Biochemical and mathematical lessons from the evolution of the SARS- CoV-2 virus: paths for novel antiviral warfare

Nicolas Cluzel, Amaury Lambert, Yvon Maday, Gabriel Turinici, Antoine Danchin

1 # Biochemical and mathematical lessons from the evolution of the SARS-
2 # CoV-2 virus: paths for novel antiviral warfare

3

4 Nicolas Cluzel[1], Amaury Lambert[2,3], Yvon Maday[1], Gabriel Turinici[4], Antoine Danchin[5,6,*]

5

6 1. Tremplin Carnot SMILES, 4 Place Jussieu, 75005 Paris, France

7 2. Laboratoire de Probabilités, Statistique & Modélisation (LPSM), Sorbonne Université, Université
8 de Paris, CNRS UMR8001, 4 place Jussieu, 75005 Paris, France

9 3. Centre Interdisciplinaire de Recherche en Biologie (CIRB), Collège de France, CNRS
10 UMR7241, INSERM U1050, PSL Research University, 11 place Marcelin Berthelot, 75005 Paris,
11 France

12 4. Ceremade, Université Paris Dauphine – PSL

13 5. Kodikos Labs / Stellate Therapeutics, Institut Cochin, 24 rue du Faubourg Saint-Jacques, 75014
14 Paris, France

15 6. School of Biomedical Sciences, Li KaShing Faculty of Medicine, Hong Kong University, 21
16 Sassoon Road, Pokfulam, SAR Hong Kong, China

17

18

19 For correspondence :

20 E-mail : antoine.danchin@normalesup.org

21 Tel: +331 4441 2551; Fax: +331 4441 2559

22

25

26

27 **Abstract**

28 **I**n the fight against the spread of COVID-19 the emphasis is on vaccination or on reactivating

29 existing drugs used for other purposes. The tight links that necessarily exist between the virus as it

30 multiplies and the metabolism of its host are systematically ignored. Here we show that the

31 metabolism of all cells is coordinated by the availability of a core building block of the cell's

32 genome, cytidine triphosphate (CTP). This metabolite is also the key to the synthesis of the viral

33 envelope and to the translation of its genome into proteins. This unique role explains why evolution

34 has led to the early emergence in animals of an antiviral immunity enzyme, viperin, that

35 synthesizes a toxic analogue of CTP. The constraints arising from this dependency guide the

36 evolution of the virus. With this in mind, we explored the real-time experiment taking place before

37 our eyes using probabilistic modelling approaches to the molecular evolution of the virus. We have

38 thus followed, almost on a daily basis, the evolution of the composition of the viral genome to link it

39 to the progeny produced over time, particularly in the form of blooms that sparked a firework of

40 viral mutations. Some of those certainly increase the propagation of the virus. This led us to make

41 out the critical role in this evolution of several proteins of the virus, such as its nucleocapsid N, and

42 more generally to begin to understand how the virus ties up the host metabolism to its own benefit.

43 A way for the virus to escape CTP-dependent control in cells would be to infect cells that are not

44 expected to grow, such as neurons. This may account for unexpected body sites of viral

45 development in the present epidemic.

46

Introduction

48 The development of the COVID-19 pandemic is being explored in a myriad of articles. Despite this

49 abundance, and because of our anthropocentrism, it is exceptional that these studies focus on the

50 virus' standpoint. Of course, much work is looking into the details of the composition and structure

51 of the SARS-CoV-2 virus genome, the proteins it codes for and its animal-infecting relatives.

52 However, there are very few major studies on how the virus exploits the metabolism of its host's

53 cells. The urgent necessity to contain the disease led investigators to emphasize vaccination or,

54 more generally, the involvement of the host's immune system. It is well known, alas, that while it

55 has sometimes been relatively easy to generate a vaccine that is both effective and harmless

56 against a widespread disease, the opposite is also true. There are still very serious and very

57 common diseases for which there is no vaccination. Vaccinating effectively assumes, in particular,

58 that the progeny of a pathogen remains the same long enough to prevent escape of the immune

59 response triggered by the vaccine. Coronaviruses are viruses made up of a long genome and an

60 envelope. The length of the genome could have led to a very high mutation rate, but these viruses,

61 thus avoiding the universal constraint of Muller's ratchet - see **Box** - have recruited a specific

62 function that proofreads and corrects replication errors [1]. This means that, while coronaviruses do

63 indeed tend to produce genetic variants over time, the number of these variants remains quite low.

64 This mutation rate may appear very limited, but the sheer number of viral particles generated

65 during an infection is enormous, while the human population currently recognized as infected

66 exceeds twenty million people. It follows that the mutation rate per nucleotide - of course very

67 heterogeneous due to the selection pressure on certain locations in the genome - is around 8×10^{-4}

68 changes per site per year [2].

69 Here, this situation was placed in the perspective of the fundamental theorem of natural selection

70 proposed by Fisher, which links the evolution of environmental fitness and genetic variance [3]. We

71 wished to use the marks left by the evolution of the virus' fitness - observed in the form of genomic

72 sequences - in the presence of the biochemical constraints that bias the choices available for

73 evolution. We had to take into account, however, that the terms of the problem are not as explicit

74 as one might have wished: fitness is not known, nor are the time markers (estimated from

75 phylogenetic trees or simply taken as physical time) and the frequency of certain strains in the

76 phylogenetic trees may be less due to natural selection than to heterogeneity in sampling and

77 sequencing depth. This motivated our use of procedures that are robust enough to cope with these

78 uncertainties. Nevertheless, the advantage of such an analysis is that it allowed us to propose

79 anticipations for the evolution of the virus. It is therefore an explicit means of feeding

80 epidemiological or clinical models with relevant observations.

