A. 2014;111(6):2223-8.
875 49. Rodrigues CHM, Myung Y, Pires DEV, Ascher DB. mCSM-PPI2: predicting the effects of mutations
876 on protein-protein interactions. Nucleic Acids Research. 2019;47(W1):W338-W44.
877 50. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast,
878 flexible, and free. J Comput Chem. 2005;26(16):1701-18.
879 51. Kutzner C, Pall S, Fechner M, Esztermann A, de Groot BL, Grubmuller H. Best bang for your buck:
880 GPU nodes for GROMACS biomolecular simulations. J Comput Chem. 2015;36(26):1990-2008.
881 52. Eargle J, Wright D, Luthey-Schulten Z. Multiple Alignment of protein structures and sequences
882 for VMD. Bioinformatics. 2006;22(4):504-6.
883 53. Vrejoiu C, Costescu A, Burlacu L. Generalized Calculus of Coefficients of Asymptotic Series Using
884 Steepest Descents Method. Stud Cercet Fiz. 1978;30(4):329-45.
885 54. Deserno M, Holm C. How to mesh up Ewald sums. II. An accurate error estimate for the particle-
886 particle-particle-mesh algorithm. Journal of Chemical Physics. 1998;109(18):7694-701.
887 55. Hess B. P-LINCS: A parallel linear constraint solver for molecular simulation. Journal of Chemical
888 Theory and Computation. 2008;4(1):116-22.
889 56. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular
890 simulations. Journal of Computational Chemistry. 1997;18(12):1463-72.
891 57. Reynolds CR, Islam SA, Sternberg MJE. EzMol: A Web Server Wizard for the Rapid Visualization
892 and Image Production of Protein and Nucleic Acid Structures. Journal of Molecular Biology.
893 2018;430(15):2244-8.
894 58. Motulsky HJ, Brown RE. Detecting outliers when fitting data with nonlinear regression - a new
895 method based on robust nonlinear regression and the false discovery rate. Bmc Bioinformatics. 2006;7.
896 59. Simmonds P. Rampant C-->U Hypermutation in the Genomes of SARS-CoV-2 and Other
897 Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories.
898 mSphere. 2020;5(3).
899 60. Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. Evidence for host-dependent
900 RNA editing in the transcriptome of SARS-CoV-2. Sci Adv. 2020;6(25):eabb5813.
901 61. Nelson CW, Hughes AL. Within-host nucleotide diversity of virus populations: Insights from next-
902 generation sequencing. Infect Genet Evol. 2015;30:1-7.
903 62. Miralles R, Gerrish PJ, Moya A, Elena SF. Clonal interference and the evolution of RNA viruses.
904 Science. 1999;285(5434):1745-7.

