Article - Discoveries

Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution

Vítor Borges[1*], Maria João Alves[2*], Massimo Amicone[3*], Joana Isidro[1*], Líbia Zé-Zé[2,4], Sílvia Duarte[5], Luís Vieira[5,6], Raquel Guiomar[7], João Paulo Gomes[1✉], Isabel Gordo[3✉]

[1]Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health Doutor Ricardo Jorge (INSA), Lisbon, Portugal.

[2]Centre for Vectors and Infectious Diseases Research, Department of Infectious Diseases, National Institute of Health Doutor Ricardo Jorge (INSA), Águas de Moura, Portugal

[3]Instituto Gulbenkian de Ciência, Oeiras

[4]BioISI - Biosystems & Integrative Sciences Institute, Faculty of Sciences, University of Lisbon, Portugal

[5]Innovation and Technology Unit, Department of Human Genetics, National Institute of Health Doutor Ricardo Jorge (INSA), Lisbon, Portugal.

[6]Centre for Toxicogenomics and Human Health (ToxOmics), Genetics, Oncology and Human Toxicology, Nova Medical School|Faculdade de Ciências Médicas, Universidade Nova de Lisboa, Lisbon, Portugal.

[7]National Reference Laboratory for Influenza and other Respiratory Viruses, Department of Infectious Diseases, National Institute of Health Doutor Ricardo Jorge (INSA), Lisbon, Portugal.

*These authors contributed equally to this work.

✉e-mail: igordo@igc.gulbenkian.pt; j.paulo.gomes@insa.min-saude.pt

Abstract

"How predictable is evolution?" is a key question in evolutionary biology. Experimental evolution has shown that the evolutionary path of microbes can be extraordinarily reproducible. Here, using experimental evolution in two circulating SARS-CoV-2, we estimate its mutation rate and demonstrate the repeatability of its evolution when facing a new cell type but no immune or drug pressures. We estimate a genomic mutation rate of 3.7×10^{-6} nt^{-1} $cycle^{-1}$ for a lineage of SARS-CoV-2 with the originally described spike protein (CoV-2-D) and of 2.9×10^{-6} nt^{-1} $cycle^{-1}$ for a lineage carrying the D614G mutation that has spread worldwide (CoV-2-G). We further show that mutation accumulation is heterogeneous along the genome, with the spike gene accumulating mutations at a mean rate 16×10^{-6} nt^{-1} per infection cycle across backgrounds, five-fold higher than the genomic average. We observe the emergence of mutators in the CoV-2-G background, likely linked to mutations in the RNA-dependent RNA polymerase and/or in the error-correcting exonuclease protein. Despite strong bottlenecks, several *de novo* mutations spread to high frequencies by selection and considerable convergent evolution in spike occurs. These results demonstrate the high adaptive potential of SARS-CoV-2 during the first stages of cell infection in the absence of immune surveillance.

Introduction

Mutation is the principal process driving the origin of genetic diversity. The genomic mutation rate and spectrum are fundamental parameters to understand how predictable molecular evolution can be. It is also important to determine how natural selection fixes or purges new variants from natural populations. While the genomic mutation rate per cell per generation is close to a constant for DNA based microbes(Drake, 1991), for RNA viruses there is a remarkable variation(Drake, 1993; Drake et al., 1998). Mutation rates, expressed as nucleotide substitutions per site per cell infection (s/n/c), vary between 10^{-6} to 10^{-3} for the several positive ssRNA viruses which have been studied(Sanjuán et al., 2010). However, the mutation rate of the human beta-coronavirus SARS-CoV-2, which is the cause of the COVID-19 pandemic(Zhu et al., 2020), is currently unknown(Bar-On et al., 2020). This rate is critical for accurate inferences of SARS-CoV-2 demographic history(van Dorp et al., 2020) and potential molecular adaptations(Lässig et al., 2017), as it strives in the human population.

Microbial populations that evolve and adapt to novel environments under strong selection pressures typically show a pattern of evolutionary convergence(Bull et al., 1997; Tenaillon et al., 2012). Laboratory mutation accumulation experiments combined with high-throughput sequencing are one of the best methods to estimate mutation rates, determine how they vary along the genome(Lynch et al., 2016) and study the extent to which convergent evolution occurs(Bull et al., 1997). With an experimental evolution design where sequential propagation of the SARS-CoV-2 virus occurs under strong bottlenecks and at low multiplicity of infection (MOI), the intensity of selection is minimized(Chao, 1990; Duarte et al., 1992). Thus, most mutations observed reflect the rate at which they originate more closely than the ones observed under the complex selective pressures that may ensue in a real host infection. Using experimental evolution of SARS-CoV-2 propagated under relatively benign conditions, the spectrum of emergence of non-lethal mutations can be estimated. Such information is necessary to close a critical knowledge gap in the basic biology of this virus and quantify how predicable its evolution can be. It may also help determining some potential genomic constraints of the virus, which are key to the design of evolution proof vaccines and antiviral drugs.

Here we perform experimental evolution in two natural isolates of SARS-CoV-2(Borges et al., 2020). We estimate SARS-CoV-2 genomic mutation rate and the extent to which it varies along the genome, via whole genome sequencing.

Results and Discussion

Two SARS-CoV-2 viruses were isolated from two non-related patients (see Materials and Methods) for continuous propagation in Vero cells. The viruses were chosen according to their polymorphism at amino acid position 614 of the spike protein: CoV-2-D carries a D (belonging to Nextstrain clade 19A) in this position and CoV-2-G (belonging to Nextstrain clade 20B) carries a derived mutation which changes this amino-acid to a G. D614G in spike emerged early in the pandemic, increased the infectivity of the virus and became prevalent worldwide(Korber et al., 2020). The rational of the design of the experimental evolution propagations (**Fig. 1a**) was motivated by the estimates of the mutation rate for the beta-coronavirus mouse hepatitis virus (MHV). For MHV the rate is about $3.5x10^{-6}$ nt^{-1} $cycle^{-1}$ and an average number of virions produced by a single infected cell of the order of 1000 has been estimated(Bar-On et al., 2020, p.; Sanjuán et al., 2010). With a MOI of 1:10, *i.e.* 2000 virus for

86 $2x10^4$ cells at each passage, a 24h passage corresponds to ~1 infection cycle. Thus, 15 passages
87 of 2x96 experimental evolution populations should lead to the emergence of hundreds of
88 mutations.
89
90 The two clinical isolates Portugal/PT0054/2020 and Portugal/PT1136/2020 were used to
91 produce the ancestors of the experimental evolution, CoV-2-D and CoV-2-G, which seeded the
92 two mutation accumulation experiments. The clinical isolates and the corresponding ancestors
93 were sequenced using Illumina technology at a high depth of coverage (mean = 4835x) and
94 mapped to the Wuhan-Hu-1/2019 reference genome sequence(Wu et al., 2020). The clinical
95 isolate Portugal/PT1136/2020 and ancestral (CoV-2-G) viral populations share eight fixed or
96 nearly fixed nucleotide changes: C241T (5' UTR); C3037T (orf1ab); C14408T (orf1ab);
97 C19041T (orf1ab); A23403G (S gene), which causes the D614G replacement in spike protein;
98 and, the 20B clade-defining triplet mutation GGG to AAC in genome positions 28881-3 (N
99 gene) (**Fig. 1b**). Another mutation present at lower frequency was also shared and 10 mutations
100 unique to the CoV-2-G were detected (**Fig. 1b**). The clinical isolate Portugal/PT0054/2020 and
101 the ancestral CoV-2-D share the fixed mutation G1440A (orf1ab) and the mutation G2891A
102 (orf1ab), which is at high frequency in both populations (98% and 73% in the clinical and
103 ancestral populations, respectively), and two other mutations (T9843A and C12921T; orf1ab)
104 at lower frequencies. CoV-2-D ancestral line carried 10 other minor frequency mutations not
105 shared with those of the clinical isolate (**Fig. 1b**).
106
107 Propagation of the 96 CoV-2-D derived lines resulted in the accumulation of 1753 *de novo*
108 mutations that reached a frequency of at least 1%, supported by a minimum of 10 reads
109 (**Supplementary Table 1**). We assume these to be *de novo* mutations accumulated during the
110 propagations, as they were not detected in either the ancestor or the original clinical isolate
111 from which the ancestor was derived. Of these *de novo* mutations, 415 were synonymous, 971
112 were non-synonymous SNPs, 367 were other events (e.g., frameshift SNP, indels, etc)
113 (**Supplementary Table 1**). From all the mutations accumulated in the 96 lines, and the
114 frequencies at which they are observed at passage 15, we estimate a mutation rate of 0.11 per
115 genome per infection cycle, corresponding to a rate of $3.7x10^{-6}$ nt^{-1} $cycle^{-1}$ for CoV-2-D. The
116 estimates of mutation rate from the mutations accumulated in each lineage propagated are
117 shown in **Fig. 2a**.