81 In this context, it seemed to us of great interest to explore the details of how SARS-CoV-2 mutated

82 over time, in the various places where COVID-19 has spread, highlighting relevant descents in

83 relation with the host metabolism. This should allow us to anticipate some of the future of the virus'

84 progeny, with important consequences for control of the disease. The analysis of the constraints

85 that govern access to the metabolism of the nucleotides that make up the virus genome has shown

86 us that the content of cytosine (C) in its genome is subjected to strong negative pressure, leading

87 to systematic depletion, over time, in cytosine monophosphate [4]. This bias has long been

88 believed to result from a major causal effect of the "editing" of the C content of the genome by the

89 family of APOBEC deaminating enzymes [5,6]. We now know that it is the organization of the

90 metabolism of pyrimidines in animal cells, and more particularly of cytosine triphosphate [7], which

91 drives the corresponding pressure on evolution (**Figure 1**).
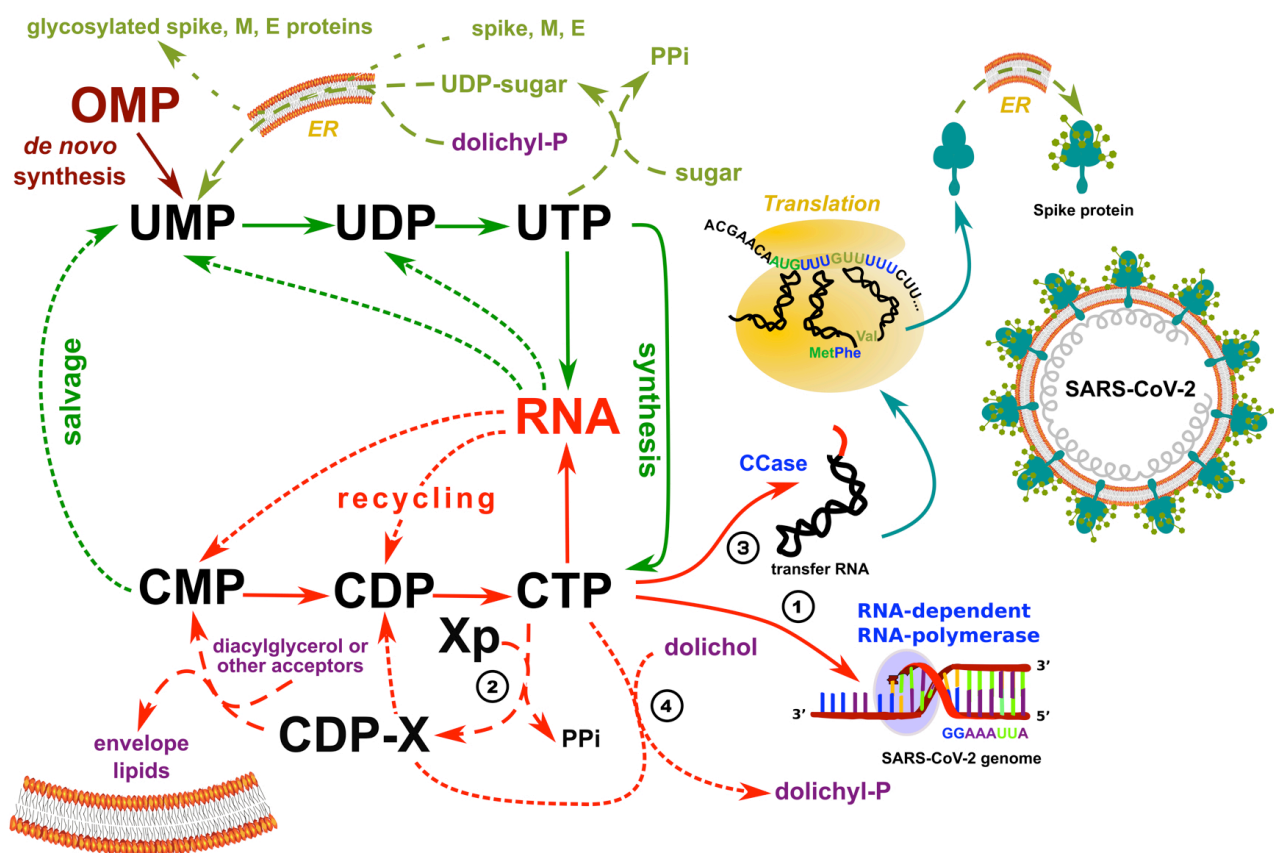
93 **Figure 1. CTP controls all crucial metabolic steps required to build up a functional SARS-**
94 **CoV-2 virus.** 1/ CTP is a precursor of the virus genome; 2/ the lipids of its envelope derive from
95 cytosine-based liponucleotide precursors; 3/ all transfer RNA molecules produced by the host must
96 be maturated to a form ending in a CCA triplet at their 3'OH end; and 4/ post-translational
97 glycosylation of viral proteins, in particular its spike protein require a dolichyl-phosphate anchor in
98 the endoplasmic reticulum (ER) and dolichol kinase is specifically dependent on CTP. See text and
99 reference [7] for details.