905 63. Wright CF, Morelli MJ, Thebaud G, Knowles NJ, Herzyk P, Paton DJ, et al. Beyond the consensus:
906 dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-
907 generation genome sequencing. J Virol. 2011;85(5):2266-75.
908 64. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiol Mol Biol Rev.
909 2012;76(2):159-216.
910 65. Ni M, Chen C, Qian J, Xiao HX, Shi WF, Luo Y, et al. Intra-host dynamics of Ebola virus during
911 2014. Nat Microbiol. 2016;1(11):16151.
912 66. Nelson CW, Hughes AL. Within-host nucleotide diversity of virus populations: insights from next-
913 generation sequencing. Infect Genet Evol. 2015;30:1-7.
914 67. Nelson CW, Moncla LH, Hughes AL. SNPGenie: estimating evolutionary parameters to detect
915 natural selection using pooled next-generation sequencing data. Bioinformatics. 2015;31(22):3709-11.
916 68. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and
917 nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986;3(5):418-26.
918 69. Day T, Gandon S, Lion S, Otto SP. On the evolutionary epidemiology of SARS-CoV-2. Curr Biol.
919 2020;30(15):R849-R57.
920 70. Woo H, Park SJ, Choi YK, Park T, Tanveer M, Cao YW, et al. Developing a Fully Glycosylated Full-
921 Length SARS-CoV-2 Spike Protein Model in a Viral Membrane. J Phys Chem B. 2020;124(33):7128-37.
922 71. Portelli S, Olshansky M, Rodrigues CHM, D'Souza EN, Myung Y, Silk M, et al. Exploring the
923 structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource. Nat Genet.
924 2020.
925 72. Lobanov M, Bogatyreva NS, Galzitskaia OV. [Radius of gyration is indicator of compactness of
926 protein structure]. Mol Biol (Mosk). 2008;42(4):701-6.
927 73. Turonova B, Sikora M, Schurmann C, Hagen WJH, Welsch S, Blanc FEC, et al. In situ structural
928 analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. Science. 2020.
929 74. Mercatelli D, Giorgi FM. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Front
930 Microbiol. 2020;11:1800.
931 75. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, et al. Deep Mutational
932 Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell.
933 2020;182(5):1295-310 e20.
934 76. Adams RM, Mora T, Walczak AM, Kinney JB. Measuring the sequence-affinity landscape of
935 antibodies with massively parallel titration curves. Elife. 2016;5.
936 77. Cuevas JM, Geller R, Garijo R, Lopez-Aldeguer J, Sanjuan R. Extremely High Mutation Rate of
937 HIV-1 In Vivo. PLoS Biol. 2015;13(9):e1002251.
938 78. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem.
939 2010;79:321-49.
940 79. van der Walt E, Martin DP, Varsani A, Polston JE, Rybicki EP. Experimental observations of rapid
941 Maize streak virus evolution reveal a strand-specific nucleotide substitution bias. Virol J. 2008;5:104.
942 80. Li Z, Wu J, Deleo CJ. RNA damage and surveillance under oxidative stress. IUBMB Life.
943 2006;58(10):581-8.
944 81. Laforge M, Elbim C, Frere C, Hemadi M, Massaad C, Nuss P, et al. Tissue damage from
945 neutrophil-induced oxidative stress in COVID-19. Nat Rev Immunol. 2020;20(9):515-6.
946 82. Pariente N, Sierra S, Lowenstein PR, Domingo E. Efficient virus extinction by combinations of a
947 mutagen and antiviral inhibitors. J Virol. 2001;75(20):9723-30.
948 83. Elena SF, Miralles R, Cuevas JM, Turner PE, Moya A. The two faces of mutation: extinction and
949 adaptation in RNA viruses. IUBMB Life. 2000;49(1):5-9.
950 84. Gojobori T, Moriyama EN, Kimura M. Molecular Clock of Viral Evolution, and the Neutral Theory.
951 P Natl Acad Sci USA. 1990;87(24):10015-8.

85. To KK, Hung IF, Ip JD, Chu AW, Chan WM, Tam AR, et al. COVID-19 re-infection by a phylogenetically distinct SARS-coronavirus-2 strain confirmed by whole genome sequencing. Clin Infect Dis. 2020.
86. Phillips AM, Ponomarenko AI, Chen K, Ashenberg O, Miao J, McHugh SM, et al. Destabilized adaptive influenza variants critical for innate immune system escape are potentiated by host chaperones. PLoS Biol. 2018;16(9):e3000008.
87. Voeten JT, Bestebroer TM, Nieuwkoop NJ, Fouchier RA, Osterhaus AD, Rimmelzwaan GF. Antigenic drift in the influenza A virus (H3N2) nucleoprotein and escape from recognition by cytotoxic T lymphocytes. J Virol. 2000;74(15):6800-7.
88. Rosenberg W. Mechanisms of immune escape in viral hepatitis. Gut. 1999;44(5):759-64.

Supporting information

Supplementary Table 1: Sequencing metadata of 76554 GISAID genomes downloaded until August 3, 2020.

For every GISAID genome, we provided GISAID genome name, GISAID unique identifier (Accession ID), geographic location, host, sequencing technology, lineage, and clade fields, among other information. The last column indicates the number of variants per genome (Major viral variants).

Supplementary Table 2: Acknowledgements from sequencing laboratories and/or consortia associated with GISAID genomes listed in Supplementary Table 1, plus genomes downloaded until September 28, 2020, containing variants V1176F, S477N and A222V occurring in the Spike protein.

Supplementary Table 3: Sequencing metadata of 17560 Sequencing Read Archive (SRA) datasets downloaded until July 28, 2020.