The propagation of the 96 lines derived from CoV-2-G resulted in 6181 *de novo* mutations (n=94 as in 2 lines the sequencing had poor coverage). Of these 724 were synonymous, 4313 were non-synonymous SNPs and 1144 were other events (e.g., frameshift SNP, indels, etc) (**Supplementary Table 1**). The much higher number of mutations in the lines of CoV-2-G, compared to CoV-2-D, was due to the emergence of mutator virus in some of the lines (see below). These mutator lines had a median of number of *de novo* mutations of 301 per line (IQR=287-343), when compared with 19 (IQR=16-22) observed for non-mutators. Excluding those lines (*n*=15) led us to an estimate of the mutations rate of 2.9×10^{-6} nt^{-1} $cycle^{-1}$ for the lineage CoV-2-G (**Fig. 2a**). Interestingly the rate for lineages CoV-2-G is slightly but significantly lower than that of lineages CoV-2-D ($P<10^{-7}$, T-test) (**Fig. 2a**). Assuming that all mutations accumulated are neutral, this difference could indicate a reduction in the rate of mutation for the CoV-2-G viral background. However, if selection is acting on some of the mutations (see below), then the differences may just reflect differences in selective pressures in the two viral backgrounds.

Both genomic backgrounds show a strong mutation bias of C->T, a well-known bias of SARS-CoV-2(Matyášek and Kovařík, 2020; Rice et al., 2021) (**Fig. 2b**). A lower rate than expected was observed for C->G and G->C mutations. Interestingly, in the lines that evolved mutators the mutation bias changed from C->T to G->T (**Fig. 2b**). The genetic cause of the mutator phenotype is difficult to determine but it could likely be hidden within the mutations that occurred in the nsp12 RNA-dependent RNA polymerase and/or its nsp14. Indeed, we find 8 non-synonymous mutations in nsp12, one leading to a stop at amino-acid 670, that are specific to the lineages with mutators (**Supplementary Table 2**), and 9 non-synonymous mutations specific to the mutators in the error-correcting exonuclease protein - nsp14 (one mutation leading to a stop at amino acid 78) (**Supplementary Table 3**). Any of these mutations could potentially lead to the observed change in mutation rate, but none of these mutations are similar to those that have been associated with an increased mutational load of the circulating viruses(Eskier et al., 2020). Thus the *in vitro* and *in natura* results suggest that many changes in either nsp12 or nsp14 could cause the genomic mutation rate of SARS-CoV-2 to increase.

Heterogeneity of the mutation rate can occur across the genome of SARS-CoV-2, irrespectively of the emergence of mutators. Such heterogeneity can be caused solely by variation in mutation rate, due to the mutation bias (**Fig. 2b**), and/or it can arise due to selection acting on some gene but not in others(Sasaki et al., 2021). The distribution of allele frequencies

152 in a sample, *i.e.* the site frequency spectrum, has a well-known theoretical expectation under a
153 simple equilibrium neutral model of molecular evolution (Chap. 5 pg. 233 of (Charlesworth
154 and Charlesworth, 2010)). This distribution is sensitive to the action of selection and also to
155 complex demographic events, such as population bottlenecks. Given the strong bottlenecks
156 occurring in the experimental evolution setup (**Fig. 1a**) and the slow evolutionary time elapsed
157 during the 15 infection cycles, the simple equilibrium neutral theoretical expectation may not
158 apply. To obtain a non-equilibrium expectation of the frequency spectrum of mutations, we
159 performed forward-simulations to model the acquisition of neutral mutations at a rate 3.3×10^{-6}
160 nt^{-1} $cycle^{-1}$, as inferred from the data obtained. We observe that the site frequency spectrum of
161 the mutations detected on the evolved lines of CoV-2-D or CoV-2-G deviates significantly
162 from that expected from simulated population genetics neutral model of this type of viral
163 propagations (**Fig. 3a**). Under neutrality the bump in the distribution of mutations for the higher
164 frequencies is not expected to occur. Furthermore, and to account for some contamination that
165 can occur when the lines were propagated, and/or during the sequencing process, we performed
166 another set of simulations where migration is allowed to occur during the passages. We find
167 that even when considering a contamination level of 10% (migration rate m=0.1), the neutral
168 site frequency spectrum is still incompatible with the experimental data (**Fig. 3b**). This strongly
169 suggests that selection may have shaped the frequency dynamics of the mutations that were
170 detected, despite the strong bottlenecks that were used in the propagations (see **Supplementary**
171 **Fig. S2** showing the mutations that reached the highest frequencies). We also analyzed how
172 migration would affect our estimation of the mutation rate. We observed that 10% of between
173 well contamination can cause a slight overestimation of the mutation rate, but within the 95%CI
174 of our initial estimate (**Fig. 3b and 2a**).
175
176 Our experiment was designed to minimize selection on the virus(Chao, 1990), however several
177 other patterns indicate that selection played a significant role. Firstly, considerable variation in
178 the rate of mutation accumulation is observed across the genome (**Fig. 4a-c, Supplementary**
179 **Fig. S1**). When comparing the mutation rates across the different gene regions in the CoV-2-
180 D background, we find that the S gene, coding for the spike protein, has the highest rate of
181 mutation accumulation. Even excluding spike, there is significant rate variation across the
182 genes (Kruskal-Wallis chi-squared = 238.09, df = 7, p-value $< 10^{-16}$) (**Fig. 4**). In the CoV-2-G
183 background, we again find that the S gene has the highest rate of mutation accumulation (**Fig.**
184 **4a-c**), when excluding the lines with mutators. Remarkably, the mutation rate estimated for the
185 spike gene was 17.1×10^{-6} (95% CI 16.1×10^{-6}, 18.1×10^{-6}) for the background CoV-2-D. This