100 Indeed, due to the extreme asymmetry of the replication of the virus - which replicates 50 to 100

101 times from its complementary template [8] - a genome editing effect of these highly context-

102 dependent enzymes would only be significant when a C into U is modified on the negative RNA

103 template, which would lead to a major enrichment in A of the viral genome, or possibly from a U →
104 C transition due to another class of deaminating enzymes acting on double stranded RNA, ADAR,
105 that deaminates adenine into inosine [9]. Furthermore, both APOBEC and ADAR are highly specific
106 enzymes and this hardly fits with the widespread C → U transitions that we keep observing as the
107 virus evolves. Here, we have focused on the dynamics of the loss of C in the genome, and sought
108 for the locations and the causes of changes in this driving force. In the first paragraph, we
109 summarized the metabolic reasons accounting for this remarkable phenomenon. Subsequently, in
110 the body of the article, we showed that the constraint on the C content of the genome leads to
111 specific descents which can be used to reveal the existence of important functions of the virus as
112 well as the role of the host's response.

113 **A universal metabolic requisite, the biosynthesis of cytidine triphosphate (CTP), guides the**
114 **evolution of the virus**

115 What do we know about the synthesis of the building blocks that allow the generation of a viral
116 particle (a virion)? During a viral infection cells usually stop multiplying. All their resources are
117 quickly diverted in favour of the multiplication of the virus. Yet, growth is a universal property of life.
118 This means that, almost always - differentiated neurons are an exception - the cell's metabolism
119 that the virus faces is organized to allow cell growth as soon as the opportunity to multiply arises.
120 The moment it infects a cell - again, with the exception of those that do not multiply - any virus will
121 therefore have to manage the metabolic pressure that organizes the availability of the building
122 blocks necessary for its construction. In our usual physical space (three-dimensional), growing
123 introduces an inevitable constraint. The cell must put together the growth of its cytoplasm (three-
124 dimensional, therefore), that of the membrane that encloses it (two-dimensional) and that of its
125 genome (one-dimensional, because nucleic acids are linear polymers). However, it is a common
126 metabolism, developed mainly in the cytoplasm, which produces the building materials needed to
127 build up these three major compartments. So, here we have a question similar to the one asked by
128 economists when they raise the question of "non-homothetic" growth [10]. Unfortunately, because
129 life developed from a primitive metabolism in several stages over 3.5 billion years [10], we might
130 fear that many organisms had found an idiosyncratic solution to this constraint, as often witnessed
131 in the huge diversity of life forms. Unexpectedly, it appears that the solution to this quandary is
132 universal: a single metabolite, the nucleotide cytidine triphosphate (CTP), has been recruited to
133 this purpose [4,7].

134 The key role of CTP appears in four essential places in cellular metabolism, and these places are
135 essential for the formation of new virions. 1/ It is the immediate precursor of one of the four
136 nucleotides forming the genome of the virus; 2/ CTP is required for the synthesis of liponucleotide
137 precursors of the viral envelope; 3/ human transfer RNAs are synthesized from 415 genes which
138 do not encode their 3'OH-CCA terminal end - this sequence is synthesized from CTP by a specific

139 nucleotidyltransferase [12]; and finally 4/ the "decoration" of proteins by complex glycosylations is

140 performed in parallel with their translation in the endoplasmic reticulum (ER) *via* the anchoring of

141 substrates by dolichyl-phosphate, produced by a kinase which uses CTP, not ATP, as its phosphate

142 donor [13]. In addition, intermediate metabolism is based on an original organization of the

143 metabolism of pyrimidines, which systematically recycles and salvages them *via* uridine

144 triphosphate (UTP) which makes CTP a pivot metabolite and limits considerably its availability

145 (**Figure 1**). As a result, accidental replication errors will tend to replace cytosine with uracil in the

146 genome.

147 **General evolution of the SARS-CoV-2 virus**

148 Using the available sequence data gathered in the SARS-CoV-2 GISAID database

149 (https://www.gisaid.org) we have, like others [14,15], reconstituted a phylogenetic tree of the

150 evolution of the virus. As the sequences of each viral genome, as well as the date of identification

151 of these sequences are known with fairly great precision, this tree makes it possible to explore the

152 orderly lineage of the mutations which appear over time. In particular, unless we can suspect a

153 recombination event due to the infection of the same patient by two or more viruses, when two

154 identical mutations appear in separate branches of the tree, we can assume that this is the result

155 of evolutionary convergence [16]. The reasons for this convergence are discussed on a case by

156 case basis when analysing each relevant mutation. A second observation, which needs to be put in

157 perspective (see below), is that the shape of the tree is not at all homogeneous. We noticed indeed

158 the presence of "blooms" where, at a particular node of the tree, a large number of branches

159 appear, demonstrating an "explosive" appearance of new mutations (**Figure 2**). We have therefore