For every SRA dataset, we provided NCBI run accession, Assay type (indicates if amplicon, RNA-seq u other sequencing corresponds), sequencing size (bases, in nucleotides), Biosample accession ID, Center Name (depositor), release date, SRA study accession, BioProject and geographic location, among other information. The last column indicates the number of variants per sample (Major viral variants, viral frequency > 0.5).

Supplementary Table 4: Sequencing metadata of 543 and 393 GISAID genomes containing variants N: I292T and S: V1176F, respectively.

We accessed GISAID database on September 28, 2020 and we downloaded genomes containing the variant I292T or V1176F. The associated metadata from both cohorts are presented in this table. For every GISAID genome, we provided GISAID genome name, GISAID unique identifier (Accession ID), collection information, geographic location, host, sequencing technology, lineage, and clade, among other information. We used GISAID unique identifiers to overlap both groups. No overlap was found.

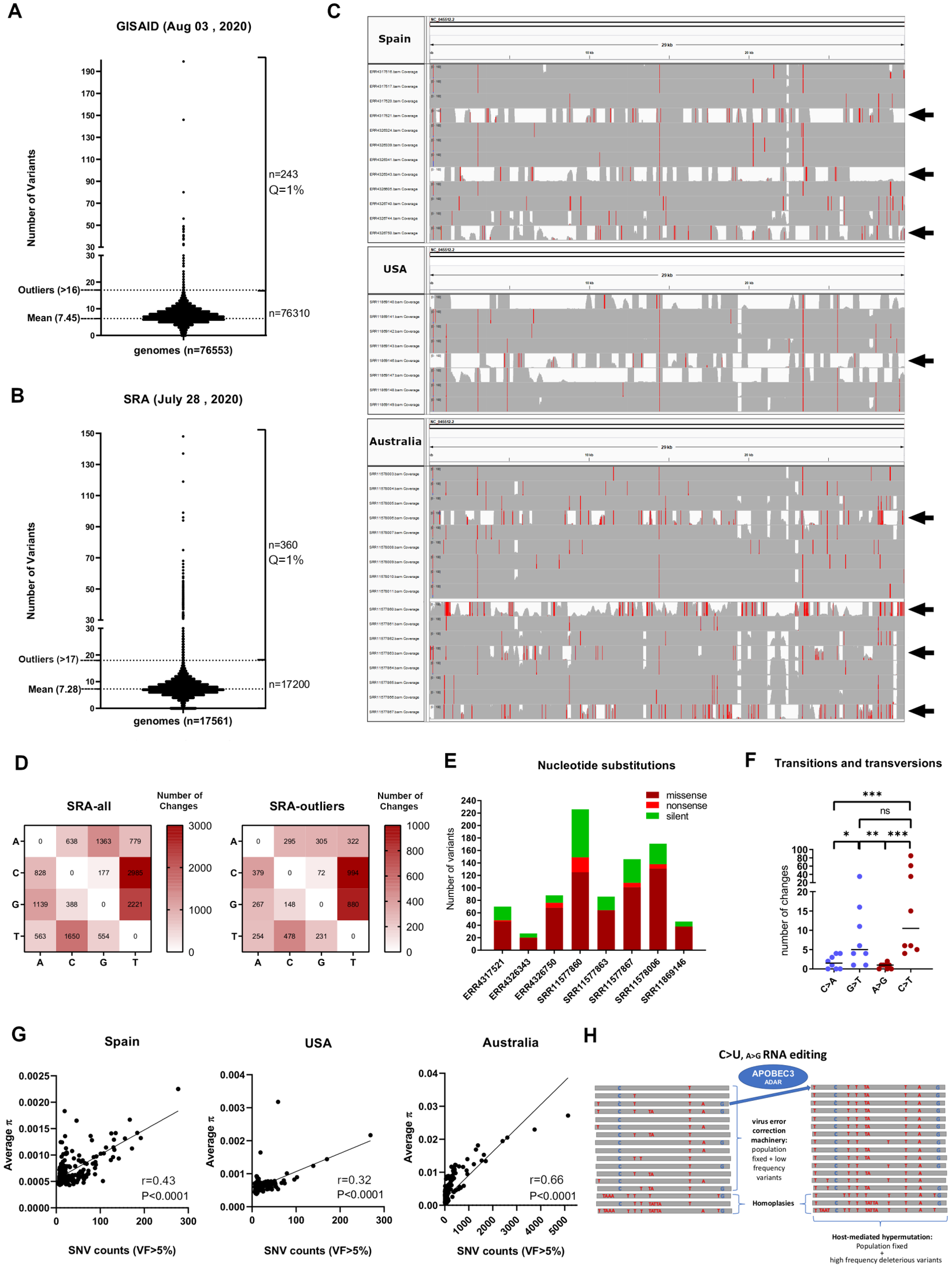
Supplementary Table 5: Worldwide fatality ratios per country, related to Supplementary Figure 2.