186 rate is 4.6 times the genome average suggesting the action of selection in these lines despite
187 the strong bottlenecks in the experiment. The rate for spike for CoV-2-G was $15.5x10^{-6}$ (95%
188 CI $14.5x10^{-6}$, $16.5x10^{-6}$), considering lines with both mutator and non-mutator viruses. This is
189 significantly different from that estimated for the CoV-2-D background (P=0.024, T-test).
190 When excluding the mutator lines, that only were detected when evolving the CoV-2-G
191 background, the rate of mutation accumulation in spike is $13.5x10^{-6}$ $nt^{-1}/cycle^{-1}$ (95% CI
192 $13.1x10^{-6}$, $13.9x10^{-6}$) significantly lower than that for background CoV-2-D ($P<10^{-7}$, T-test).
193 Indeed, the spike gene evolved ~2 times faster in the mutator lines than the non-mutators (**Fig.**
194 **4c**, **Supplementary Fig. S1**). Secondly, the accumulation of non-synonymous *vs* synonymous
195 SNPs was also indicative of selection in some coding regions. We find that spike accumulated
196 the highest number of non-synonymous mutations in both genetic backgrounds and showed a
197 ratio of non-synonymous to synonymous mutations well above 1, 14 for CoV-2-D and 22 for
198 the CoV-2-G background (**Supplementary Fig. S3** and **Supplementary Table 4**). This further
199 shows that natural selection shaped the evolution of the spike when the virus propagates in
200 cells.
201
202 One important signature of selection is the level of genetic convergence across independent
203 populations evolving in the same environment. We quantified the level of convergence at the
204 nucleotide and amino levels between CoV-2-D and CoV-2-G. Convergence between the two
205 backgrounds reflects true independent origin of the mutations, as the lines of each background
206 were propagated and processed for sequencing independently. In contrast, convergence within
207 replicates of the same background can result from the same mutation originating independently
208 or from migration due to some possible cross-contamination. Conditioning on the same
209 mutation occurring across backgrounds, we find that for spike, 17 sites and 3 small regions out
210 of its 3822 nucleotides show evolutionary convergence (**Fig. 4d**). At the amino level 20 specific
211 sites and 3 regions were hit independently (**Fig. 4d, Supplementary Table 5)** Consistent with
212 previous reports(Sasaki et al., 2021) we find high evolutionary convergence at the S1/S2
213 cleavage site: three distinct deletions (675-QTQTN-679 del; 679-NSPRRAR-685 del and 679-
214 NSPRRARSVA-688 delTNSPRRARSVA678-688T, emerge multiple times in both
215 backgrounds. Such changes have been previously shown to emerge rapidly in Vero cells and
216 to be important for the virus cell tropism(Sasaki et al., 2021). Apart from these deletions,
217 mutations in 682 site were also highly convergent, most likely because they trigger a similar
218 functional effect, i.e., knock out of the furin cleavage site(Liu et al., 2020). Notably, another
219 deletion in this region (678-TNSPRRARS-686 del) was frequently observed, still it was

220 exclusive of CoV-2-D lines (*n*= 58), suggesting that the conformation changes mediated by
221 Spike D614G may influence the directionality of the evolution towards the knock out of the
222 furin-cleavage site(Gobeil et al., 2021).
223
224 Evolutionary convergence could also be found for the structural genes N, E and M
225 (**Supplementary Table 6**). The convergence was however not as high as it was for the S gene:
226 3 cases of convergence at the nucleotide level for N, leading to the P13L, T379I amino acid
227 changes and a stop at Q418; 6 cases of convergence at the nucleotide level for E, leading to the
228 T9I, L19F, V24M, T30I, L37F and S68F amino acid changes; 2 cases for the gene M, one of
229 which leading to the H125Y amino acid change. The cases of convergence suggest that
230 adaptation could also have occurred in these genes.
231
232 The observed emergence of mutators in the CoV-2-G background (**Fig. 2a**) is worrisome as
233 despite these being found at low frequency (~1-2% after 15 infection cycles), they carry many
234 mutations in the spike protein including in the receptor binding domain -RBD- (amino acid
235 changes at positions 328, 339, 364, 416, 454, 465, 474, 479, 482, 522 and 524) and multi
236 cleavage site regions (positions 798 and 799) (**Supplementary Table 1**). We did not observe
237 any mutation causing either a N501Y or a E484K change, which are mutations of concern due
238 to the fitness advantages they confer to SARS-CoV-2 as it infects the human population.
239 However, we do find changes in amino-acids 80 and 215 that are mutated in the South African
240 variant, in 190 and 655(Makoni, 2021), mutated in the P1 variant originated in Brazil and the
241 D253G found in an emerging lineage of viral isolates in the New York region (**Fig 4e**). The
242 emergence of mutators also shows that the genomic mutation rate of SARS-CoV-2 can increase
243 without significant loss of viability, at least in the short run. This implies that strategies to
244 reduce viral fitness using mutagens should be tested with great precaution(Jensen et al., 2020;
245 Jensen and Lynch, 2020).
246
247 The SARS-CoV-2 betacoronavirus, first observed in the Wuhan province of China(Zhu et al.,
248 2020) has infected at least 111 million people causing more than a 2 million toll of deaths in
249 the human population (as of February 2021). Since it was first sequenced(Wu et al., 2020) the
250 virus has been accumulating 0.44 substitutions per week at close to linear rate. Here we
251 estimate that its rate of spontaneous mutation is 0.1 per genome per cell infection and that
252 viruses with an increased genomic mutation rate can emerge within 15 days of its propagation
253 in cells. Despite the bottlenecks at each passage, new beneficial mutations in spike did spread

254 to high frequencies and considerable convergent evolution was detected across different
255 genomic backgrounds. Overall the results show the remarkable ability of SARS-CoV-2
256 adaptation to new environments in particular via convergent evolution of its spike protein in
257 cells and is fully consistent to its rapid adaptation to different hosts(Gu et al., 2020; Tegally et
258 al., 2020).
259

260 **Materials and Methods**

261

262 **Virus growth and *in vitro* assay**

263 Vero E6 (african green monkey, *Cercopithecus aethiops* kidney epithelial cells, ATCC® CRL
264 1586™) cells were cultured at 37ºC and 5% CO_2 in Minimum Essential Medium (MEM 1X,
265 Gibco®) supplemented with 10% foetal bovine serum (FBS), penicillin (100 units/ml) and
266 streptomycin (100 µg/ml) + fungizone. All work with infectious SARS-CoV-2 strains
267 Portugal/PT0054/2020 and Portugal/PT1136/2020, isolated by the National Institute of Health
268 Doutor Ricardo Jorge (INSA), was done inside a class III microbiological safety cabinet in a
269 containment level 3 facility at the Centre for Vectors and Infectious Diseases Research (INSA).
270 A SARS-CoV-2 stock was produced by infecting Vero E6 cells (freshly grown for 24h) and
271 incubating the cells for 72 h. The culture supernatant was stored in aliquots at -80ºC. The
272 $TCID_{50}$ of viral stock was calculated according to the method of Reed and Muench (Reed and
273 Muench, 1938).
274 The *in vitro* assay was developed for 15 consecutive days in 96-well plates, one for each SARS-
275 CoV-2 strain in study. Two 96-well plates fully inoculated with 50 µl of Vero E6 cells ($2.0x10^4$
276 cells) grown for 24h were infected with 50 µl of SARS-CoV-2 strains Portugal/PT0054/2020
277 and Portugal/PT1136/2020 viral suspension (($2.0x10^3$ viruses) at a multiplicity of infection
278 (MOI) of 0.1. MEM supplemented with 10% FBS, penicillin (100 units/ml) and streptomycin
279 (100 µg/ml) + fungizone was added to each well (50 µl) and the plates were incubated for 24
280 h. Each well had a final volume of 150 µl. During 15 days, with 24 h intervals serial passages
281 of each of the 96 previous day inoculation well were done into a new 96-well plate, using the
282 same procedure (MOI, Vero E6 cells, volumes and concentration) and incubated in the same
283 conditions. At day 15, total nucleic acids were extracted from 100 µl of viral suspension of
284 each well in each plate (96 samples of the day 15th of serial passages of SARS-CoV-2 strain
285 Portugal/PT0054/2020 and 96 samples of the day 15th of serial passages of SARS-CoV-2 strain
286 Portugal/PT1136/2020) using the automated platform NUCLISENS easyMAG (Biomérieux).