160 devised a statistical approach that allowed us to characterize them explicitly.
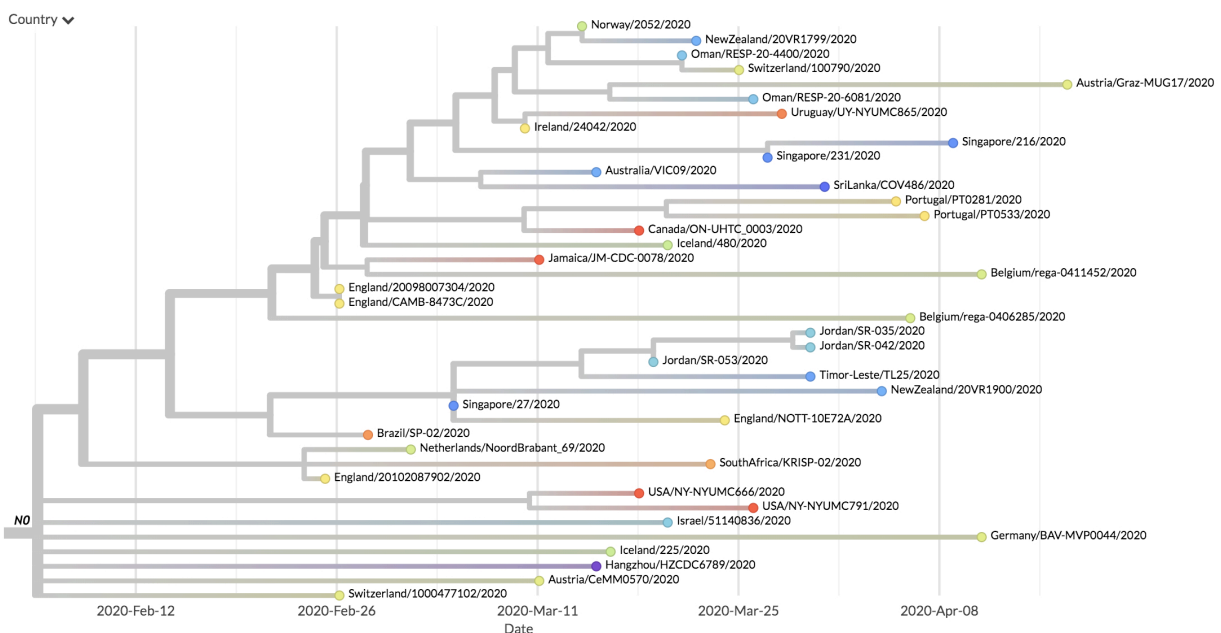
162 **Figure 2. An example of bloom detected by our statistical approach.** At node N0, there are 25
163 different states in the 40 samples of the subtree and a high number of branches. This behaviour
164 differs significantly from that of the other sub-trees.

165 The causes of these blooms are multiple, but the adaptation of important viral functions can be at

166 their origin, and we retained a few cases of this kind for further discussion (see **Materials and**

167 **Methods** for the statistical definition of blooms).

Description and analysis of the evolution of the C content of the genome

169 Generally speaking, the coronavirus genome tends to evolve by adapting its C content to the

170 metabolism of its host. More specifically SARS-CoV-2 evolves towards forms less rich in C as the

171 epidemic develops [7]. However, this development is not homogeneous.

172 In the two data sets of interest, 77% of the transitions between pyrimidines are represented by

173 transitions from cytosine to uracil. These transitions represent 48% of all substitutions identified in

174 the first set (respectively, 49% in the second). An important imbalance can also be noted at the

175 level of the transversions, knowing that more than 73% of those pertain to a substitution from

176 purine to pyrimidine in the first set (respectively, 74%). However, only 20% of these 73% lead to

177 the occurrnce of cytosine (respectively, 17%), indicating once again a tendency to favour the

178 generation of uracil, thus demonstrating that the major constraint of the mutagenic process is the

179 availability of each one of the nucleoside triphosphates in the cell. This inhomogeneity is also

180 salient at the tree level. At the level of branch B4 (20% of the samples), the tendency is strongly