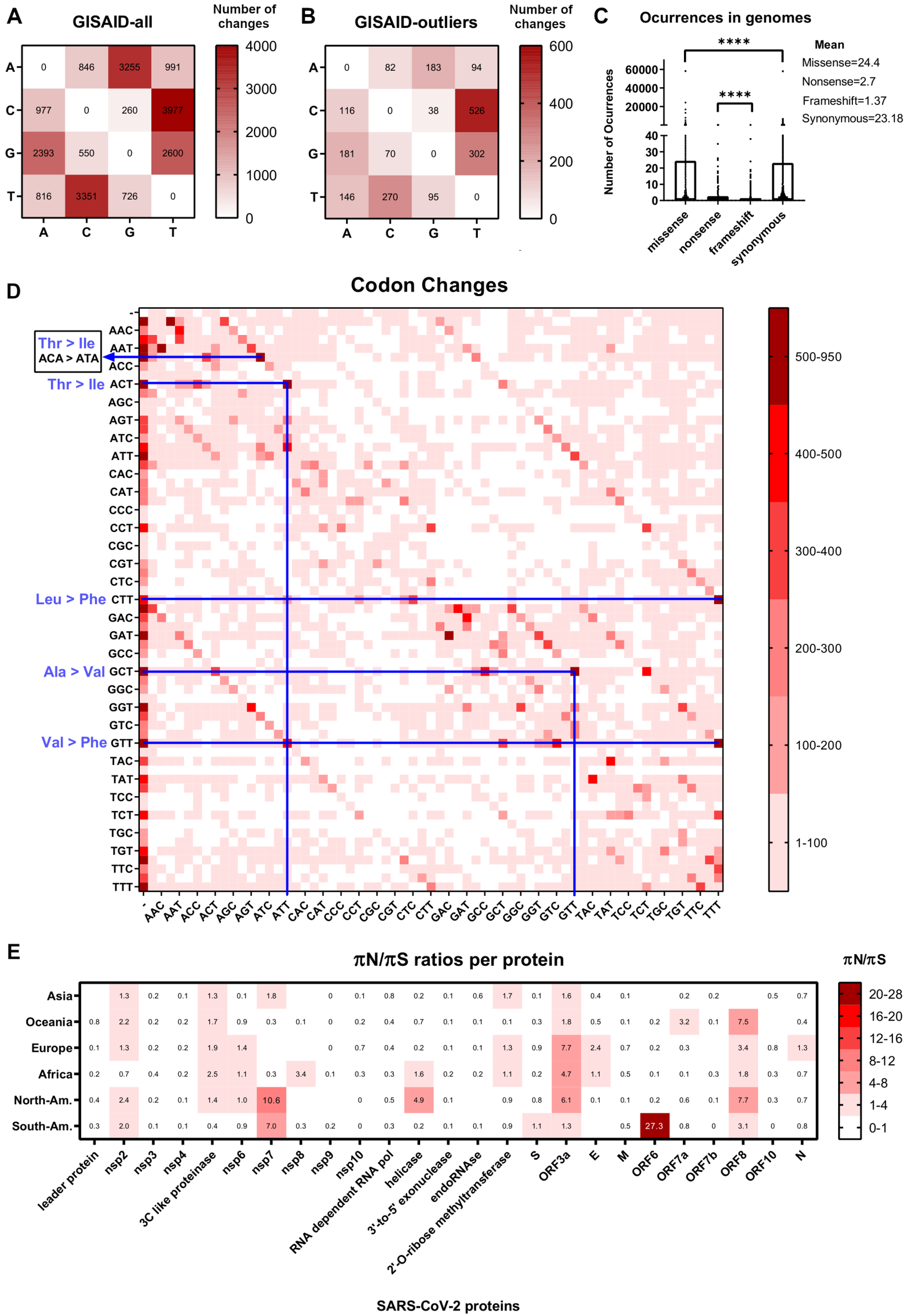
Worldwide fatality ratios among 49 countries obtained from John Hopkins coronavirus resource center (https://coronavirus.jhu.edu/map.html), accessed on September 28, 2020. At right, we calculated viral allele frequencies of several SARS-CoV-2 variants per country, based on the GISAID database, also accessed on September 28, 2020. We analyzed the following