287 Confirmation of nucleic acid integrity and rough concentration estimative was made before
288 sequencing experiment by RT-qPCR of 8 random chosen samples from each plate at day 15
289 (CoV-2-D and CoV-2-G) using Novel Coronavirus (2019-nCoV) RT-PCR Detection Kit
290 (Fosun Diagnostics). Samples from inoculation suspension (day 1) were also analyzed. All
291 samples presented values of 7-10 Ct (Cycle threshold).
292
293 **SARS-CoV-2 genome sequencing and bioinformatics analysis**
294 Genome sequencing was performed at INSA following an amplicon-based whole-genome
295 amplification strategy using tiled, multiplexed primers(Quick et al., 2017), according to the
296 ARTIC network protocol (https://artic.network/ncov-2019;
297 https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w) with slight
298 modifications, as previously described(Borges et al., 2020). Briefly, after cDNA synthesis,
299 whole-genome amplification was performed using two separate pools of tiling primers [pools
300 1 and 2; primers version V3 (218 primers) was used for all samples: https://github.com/artic-
301 network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019]. The two pools of
302 multiplexed amplicons were then pooled for each sample, followed by post PCR clean-up and
303 Nextera XT dual-indexed library preparation, according to the manufacturers' instructions.
304 Sequencing libraries were paired-end sequenced (2x150bp) on an Illumina NextSeq 550
305 apparatus, as previously described(Borges et al., 2018). Sequence read quality analysis and
306 mapping was conducted using the bioinformatics pipeline implemented in INSaFLU
307 (https://insaflu.insa.pt/; https://github.com/INSaFLU; https://insaflu.readthedocs.io/en/latest/;
308 as of 10 March 2021), which is a web-based (and also locally installable) platform for
309 amplicon-based next-generation sequencing data analysis(Borges et al., 2018). Briefly, we
310 performed raw reads quality analysis using FastQC v0.11.9
311 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc), followed by quality
312 improvement using Trimmomatic v.0.27
313 (http://www.usadellab.org/cms/index.php?page=trimmomatic; HEADCROP:30 CROP:90
314 SLIDINGWINDOW:5:20 LEADING:3 TRAILING:3 MINLEN:35 TOPHRED33)(Bolger et
315 al., 2014), with reads being conservatively cropped 30bp at both ends for primer clipping.
316 Reference-based mapping was performed against the Wuhan-Hu-1/2019 reference genome
317 sequence (https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3; NC_045512.2)(Wu et al.,
318 2020) using the Burrow-Wheeler Aligner (BWA_MEM) v.0.7.12 (r1039) (http://bio-
319 bwa.sourceforge.net/)(Li and Durbin, 2009) integrated in multisoftware tool Snippy
320 (https://github.com/tseemann/snippy) available in INSaFLU. The obtained median depth of

321 coverage throughout the genome for CoV-2-D and CoV-2-G samples (except two samples
322 excluded due to low coverage) was 4807 (IQR=3969-5242) and 5154 (IQR=4802-5439),
323 respectively. Variant (SNP/indels) calling was performed over BAM files using LoFreq v.2.1.5
324 (*call* mode, including *--call-indels)*(Wilm et al., 2012), with indel qualities being assessed
325 using Dindel(Albers et al., 2011). Mutation frequency analysis was dynamic and contingent on
326 the depth of coverage of each processed site, e.g, minor mutations at "allele" frequencies of
327 10%, 2% and 1% (minimum cut-off used) were validated for sites with depth of coverage of at
328 least 100-fold, 500-fold and 1000-fold, respectively. The median depth coverage per site for
329 all validated mutations in CoV-2-D and CoV-2-G samples was 4219 (IQR=2508-6649) and
330 6424 (IQR=3076-10104), respectively. In order to assess if proximal SNPs and/or indels belong
331 to the same mutational event (and thus, avoid overestimating the mutation rate), we identified
332 all consecutive mutations separated by ≤12 bp. The mutations more likely to represent a single
333 mutation event, i.e., those with similar frequencies (differing by ≤ 2.5%), were further visually
334 inspected using IGV (http://software.broadinstitute.org/software/igv/) to confirm/exclude their
335 co-localization in the same reads. In total, this curation led to the identification 37 SNPs/indels
336 that were collapsed into 13 complex or multi-nucleotide polymorphisms (MNP). The effect of
337 mutations on genes and predicted protein sequences was determined using Ensembl Variant
338 Effect Predictor (VEP) version 103.1 (https://github.com/Ensembl/ensembl-vep; available as a
339 self-contained Docker image)(McLaren et al., 2016). To obtain a refined annotation including
340 all ORF1ab sub-peptides, the GFF3 genome annotation file (relative to the reference Wuhan-
341 Hu-1/2019 genome of SARS-CoV-2, acc. no. NC_045512.2) available in the coronapp
342 COVID-19 genome annotator (http://giorgilab.unibo.it/coronannotator/) (Mercatelli et al.,
343 2020) was adapted to generate an annotation GTF file for input for the *--gtf* parameter. The
344 parameter *--distance* was set to 0. **Supplementary Table 1** summarizes all mutations detected
345 in this study and their distribution across clinical, ancestral cultures and end-point cultured
346 lines (15th passage). SARS-CoV-2 consensus sequences obtained directly from clinical samples
347 for CoV-2-D (Portugal/PT0054/2020) and CoV-2-G (Portugal/PT1136/2020) viruses are
348 available in GISAID under the accession numbers EPI_ISL_421457 and EPI_ISL_511683,
349 respectively. Reads generated at the end of the experimental evolution study were deposited in
350 the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/data/view/PRJEB43731).
351

352 **Simulations of the neutral mutation accumulation**

353 To obtain a non-equilibrium neutral expectation of the frequency spectrum of mutations, we
354 performed forward-simulations to model mutation accumulation using the mutation rate

355 inferred from the experiment. We model a virus with a genome of size L=30000 (~SARS-
356 CoV-2) and a bi-allelic model per site is assumed. An initially isogenic population undergoes
357 15 cycles of growth, mutation and bottleneck (as in the experiment) under a neutral model. The
358 viral life cycle assumed is as follows:

359 1. A clonal population starts with an inoculum size of 2000 viral particles.
360 2. To simulate the replication that occurs within a cell, each viral genome replicates X
361 times, where X is the burst size. We assume an average burst size of 1000 and assume
362 X to be Poisson distributed.
363 3. For each of the replicating genomes we introduce a Poisson number of mutations with
364 mean 0.1 (corresponding to a rate of 3.3×10^{-6} nt^{-1} $cycle^{-1}$, similar to what we inferred
365 from the data). We assume mutations to emerge with uniform probability in the parental
366 genome and we allow for back-mutation. The total population size after replication is
367 on average 2×10^{6}, implying that an average of 2×10^{5} mutations emerge every
368 generation.
369 4. After replication and mutation, the population undergoes 1:1000 dilution (bottleneck)
370 as in the experiment.
371 5. Repeat steps 2-4, 15 times for 100 independent populations.

372 We validated the simulation code by confirming expected outcomes: mutations accumulate
373 linearly over time and the posterior estimation of the mutation rate retrieves the original value
374 (right plot in **Fig. 3b**).
375 After 15 cycles we collect the artificial genomes from 100 independent simulations, and
376 compute the simulated site frequency spectrum as in the experiment. Neutral processes do not
377 explain the high frequency mutations in the observed distribution (**Fig. 3a**). In order to test
378 whether cross-well contamination could justify the right tail of the site frequency distribution,
379 we modified the previous algorithm by introducing migration. At each cycle t, after each
380 bottleneck event, a fraction of viral genomes (m=0.1) is replaced by migrants sampled from a
381 pool of genomes that have undergone *t* cycles of growth. The process of migration under
382 neutrality can lead to an overestimation of the mutation rate but cannot not explain mutations
383 at high frequency (**Fig. 3**). The algorithm was written in R (version 3.6.1) and the results
384 analyzed in RStudio.
385

386 **Data and materials availability**

387 SARS-CoV-2 consensus genome sequences obtained directly from clinical samples for CoV-
388 2-D (Portugal/PT0054/2020) and CoV-2-G (Portugal/PT1136/2020) viruses are available in
389 GISAID under the accession numbers EPI_ISL_421457 and EPI_ISL_511683, respectively.
390 Reads generated throughout the experimental evolution in this study were deposited in the
391 European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/data/view/PRJEB43731).
392 The code for the neutral model was deposited on github/AmiconeM/neutralviralpassage.
393
394
395