181 marked to lose less C as compared to the rest of the tree (**Figure 3**).
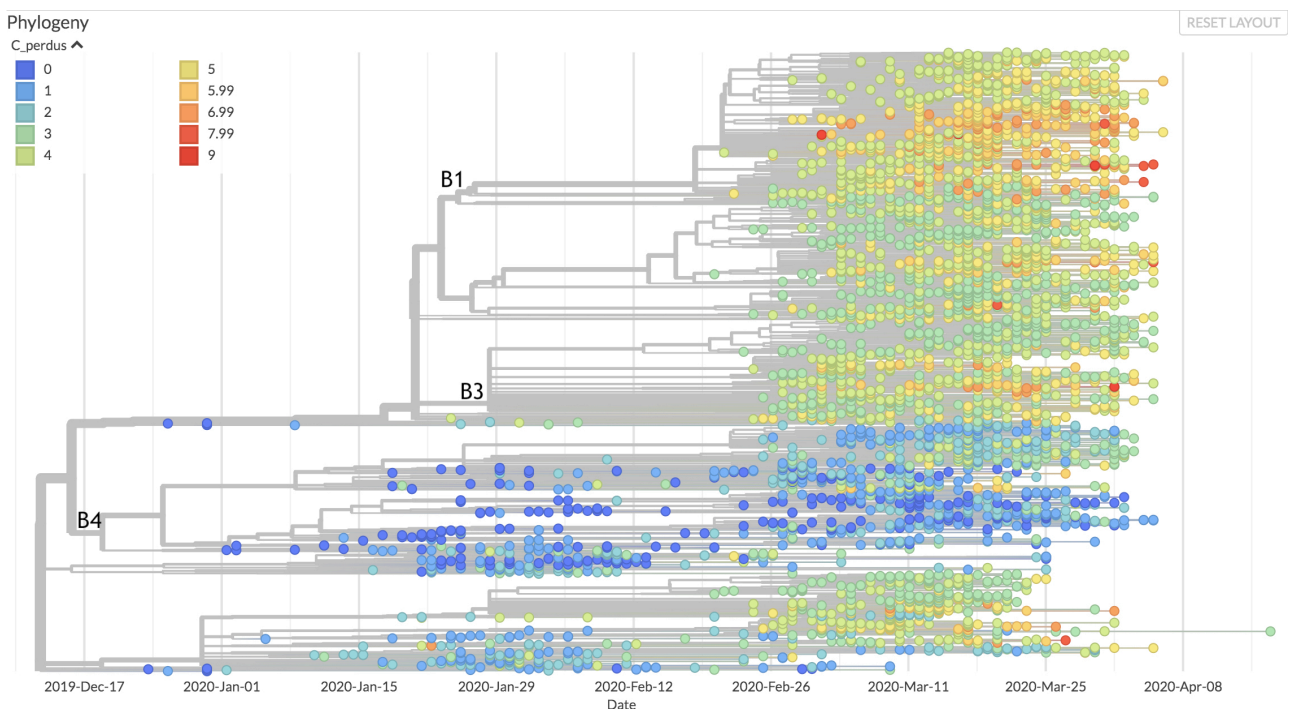
183 **Figure 3: Heat map of C losses from the original sequence.** Branches 1 and 4 can be readily
184 discriminated by their extreme values.

185 Interestingly, this branch is also the one that comprises on average the strains with the least
186 divergence from the original strain of the virus. By contrast, in branch B1, the loss of C looks larger.
187 The rate of virus mutation also seems to be accelerating in this branch, with a rate of transversions
188 20% higher than the rest of the tree (and also higher transition rates, but in more anecdotal
189 proportions). Finally, for branch B3, the main site of blooms, a 29% decrease in the transition rate
190 of pyrimidines and a 30% decrease in the rate of purines compared to the rest of the tree is
191 noteworthy.

192 This inhomogeneity can be the consequence of many constraints:

193 1/ The very structure of the genome, which must fold into a compact capsid envelope requires
194 certain regions to maintain the presence of specific C residues. This is the case of the regions
195 which control the origin of replication [8] or transcription, AACGAAC, for example [17]. In the case
196 of the translated regions, the pressure on the presence of C varies depending on its position in the
197 codon trinucleotides. When C is located at the first position of a codon, it is used to input arginine,
198 glutamine, histidine, leucine or proline into proteins. Histidine and glutamine are coded in two
199 codon families, discussed below. For arginine, the selection pressure is lower because the CGN
200 codons can be replaced by AGR codons - we used here the IUPAC convention for labelling
201 nucleotides or aminoacids, e.g. N is for aNy, R for puRine, etc.
202 (https://www.bioinformatics.org/sms/iupac.html). The selection pressure on the leucine content is
203 also lower, since in addition to the CUN codons, this amino acid can be input using the UUR
204 codons. In the second codon position, C is again used to code for proline, but also threonine
205 (ACN), alanine (GCN) and serine (UCN). Again, the latter amino acid escapes a large part of the
206 constraint imposed by the availability of C because it can also use the AGY codons. Finally, the
207 third position of the codons is much less constrained because it can be replaced by U but also by A
208 or G in the families with four codons (alanine, proline, threonine, valine). The two codon families
209 UGY, AGY and NAY are discriminated along a pyrimidine / purine axis. A pyrimidine is used to
210 maintain the same nature of the coded residue, as the codon uses a U or C as the 3 'end
211 (aspartate, asparagine, cysteine, histidine and tyrosine). Finally, isoleucine is coded by three
212 codons (AUH), and ending in U or C is taken into account by relevant tRNAs [18];

213 2/ The function of the virus proteins can impose the presence of certain amino acids in their
214 sequence. For example, the proline residue encoded by the CCN codons is not strictly an amino
215 acid, but is essential for the folding of key domains of viral proteins [19];

216 3 / Further stressing the importance of CTP, during evolution, innate antiviral immunity recruited the
217 activity of an enzyme, viperin, which modifies CTP into a form toxic to the development of the virus,

218 3'-deoxy-3',4'-didehydro-CTP (ddhCTP) [20]. An interesting consequence of this pathway is that

219 decreasing the C content of the genome will allow the virus replication process to be less sensitive

220 to the presence of this nucleobase. It follows that, during the transfer of a virus relatively rich in C

221 from an animal host to human beings, the evolution towards the loss of C may be transiently

222 concomitant with an increase in its pathogenicity. In the long term, however, the loss of C severely

223 restricts the evolutionary landscape of the virus and most likely will tend to its attenuation [21].

224 **Examples of correlations allowing us to propose a function for viral proteins**

225 Thousands of mutations have been identified at this date. It is possible to follow their emergence

226 along the tree of its phylogenetic evolution of the virus and then highlight some interesting features

227 that may allow us to anticipate some of its future.