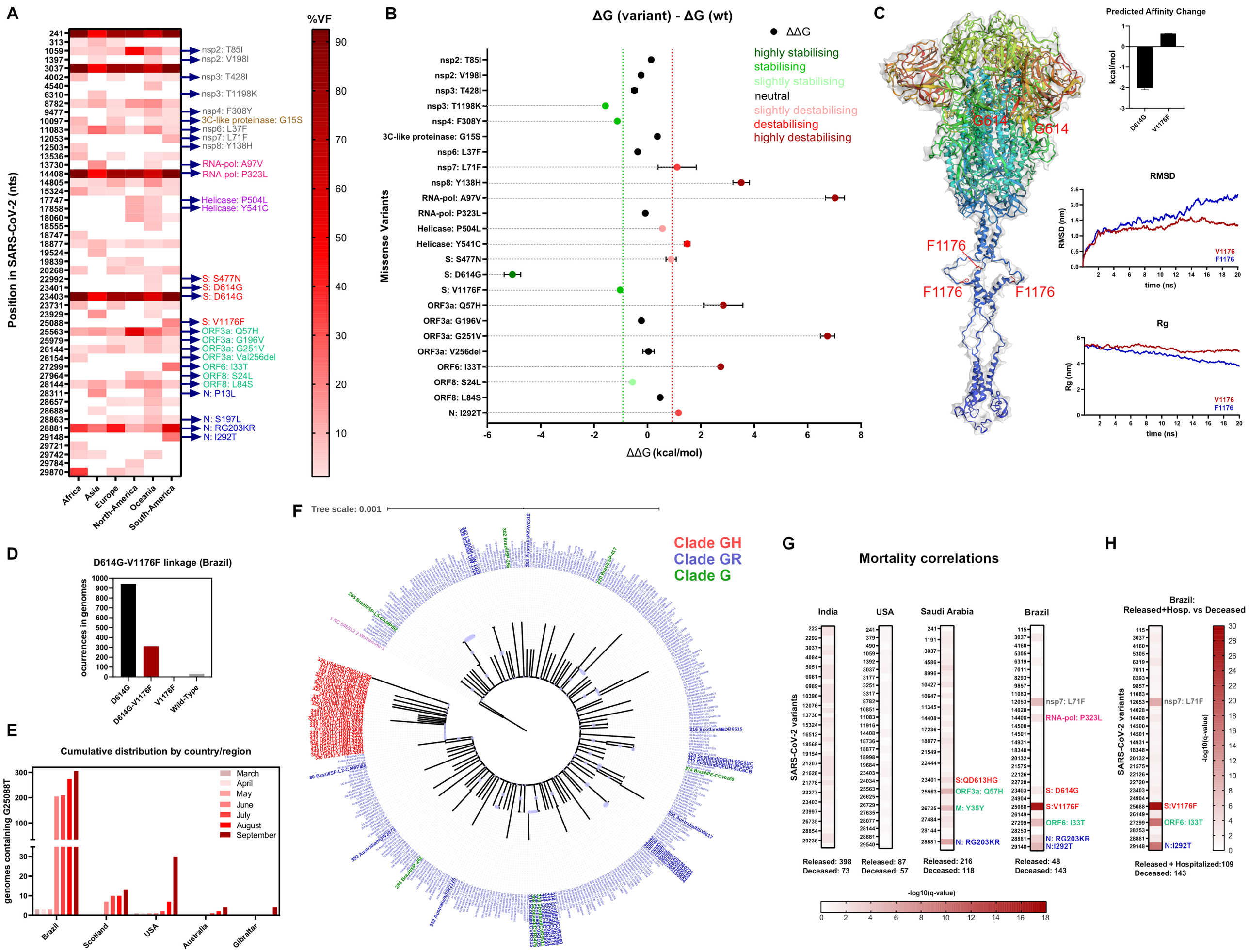
variants: ORF3a (Q57H), N (RG203KR), RNA-dependent RNA polymerase (P323L), S (D614G), S (V1176F), nsp7 (L71F) and N (I292T), respectively.