396 **References**

397 Albers, C.A., Lunter, G., MacArthur, D.G., McVean, G., Ouwehand, W.H., Durbin, R., 2011.
398 Dindel: accurate indel calls from short-read data. Genome Res 21, 961–973.
399 https://doi.org/10.1101/gr.112326.110
400 Bar-On, Y.M., Flamholz, A., Phillips, R., Milo, R., 2020. SARS-CoV-2 (COVID-19) by the
401 numbers. Elife 9. https://doi.org/10.7554/eLife.57309
402 Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina
403 sequence data. Bioinformatics 30, 2114–2120.
404 https://doi.org/10.1093/bioinformatics/btu170
405 Borges, V., Isidro, J., Cortes-Martins, H., Duarte, S., Vieira, L., Leite, R., Gordo, I., Caetano,
406 C.P., Nunes, B., Sá, R., Oliveira, A., Guiomar, R., Portuguese network for SARS-
407 CoV-2 genomics, Gomes, J.P., 2020. Massive dissemination of a SARS-CoV-2 Spike
408 Y839 variant in Portugal. Emerg Microbes Infect 9, 2488–2496.
409 https://doi.org/10.1080/22221751.2020.1844552
410 Borges, V., Pinheiro, M., Pechirra, P., Guiomar, R., Gomes, J.P., 2018. INSaFLU: an
411 automated open web-based bioinformatics suite "from-reads" for influenza whole-
412 genome-sequencing-based surveillance. Genome Med 10, 46.
413 https://doi.org/10.1186/s13073-018-0555-0
414 Bull, J.J., Badgett, M.R., Wichman, H.A., Huelsenbeck, J.P., Hillis, D.M., Gulati, A., Ho, C.,
415 Molineux, I.J., 1997. Exceptional convergent evolution in a virus. Genetics 147,
416 1497–1507.
417 Chao, L., 1990. Fitness of RNA virus decreased by Muller's ratchet. Nature 348, 454–455.
418 https://doi.org/10.1038/348454a0
419 Charlesworth, B., Charlesworth, D., 2010. Elements of evolutionary genetics. Roberts and
420 Co. Publishers, Greenwood Village, Colo.
421 Drake, J.W., 1993. Rates of spontaneous mutation among RNA viruses. Proc Natl Acad Sci
422 U S A 90, 4171–4175. https://doi.org/10.1073/pnas.90.9.4171
423 Drake, J.W., 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proc
424 Natl Acad Sci U S A 88, 7160–7164. https://doi.org/10.1073/pnas.88.16.7160
425 Drake, J.W., Charlesworth, B., Charlesworth, D., Crow, J.F., 1998. Rates of spontaneous
426 mutation. Genetics 148, 1667–1686.

Duarte, E., Clarke, D., Moya, A., Domingo, E., Holland, J., 1992. Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet. Proc Natl Acad Sci U S A 89, 6015–6019. https://doi.org/10.1073/pnas.89.13.6015

Eskier, D., Suner, A., Oktay, Y., Karakülah, G., 2020. Mutations of SARS-CoV-2 nsp14 exhibit strong association with increased genome-wide mutation load. PeerJ 8, e10181. https://doi.org/10.7717/peerj.10181

Gobeil, S.M.-C., Janowska, K., McDowell, S., Mansouri, K., Parks, R., Manne, K., Stalls, V., Kopp, M.F., Henderson, R., Edwards, R.J., Haynes, B.F., Acharya, P., 2021. D614G Mutation Alters SARS-CoV-2 Spike Conformation and Enhances Protease Cleavage at the S1/S2 Junction. Cell Reports 34, 108630. https://doi.org/10.1016/j.celrep.2020.108630

Gu, H., Chen, Q., Yang, G., He, L., Fan, H., Deng, Y.-Q., Wang, Y., Teng, Y., Zhao, Z., Cui, Y., Li, Yuchang, Li, X.-F., Li, J., Zhang, N.-N., Yang, Xiaolan, Chen, S., Guo, Y., Zhao, G., Wang, X., Luo, D.-Y., Wang, H., Yang, Xiao, Li, Yan, Han, G., He, Y., Zhou, X., Geng, S., Sheng, X., Jiang, S., Sun, S., Qin, C.-F., Zhou, Y., 2020. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. Science 369, 1603–1607. https://doi.org/10.1126/science.abc4730

Jensen, J.D., Lynch, M., 2020. Considering mutational meltdown as a potential SARS-CoV-2 treatment strategy. Heredity (Edinb) 124, 619–620. https://doi.org/10.1038/s41437-020-0314-z

Jensen, J.D., Stikeleather, R.A., Kowalik, T.F., Lynch, M., 2020. Imposed mutational meltdown as an antiviral strategy. Evolution 74, 2549–2559. https://doi.org/10.1111/evo.14107

Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., Hastie, K.M., Parker, M.D., Partridge, D.G., Evans, C.M., Freeman, T.M., de Silva, T.I., Sheffield COVID-19 Genomics Group, McDanal, C., Perez, L.G., Tang, H., Moon-Walker, A., Whelan, S.P., LaBranche, C.C., Saphire, E.O., Montefiori, D.C., 2020. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell 182, 812-827.e19. https://doi.org/10.1016/j.cell.2020.06.043

Lässig, M., Mustonen, V., Walczak, A.M., 2017. Predicting evolution. Nat Ecol Evol 1, 77. https://doi.org/10.1038/s41559-017-0077

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Liu, Z., Zheng, H., Lin, H., Li, M., Yuan, R., Peng, J., Xiong, Q., Sun, J., Li, B., Wu, J., Yi, L., Peng, X., Zhang, H., Zhang, W., Hulswit, R.J.G., Loman, N., Rambaut, A., Ke, C., Bowden, T.A., Pybus, O.G., Lu, J., 2020. Identification of Common Deletions in the Spike Protein of Severe Acute Respiratory Syndrome Coronavirus 2. J Virol 94. https://doi.org/10.1128/JVI.00790-20

Lynch, M., Ackerman, M.S., Gout, J.-F., Long, H., Sung, W., Thomas, W.K., Foster, P.L., 2016. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet 17, 704–714. https://doi.org/10.1038/nrg.2016.104

Makoni, M., 2021. South Africa responds to new SARS-CoV-2 variant. Lancet 397, 267. https://doi.org/10.1016/S0140-6736(21)00144-6