228 *Mutations leading to an early translation termination*

229 Mutations leading to premature termination of the virus protein synthesis are expected to appear

230 with high frequency. In the present context, this is all the more likely because the translation

231 termination codons UAA, UAG and UGA do not contain C, and are therefore favoured by the

232 disappearance of this nucleotide. Since most of these mutations lead to non-functional

233 polypeptides, it is generally probable that the affected viruses do not give rise to a significant

234 progeny. It follows that when these mutations are observed - and that they do not result from

235 sequencing errors - they indicate that the role of the truncated protein corresponds to a function

236 which is not critical, or that the protein has remained functional at a sufficient level to allow virus

237 reproduction. However, a few observations allowed us to offer an explanation for the fact that the

238 viruses in question may have survived. Here are three examples which reveal interesting features

239 of the virus.

240 Example 1: In a strain from Iceland, the succession of mutations G1440A (Gly392Asp, protein

241 Nsp2) and G2891A (Ala876Thr, ubiquitin-like domain of protein Nsp3) is now present in multiple

242 world locations [22]. This sequence ends up with C27661U (which modifies amino acid Gln90 into

243 a premature translation end, near the carboxy-terminal end of protein Orf7a). This viral protein is

244 found in the endoplasmic reticulum, the Golgi apparatus and the perinuclear space [23]. Several

245 variants have been identified in the course of the epidemic [24]. Remarkably, several deletions

246 have been isolated in the gene, which suggests that the function of this region is not essential [25].

247 However, we noticed that many of these mutations, as the one discussed here, keep the small

248 hydrophobic protein Orf7b gene intact, downstream of Orf7a. This very small protein is present in

249 the Golgi apparatus and is also found in the purified virus [26]. It must be noticed that it is

250 synthesized in vivo *via* a frameshift that spans the termination codon of the Orf7a frame (...GAA

251 TGA TT... becomes ...GA ATG ATT...). This can be interpreted as a conflict in this region between

252 translation of Orf7a and Orf7b, creating a cost / benefit dilemma for the expression of either one of

253   these proteins. Hence it will be important to monitor the future descent of the virus in this region as

254   it may result in interesting attenuated forms.

255   Example 2: Another succession of mutations that leads to premature translation termination of a

256   viral protein begins with G11083U (protein Nsp6, Leu37Phe). This mutation is now widely

257   distributed worldwide. It is likely to induce a more stable binding of the protein to the ER, possibly

258   favouring coronavirus infection by compromising delivery of viral components to lysosomes for

259   degradation [27]; then we have G1397A (Nsp2, Val378Ile), also likely to favour virus propagation

260   [28]; followed by G29742U (3'UTR of the virus), and U28688C (synonymous); subsequently, we

261   have the couple of mutations C884U (Nsp2 again, Arg207Cys [28]) and G8653U (Nsp4, essential

262   for envelope assembly [29]. The corresponding change (Met2796Ile) is located at the border of the

263   ER lumenal domain of the protein. It is known that, in order to function properly, the ER requires

264   the presence of oxygen [30], and reactive oxygen species (ROS) are associated to misfolding of

265   proteins in this compartment. Nsp4 has a number of cysteine residues, prone to be oxidized. The

266   role of methionine in the parent might be to act as a buffer against ROS, so that the mutant would

267   be slightly attenuated). These mutations are followed by A19073G (in the methylase domain of

268   protein Nsp14, Asp1869Gly, a position that already evolved from SARS-CoV-1 [31], hence likely to

269   be more or less neutral), then the couple with the mutation resulting in end of translation:

270   G27915U, Gly8 to end of translation at the N-terminus of Orf8 and C29077U (synonymous); the

271   succession ends with the couple of mutations leading to synonymous changes C19186U and

272   G23608U. This region of SARS-related coronaviruses is hypervariable. It changes during the

273   course of epidemics, showing that it is subject to ongoing selection pressure, sometimes producing

274   two peptides Orf8a and Orf8b [32]. It corresponds to proteins expressed at the end of the infection

275   cycle. It will be important to monitor the way they function in the course of the evolution of virulence

276   of the virus. This displays a branching that appeared in four different countries and in seven

277   samples, spanning six weeks between the first and the last mutation.

278   Example 3: Here we have a succession of mutations that begin within the 5'end of the virus

279   genome, C241U, followed by mutation C14408U (Pro314Leu) at the end of a zinc finger in

280   replicase Nsp12, which appears in many branches of the evolution tree of the virus. It is discussed

281   in details below (origin of blooms). This mutation is followed by A23403G (Asp614Gly) a widely

282   spread mutation of the spike protein (also discussed below), C3037U (synonymous), mutation

283   G25563U (Gln57His) in Orf3a forming potassium channels is supposed to negatively interfere with

284   the function of the protein [33], C1059U (Thr265Ile) in protein Nsp2, discussed previously, and the

285   triplet G4181A (Ala1305Thr) in the SUD-N domain of protease Nsp3, then mutations G4285U

286   (Glu1340Asp), and G28209U which results in an end of translation at glutamate 106 of protein

287   Orf8. As discussed previously, many mutations, including deletions in Orf8, were frequently

288   observed. This is again an indication that evolution of this regions should be carefully monitored to

289 look for attenuated forms of the virus. This particular mutation to an end of translation is significant

290 as it was found in a sample from Croatia, another one from Thailand, on two significantly separated