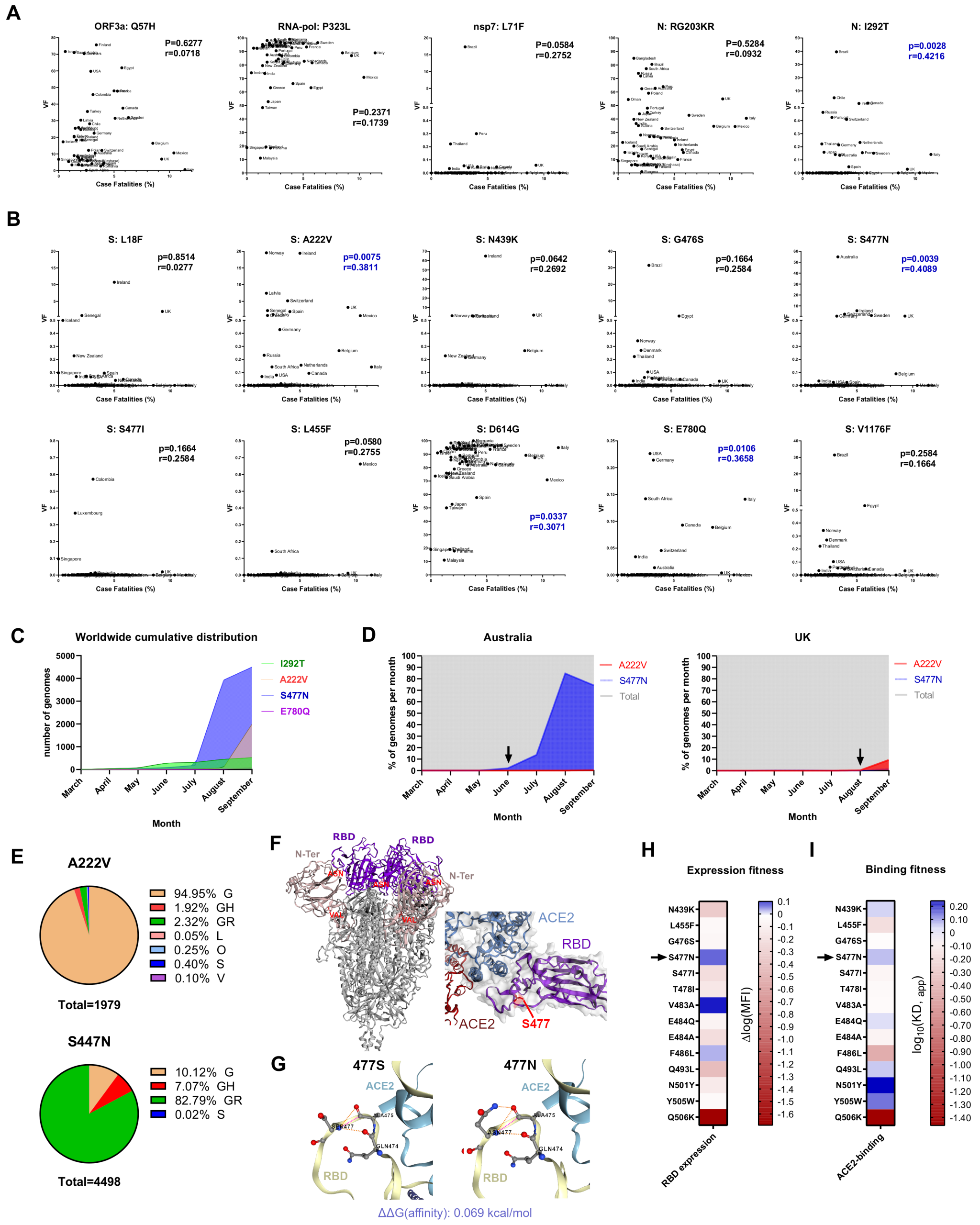
Supplementary Figure 1: Intra Host variant effects.