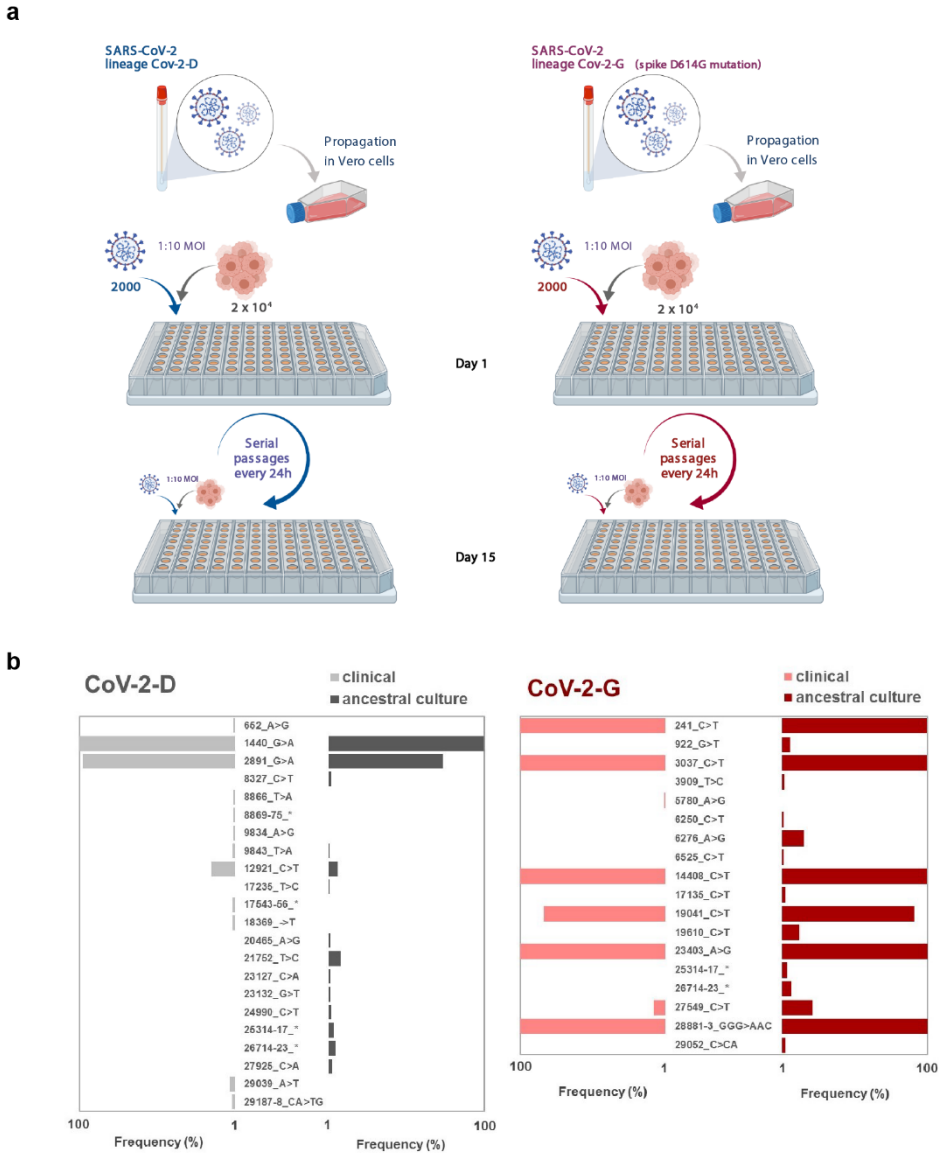
Matyášek, R., Kovařík, A., 2020. Mutation Patterns of Human SARS-CoV-2 and Bat RaTG13 Coronavirus Genomes Are Strongly Biased Towards C>U Transitions, Indicating Rapid Evolution in Their Hosts. Genes (Basel) 11. https://doi.org/10.3390/genes11070761

476 McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P.,
477 Cunningham, F., 2016. The Ensembl Variant Effect Predictor. Genome Biol 17, 122.
478 https://doi.org/10.1186/s13059-016-0974-4
479 Mercatelli, D., Triboli, L., Fornasari, E., Ray, F., Giorgi, F.M., 2020. *Coronapp* : A web
480 application to annotate and monitor SARS-CoV-2 mutations. J Med Virol jmv.26678.
481 https://doi.org/10.1002/jmv.26678
482 Quick, J., Grubaugh, N.D., Pullan, S.T., Claro, I.M., Smith, A.D., Gangavarapu, K., Oliveira,
483 G., Robles-Sikisaka, R., Rogers, T.F., Beutler, N.A., Burton, D.R., Lewis-Ximenez,
484 L.L., de Jesus, J.G., Giovanetti, M., Hill, S.C., Black, A., Bedford, T., Carroll, M.W.,
485 Nunes, M., Alcantara, L.C., Sabino, E.C., Baylis, S.A., Faria, N.R., Loose, M.,
486 Simpson, J.T., Pybus, O.G., Andersen, K.G., Loman, N.J., 2017. Multiplex PCR
487 method for MinION and Illumina sequencing of Zika and other virus genomes
488 directly from clinical samples. Nat Protoc 12, 1261–1276.
489 https://doi.org/10.1038/nprot.2017.066
490 Reed, L.J., Muench, H., 1938. A SIMPLE METHOD OF ESTIMATING FIFTY PER CENT
491 ENDPOINTS12. American Journal of Epidemiology 27, 493–497.
492 https://doi.org/10.1093/oxfordjournals.aje.a118408
493 Rice, A.M., Castillo Morales, A., Ho, A.T., Mordstein, C., Mühlhausen, S., Watson, S.,
494 Cano, L., Young, B., Kudla, G., Hurst, L.D., 2021. Evidence for Strong Mutation Bias
495 toward, and Selection against, U Content in SARS-CoV-2: Implications for Vaccine
496 Design. Mol Biol Evol 38, 67–83. https://doi.org/10.1093/molbev/msaa188
497 Sanjuán, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral mutation
498 rates. J Virol 84, 9733–9748. https://doi.org/10.1128/JVI.00694-10
499 Sasaki, M., Uemura, K., Sato, A., Toba, S., Sanaki, T., Maenaka, K., Hall, W.W., Orba, Y.,
500 Sawa, H., 2021. SARS-CoV-2 variants with mutations at the S1/S2 cleavage site are
501 generated in vitro during propagation in TMPRSS2-deficient cells. PLoS Pathog 17,
502 e1009233. https://doi.org/10.1371/journal.ppat.1009233
503 Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J.,
504 Doolabh, D., Pillay, S., San, E.J., Msomi, N., Mlisana, K., von Gottberg, A., Walaza,
505 S., Allam, M., Ismail, A., Mohale, T., Glass, A.J., Engelbrecht, S., Van Zyl, G.,
506 Preiser, W., Petruccione, F., Sigal, A., Hardie, D., Marais, G., Hsiao, M., Korsman,
507 S., Davies, M.-A., Tyers, L., Mudau, I., York, D., Maslo, C., Goedhals, D.,
508 Abrahams, S., Laguda-Akingba, O., Alisoltani-Dehkordi, A., Godzik, A., Wibmer,
509 C.K., Sewell, B.T., Lourenço, J., Alcantara, L.C.J., Pond, S.L.K., Weaver, S., Martin,
510 D., Lessells, R.J., Bhiman, J.N., Williamson, C., de Oliveira, T., 2020. Emergence
511 and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2
512 (SARS-CoV-2) lineage with multiple spike mutations in South Africa (preprint).
513 Epidemiology. https://doi.org/10.1101/2020.12.21.20248640
514 Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R.L., McDonald, P., Bennett, A.F., Long, A.D.,
515 Gaut, B.S., 2012. The molecular diversity of adaptive convergence. Science 335,
516 457–461. https://doi.org/10.1126/science.1212986
517 van Dorp, L., Richard, D., Tan, C.C.S., Shaw, L.P., Acman, M., Balloux, F., 2020. No
518 evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nat
519 Commun 11, 5986. https://doi.org/10.1038/s41467-020-19818-2
520 Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric,
521 R., Hibberd, M.L., Nagarajan, N., 2012. LoFreq: a sequence-quality aware, ultra-
522 sensitive variant caller for uncovering cell-population heterogeneity from high-
523 throughput sequencing datasets. Nucleic Acids Res 40, 11189–11201.
524 https://doi.org/10.1093/nar/gks918

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E.C., Zhang, Y.-Z., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269. https://doi.org/10.1038/s41586-020-2008-3

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., Tan, W., China Novel Coronavirus Investigating and Research Team, 2020. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med 382, 727–733. https://doi.org/10.1056/NEJMoa2001017

549 **Figures**



550

551 **Fig. 1| Experimental evolution design and ancestor backgrounds. a,** Schematic of the
552 experimental design of the mutation accumulation experiments where two viral backgrounds
553 were propagated in Vero cells (created with BioRender.com). **b,** Mutations and their
554 frequencies in the clinical isolates and their corresponding derived ancestors for the
555 propagations. For the sake of figure simplicity, the complex mutations indicated in the figure
556 with an asterisk correspond to: 8869-75_TTTGCCT>CAAACCA; 17543-
557 56_TGTTCCTCGGAACT>AGTTCCGAGGAACA; 25314-17_GATC>TATG; 26714-
558 23_TTTTGTGCTT>-GTTGTAC--;