291 branches and with one month difference. The sequence of mutations here corresponds to the

292 Thailand sample.

293   *Reversal of the tendency of the viral genome to lose its cytosine residues*

294 We have here retained two examples of a situation where, from an upstream branching point in the

295 evolution tree, it appears that the descendants of the virus stop losing their cytosines, and may

296 even tend to regain them. These examples are as follows (**Figure 4**).
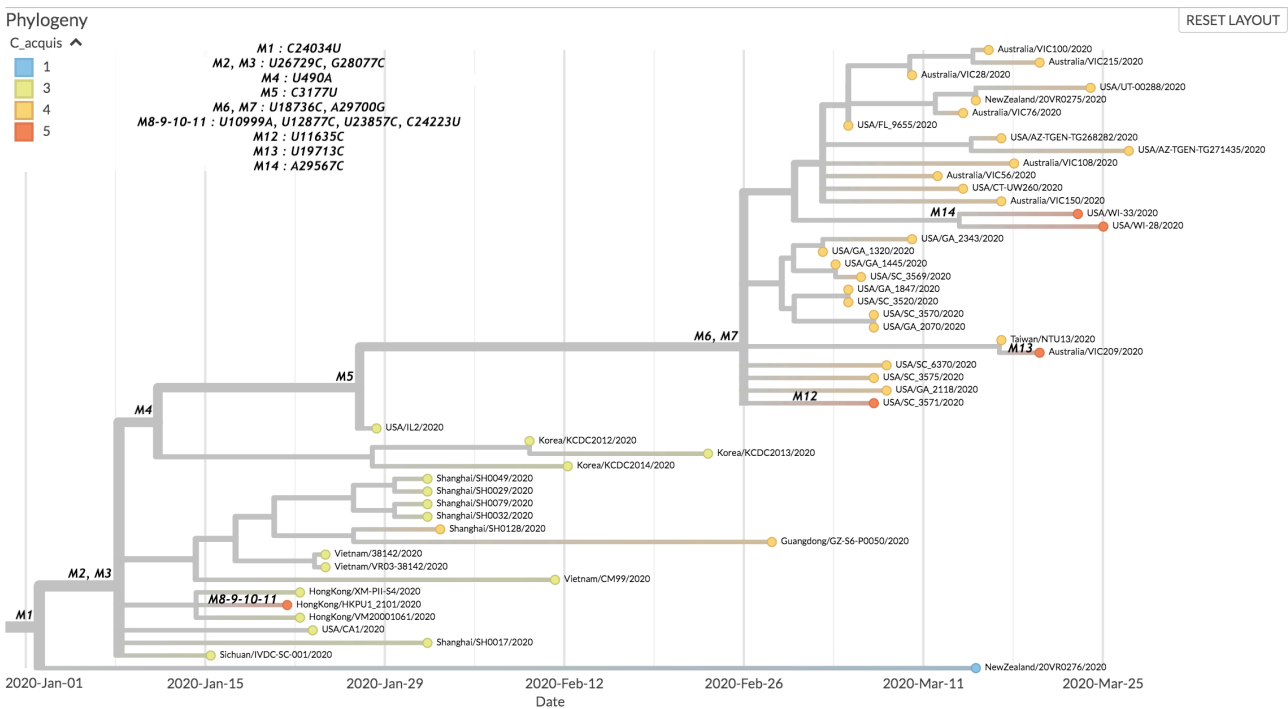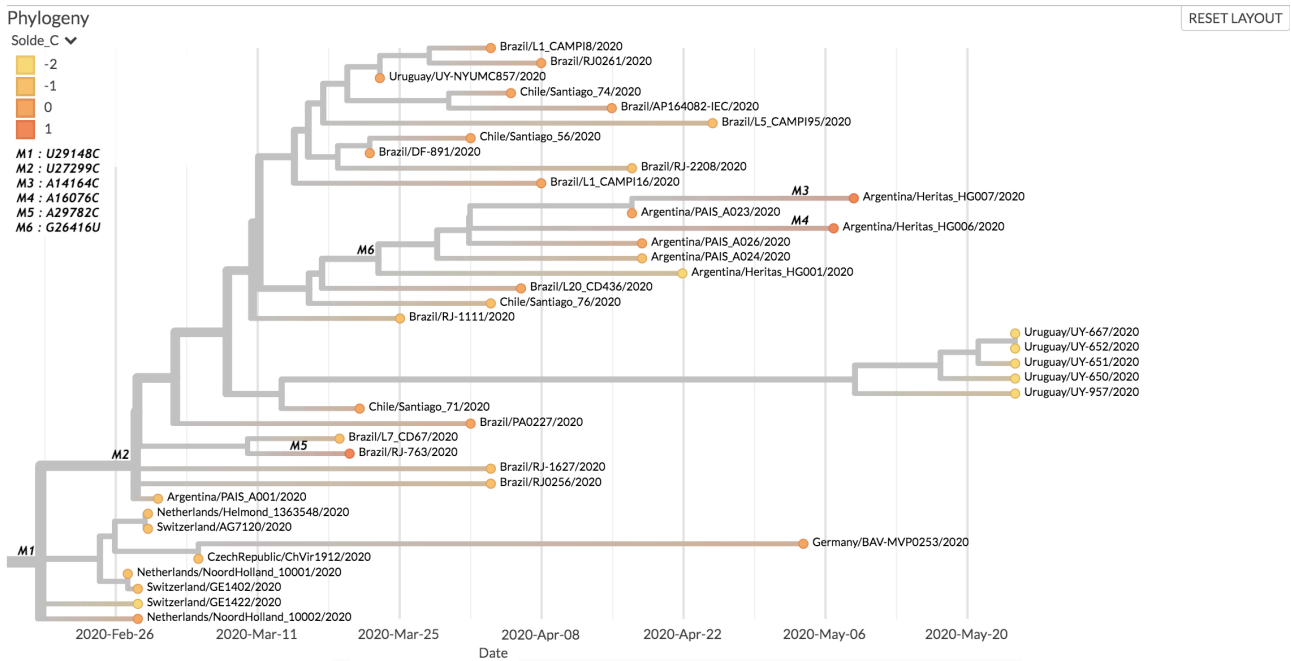
Figure 4. Two sub-trees in the first dataset where the tendency of the genome to lose its cytosine residues is reversed. **Upper panel. First sub-tree.** The sub-samples displayed are those that have acquired the most C, apart from a few isolated samples on other branches. The node with the M1 mutation directly follows those respectively associated with the C8782U and U28144C mutations. **Lower panel. Second sub-tree.** This tree contains a majority of strains with a neutral C balance (both gained and lost), as well as 3 strains with more C gained than lost.