23,269 aggregated variants from 76554 GISAID genomes were merged as indicated in Figure 2A. These variants were classified by SnpEff program, available in the Galaxy server. We plotted a heatmap with the number of changes per consequence type (see x-axis) against every SARS-CoV-2 protein (see y-axis). We also calculated the overall missense/silent ratio occurring in SARS-CoV-2 proteins (1.82).


A
GISAID (Aug 03 , 2020)
Number of Variants
n=243
Q=1%
Outliers (>16)
Mean (7.45)
n=76310
genomes (n=76553)
B
SRA (July 28 , 2020)
Number of Variants
n=360
Q=1%
Outliers (>17)
Mean (7.28)
n=17200
genomes (n=17561)
C
Spain
USA
Australia
29 kb
D
SRA-all
Number of Changes
A 0 638 1363 779
C 828 0 177 2985
G 1139 388 0 2221
T 563 1650 554 0
SRA-outliers
Number of Changes
A 0 295 305 322
C 379 0 72 994
G 267 148 0 880
T 254 478 231 0
E
Nucleotide substitutions
missense
nonsense
silent
Number of variants
ERR4317521
ERR4326343
ERR4326750
SRR11577860
SRR11577863
SRR11577867
SRR11578006
SRR11869146
F
Transitions and transversions
number of changes
C>A
G>T
A>G
C>T
ns
G
Spain
USA
Australia
Average π
SNV counts (VF>5%)
r=0.43
P<0.0001
r=0.32
P<0.0001
r=0.66
P<0.0001
H
C>U, A>G RNA editing
APOBEC3
ADAR
virus error correction machinery:
population fixed + low frequency variants
Homoplasies
Host-mediated hypermutation:
Population fixed
+
high frequency deleterious variants


A
GISAID-all
Number of changes
4000
3000
2000
1000
0
A C G T
A 0 846 3255 991
C 977 0 260 3977
G 2393 550 0 2600
T 816 3351 726 0
B
GISAID-outliers
Number of changes
600
400
200
0
A C G T
A 0 82 183 94
C 116 0 38 526
G 181 70 0 302
T 146 270 95 0
C
Ocurrences in genomes
Number of Ocurrences
60000
40000
20000
40
30
20
10
0
****
****
Mean
Missense=24.4
Nonsense=2.7
Frameshift=1.37
Synonymous=23.18
missense
nonsense
frameshift
synonymous
D
Codon Changes
Thr > Ile
ACA > ATA
Thr > Ile
Leu > Phe
Ala > Val
Val > Phe
500-950
400-500
300-400
200-300
100-200
1-100
E
πN/πS ratios per protein
πN/πS
20-28
16-20
12-16
8-12
4-8
1-4
0-1
Asia 1.3 0.2 0.1 1.3 0.1 1.8 0 0.1 0.8 0.2 0.1 0.6 1.7 0.1 1.6 0.4 0.1 0.2 0.2 0.5 0.7
Oceania 0.8 2.2 0.2 0.2 1.7 0.9 0.3 0.1 0 0.2 0.4 0.7 0.1 0.1 0.1 0.3 1.8 0.5 0.1 0.2 3.2 0.1 7.5 0.4
Europe 0.1 1.3 0.2 0.2 1.9 1.4 0 0.7 0.4 0.2 0.1 0.1 1.3 0.9 7.7 2.4 0.7 0.2 0.3 3.4 0.8 1.3
Africa 0.2 0.7 0.4 0.2 2.5 1.1 0.3 3.4 0.1 0.3 0.3 1.6 0.2 0.2 1.1 0.2 4.7 1.1 0.5 0.1 0.1 0.3 1.8 0.3 0.7
North-Am. 0.4 2.4 0.2 0.1 1.4 1.0 10.6 0 0.5 4.9 0.1 0.9 0.8 6.1 0.1 0.1 0.2 0.6 0.1 7.7 0.3 0.7
South-Am. 0.3 2.0 0.1 0.1 0.4 0.9 7.0 0.3 0.2 0 0.3 0.2 0.1 0.1 0.9 1.1 1.3 0.5 27.3 0.8 0 3.1 0 0.8
leader protein
nsp2
nsp3
nsp4
3C like proteinase
nsp6
nsp7
nsp8
nsp9
nsp10
RNA dependent RNA pol
helicase
3'-to-5' exonuclease
endoRNAse
2'-O-ribose methyltransferase
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
ORF10
N
SARS-CoV-2 proteins