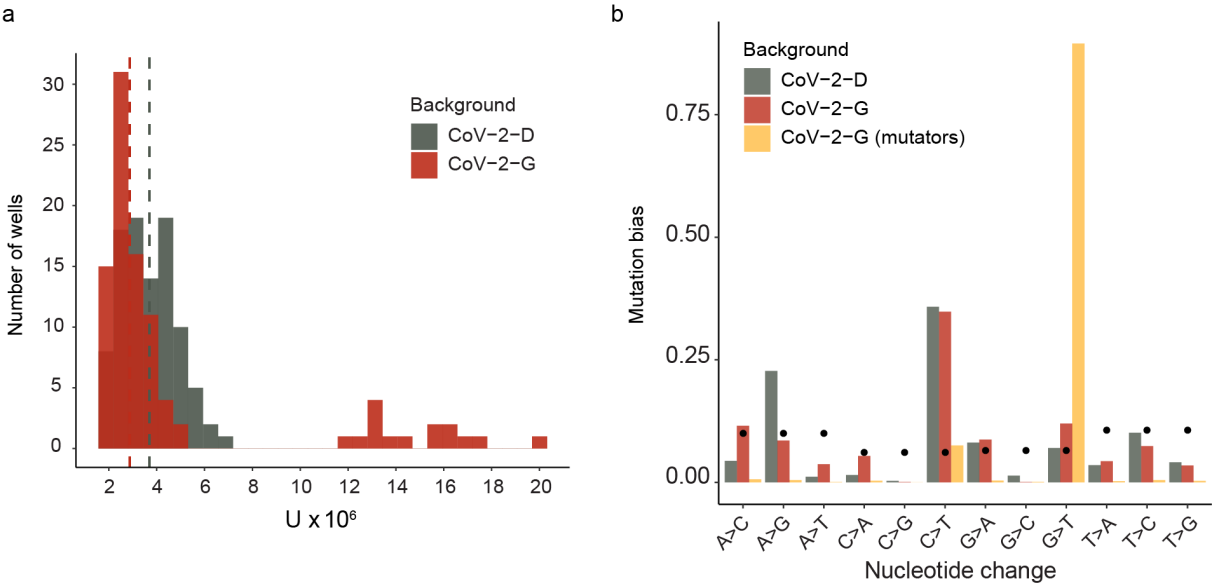


Fig. 2| Mutation rate per genome per infection cycle of SARS-CoV-2. Mutation rate estimate (*U*) of Sars-CoV-2 was calculated from the number of mutations accumulated (*M*) and their frequency (*f*) as: $U = \frac{\sum_i^M f}{P*G}$, where *P* is the number of passages (*P*=15) and G is the genome size (*G*= 29903). **a,** Distribution of *U* for the lines derived from the genomic background CoV-2-D (background 19A), which give rise to an estimated U=3.70x10^{-6} (+ 0.11 x10^{-6}, s.e.m., gray dashed line) and for the lines derived from the genomic background CoV-2-G (background 20B). Some of CoV-2-G lines evolved an elevated mutation rate and thus have mutator viruses. From the lines that did not evolve mutator clones an estimated U=2.9x10^{-6} (+ 0.12 x10^{-6}, s.e.m., red dashed lines) can be derived. The non-mutator lines of CoV-2-D (background 19A) have a higher mutation rate than those of CoV-2-G. **b,** Fraction of substitutions observed in CoV-2-D and CoV-2-G. Dots indicated the expectation given the genome composition under equal mutation probability for each type of nucleotide change.

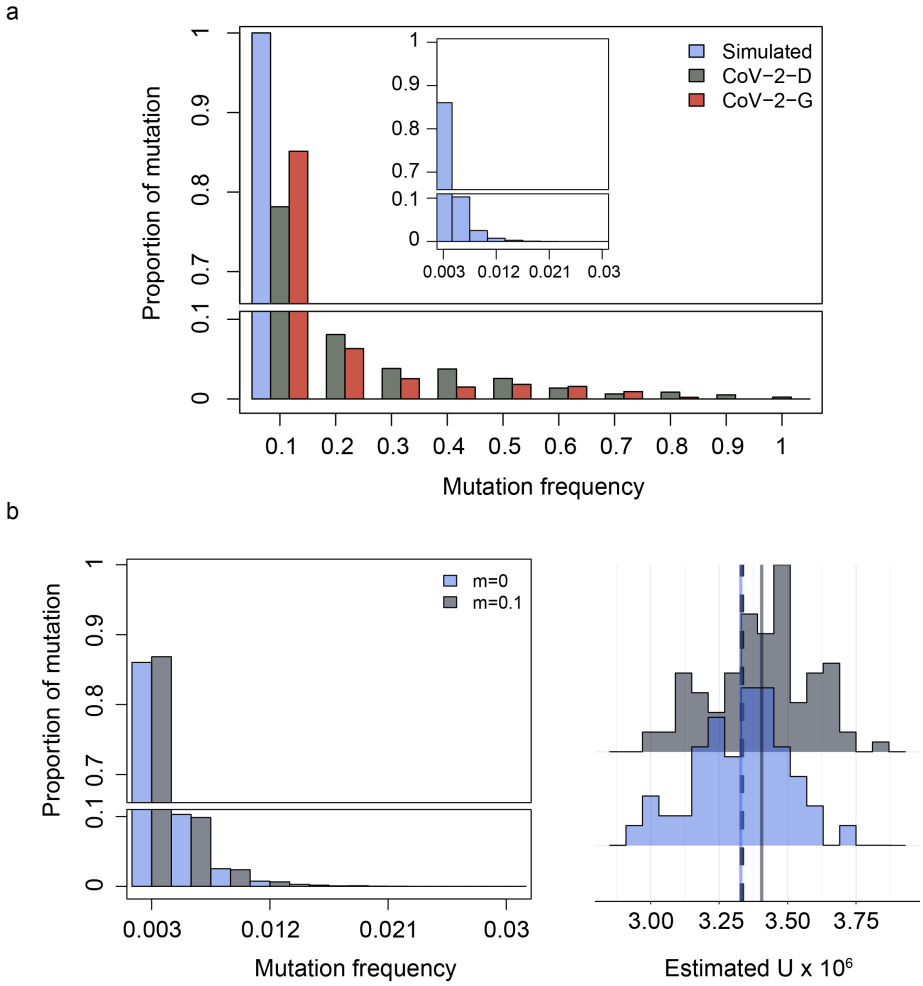
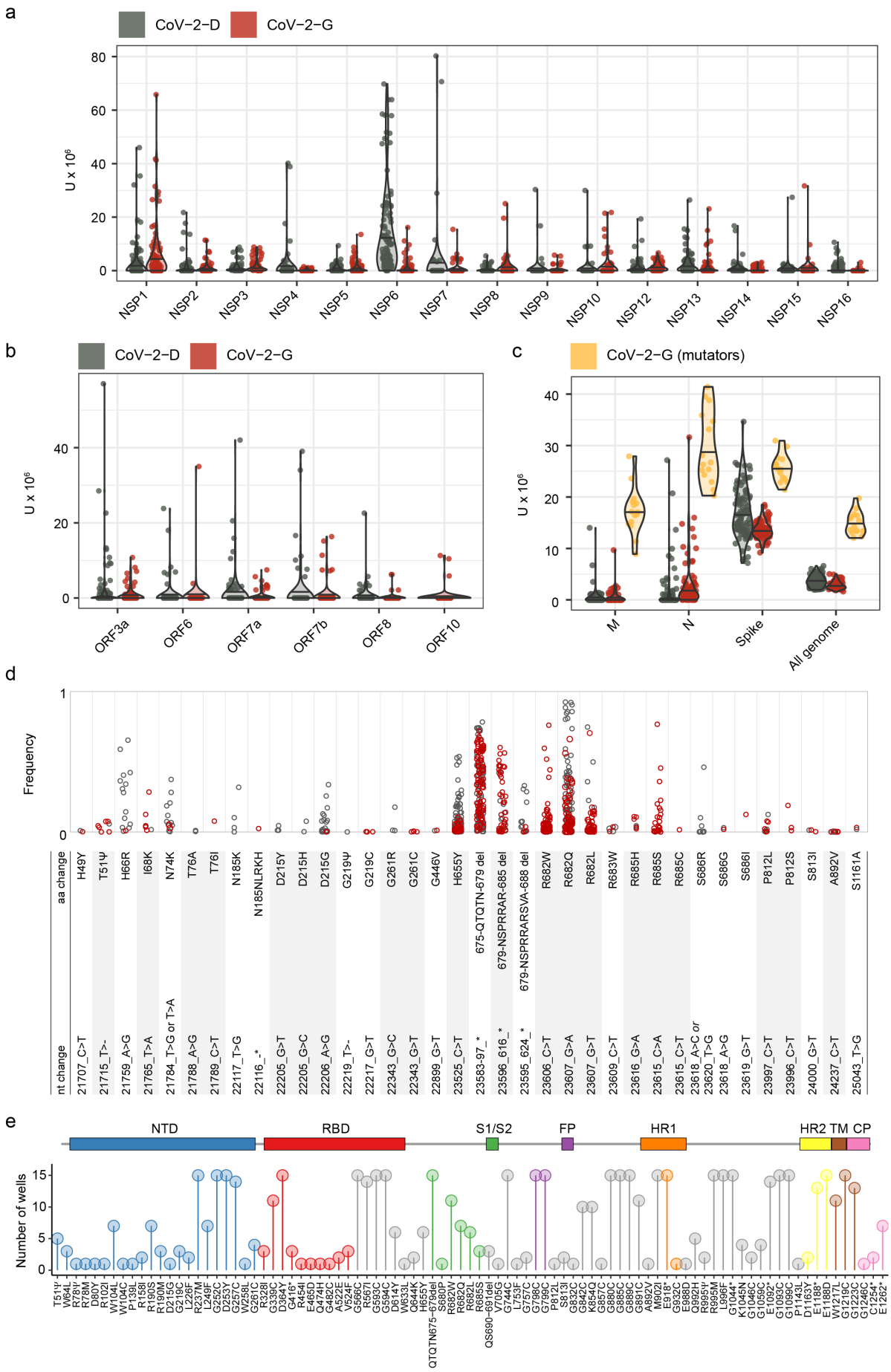


Fig. 3| Site frequency spectrum of the mutations observed after 15 infection cycles. a, The observed site frequency spectrum after 15 cycles of propagation, for CoV-2-D and CoV-2G, and the site frequency spectrum predicted under a neutral model of mutation accumulation. The bump observed at high frequencies in the data is not compatible with the expectation of a neutral model. **b,** The neutral site frequency spectrum expected under 10% cross contamination (m=0.1) between wells at each infection cycle and its effect for estimation of mutation rate. The dashed line represents the simulated *U* while the continuous lines represent the average estimated *U*.


a
CoV−2−D
CoV−2−G
U x 10^6
NSP1
NSP2
NSP3
NSP4
NSP5
NSP6
NSP7
NSP8
NSP9
NSP10
NSP12
NSP13
NSP14
NSP15
NSP16
b
CoV−2−D
CoV−2−G
U x 10^6
ORF3a
ORF6
ORF7a
ORF7b
ORF8
ORF10
c
CoV−2−G (mutators)
U x 10^6
M
N
Spike
All genome
d
Frequency
aa change
nt change
H49Y 21707_C>T
T51Ψ 21715_T>-
H66R 21759_A>G
I68K 21765_T>A
N74K 21784_T>G or T>A
T76A 21788_A>G
T76I 21789_C>T
N185K 22117_T>G
N185NLRKH 22116_*
D215Y 22205_G>T
D215H 22205_G>C
D215G 22206_A>G
G219Ψ 22219_T>-
G219C 22217_G>T
G261R 22343_G>C
G261C 22343_G>T
G446V 22899_G>T
H655Y 23525_C>T
675-QTQTN-679 del 23583-97_*
679-NSPRRAR-685 del 23596_616_*
679-NSPRRARSVA-688 del 23595_624_*
R682W 23606_C>T
R682Q 23607_G>A
R682L 23607_G>T
R683W 23609_C>T
R685H 23616_G>A
R685S 23615_C>A
R685C 23615_C>T
S686R 23618_A>C or 23620_T>G
S686G 23618_A>G
S686I 23619_G>T
P812L 23997_C>T
P812S 23996_C>T
S813I 24000_G>T
A892V 24237_C>T
S1161A 25043_T>G
e
NTD
RBD
S1/S2
FP
HR1
HR2 TM CP
Number of wells

597

598 **Fig. 4 Heterogeneity of mutation rates across genes and selection on spike. a, b, c,**
599 Mutation rates across the different gene regions for CoV-2-D and CoV-2-G. Mutation rate
600 estimates were calculated from the number of mutations accumulated and their frequency.
601 Mutation rate of the spike (U_s) for the lines derived from the genomic background CoV-2-D
602 gives rise to an estimated U_s=17.1x10^{-6} ($\pm$ 1.2 x10^{-6}, s.e.m.), ~5 times higher than the
603 genomic average. **d,** Evidence for selection on the spike gene at the level of nucleotide site
604 and amino-acid. Selection can be inferred by the level of convergent evolution observed
605 across the two genomic backgrounds (see **Supplementary Table 1** for detailed annotations).
606 **e,** Non-synonymous mutations on the spike detected in the populations where the mutators
607 were observed (the number of wells where the mutations were observed are in the Y-axis).
608 The color annotation represents the N-terminal domain (NTD, 14–305), the receptor-binding
609 domain (RBD, 319–541), the cleavage site (S1/S2, 669-688), the fusion peptide (FP, 788–
610 806), the heptapeptide repeat sequences (HR1, 912–984 and HR2, 1163–1213), the TM
611 domain (1213–1237), and cytoplasm domain (CP, 1237–1273).
612
613
614 **Acknowledgments:** We would like to thank the personnel of the IGC Genomic Facility for
615 their assistance.
616
617 **Funding:** M.A. was supported by "Fundação para a Ciência e Tecnologia" (FCT),
618 fellowships PD/BD/138735/2018, respectively. Research was supported by FCT Project
619 PTDC/BIA-EVL/31528/2017 to I.G. and by funds from Portuguese NIH.
620
621 **Author contributions:** IG, MJA and JPG designed the project. MJA and LZZ performed the
622 culture and RNA extraction experiments. VB and JI performed the pre-sequencing wet-lab
623 procedures, bioinformatic analysis and data analysis. SD and LV performed and supervised the
624 wet-lab sequencing procedures. IG, MA performed the data analysis and the simulations. MJA,
625 LV and JPG provided materials and reagents. IG wrote the initial draft of manuscript. VB,
626 MJA, MA and JI contributed equally to this work. All authors contributed in the final writing
627 of the manuscript and gave final approval for publication.

Ethical statement: The Portuguese NIH is authorized by the Portuguese Authorities' (General-Directorate of Health and the Authority for Working Conditions) to handle and propagate Risk group 2 and 3 microorganisms. All culture procedures were performed inside a class III microbiological safety cabinet in a containment level 3 facility. This study is covered by the ethical approval issued by the Ethical Committee ("Comissão de Ética para a Saúde") of the Portuguese National Institute of Health.

Competing interests: The authors declare no competing interests.

Additional information: Supplementary information is available for this paper. Correspondence and requests for materials should be addressed to igordo@igc.gulbenkian.pt; j.paulo.gomes@insa.min-saude.pt.