A
%VF
Position in SARS-CoV-2 (nts)
nsp2: T85I
nsp2: V198I
nsp3: T428I
nsp3: T1198K
nsp4: F308Y
3C-like proteinase: G15S
nsp6: L37F
nsp7: L71F
nsp8: Y138H
RNA-pol: A97V
RNA-pol: P323L
Helicase: P504L
Helicase: Y541C
S: S477N
S: D614G
S: D614G
S: V1176F
ORF3a: Q57H
ORF3a: G196V
ORF3a: G251V
ORF3a: Val256del
ORF6: I33T
ORF8: S24L
ORF8: L84S
N: P13L
N: S197L
N: RG203KR
N: I292T
Africa
Asia
Europe
North-America
Oceania
South-America
B
ΔG (variant) - ΔG (wt)
ΔΔG
highly stabilising
stabilising
slightly stabilising
neutral
slightly destabilising
destabilising
highly destabilising
Missense Variants
ΔΔG (kcal/mol)
C
G614
G614
F1176
F1176
F1176
Predicted Affinity Change
kcal/mol
D614G
V1176F
RMSD
RMSD (nm)
time (ns)
V1176
F1176
Rg
Rg (nm)
time (ns)
V1176
F1176
D
D614G-V1176F linkage (Brazil)
ocurrences in genomes
D614G
D614G-V1176F
V1176F
Wild-Type
E
Cumulative distribution by country/region
genomes containing G25088T
March
April
May
June
July
August
September
Brazil
Scotland
USA
Australia
Gibraltar
F
Tree scale: 0.001
Clade GH
Clade GR
Clade G
G
Mortality correlations
India
USA
Saudi Arabia
Brazil
SARS-CoV-2 variants
S:QD613HG
ORF3a: Q57H
M: Y35Y
N: RG203KR
nsp7: L71F
RNA-pol: P323L
S: D614G
S:V1176F
ORF6: I33T
N: RG203KR
N:I292T
Released: 398
Deceased: 73
Released: 87
Deceased: 57
Released: 216
Deceased: 118
Released: 48
Deceased: 143
-log10(q-value)
H
Brazil:
Released+Hosp. vs Deceased
SARS-CoV-2 variants
nsp7: L71F
S:V1176F
ORF6: I33T
N:I292T
-log10(q-value)
Released + Hospitalized:109
Deceased: 143


A
ORF3a: Q57H
P=0.6277
r=0.0718
RNA-pol: P323L
P=0.2371
r=0.1739
nsp7: L71F
P=0.0584
r=0.2752
N: RG203KR
P=0.5284
r=0.0932
N: I292T
p=0.0028
r=0.4216
VF
Case Fatalities (%)
B
S: L18F
p=0.8514
r=0.0277
S: A222V
p=0.0075
r=0.3811
S: N439K
p=0.0642
r=0.2692
S: G476S
p=0.1664
r=0.2584
S: S477N
p=0.0039
r=0.4089
S: S477I
p=0.1664
r=0.2584
S: L455F
p=0.0580
r=0.2755
S: D614G
p=0.0337
r=0.3071
S: E780Q
p=0.0106
r=0.3658
S: V1176F
p=0.2584
r=0.1664
C
Worldwide cumulative distribution
number of genomes
I292T
A222V
S477N
E780Q
March
April
May
June
July
August
September
Month
D
Australia
UK
% of genomes per month
A222V
S477N
Total
E
A222V
94.95% G
1.92% GH
2.32% GR
0.05% L
0.25% O
0.40% S
0.10% V
Total=1979
S447N
10.12% G
7.07% GH
82.79% GR
0.02% S
Total=4498
F
RBD
N-Ter
ASN
VAL
ACE2
S477
G
477S
477N
ACE2
RBD
ΔΔG(affinity): 0.069 kcal/mol
H
Expression fitness
N439K
L455F
G476S
S477N
S477I
T478I
V483A
E484Q
E484A
F486L
Q493L
N501Y
Y505W
Q506K
Δlog(MFI)
RBD expression
I
Binding fitness
log10(KD, app)
ACE2-binding