In dataset 1, there are two sub-trees, the first of which is more of an Asian sub-tree with the root of the node associated with the M2 and M3 mutations. The second contains samples from North America and Oceania, and its root node is related to the M6 and M7 mutations. The first tree arises from the succession of C8782U (synonymous), U28144C (Leu84Ser) mutations in the Orf8 protein, whose function was discussed above. It defines a major clade of variants of the virus [24], C24034U (synonymous), and finally the doublet U26729C (synonymous), G28077C (Val62Leu), in the Orf8 protein again. As this is the origin of the observed phenomenon, we are led to believe that it is the alteration of the role of Orf8 (8a or 8b) that is responsible. The Orf8 region is particularly variable and has been clearly implicated in interspecies transmission [34]. A common hypothesis is that the alteration of this gene corresponds to a loss of active function in chiropteran ancestors [35]. Since these are generally richer in cytosine than the human forms [21], one might ask whether one of the functions of this protein is to modulate the activity of CTP synthase.

In fact, the second branch comes from the same descent, to which is added the U490A (Asp75Glu) mutation in the Nsp1 protein, which controls the specific translation of viral RNA [36], systematically associated with the mutation C3177U (Pro971Leu) in the acidic domain, without any clearly identified function, of the multifunctional protease Nsp3 [37], and finally the U18736C (Phe1757Leu) doublet of the exonuclease, N7-methyltransferase Nsp14, and A29700G in the 3'UTR region of the virus. The Phe1757Leu modification is located in the middle of a zinc binding

323   site at the interface between the two domains of the Nsp3 protein. It can therefore be surmised that
324   this mutation could subtly change the proofreading process correcting replication errors in a way
325   that would be less amenable to the entry of UTP opposite an A in the negative viral template. We
326   noted that 3 out of the 5 samples that acquired the most C did so through a transition from U to C.
327   The first one, HongKong/HKPU1_2101, shows two simultaneous transitions at positions 12877 and
328   23857. These mutations being synonymous, they are unlikely to change the replication-correction
329   mechanism. The second one, USA_SC_3571, and the third one, Australia/VIC209, show
330   transitions of the same type, also synonymous, at positions 11635 and 19713 respectively. Finally,
331   the last two samples, USA/WI-33 and USA/WI-28, were derived from the transversion from A to C
332   at position 29567, a mutation at the end of ORF9b.

333   For dataset number 2, this reversal of the trend concerns mostly Latino-American strains. The
334   succession of mutations C241U, C14408U, then A23403G discussed in relation to the generation
335   of end of translation codons in the virus genes, is followed by C3037U (synonymous), and the
336   triplet G28881A, G28882A, G28883C, overlapping the codons at position 203-204 of the N
337   nucleocapsid N gene. They mutate an arginine-glycine dipeptide into a lysine-arginine dipeptide.
338   This alters the positive charge of the protein and may help improve its role in the assembly of the
339   virus genome in the capsid, as discussed below in relation to the appearance of blooms (36). After
340   this triple modification, we see several reversals of the tendency to lose C in the genome.
341   U29148C (Ile292Thr) is found again in the nucleocapsid N gene, then U27299C (Ile33Thr) in the
342   Orf6 gene, resulting in a set of samples that have at worst gained as much C as they have lost.
343   There are also 3 samples among the 39 in the subtree that gained one more C than they lost
344   (Brazil/RJ-763, Argentina/Heritas_HG007, Argentina/Heritas_HG006). Each time, the last C
345   acquisition comes from a transversion from an adenine (in positions 14164 (Met233Leu), 16076
346   (Asp870Ala), and 29782, in the late 3'UTR of the viral genome. Overall, it is the change in the
347   nucleocapsid that appears to be most conducive to reversing the tendency to lose C. Indeed, this
348   protein, expressed at a high level during the infection, regulates the process of replication /
349   transcription of the virus and this may account for this remarkable observation [39].

350   *Emergence of blooms*

351   The succession C3037U, (C241U, A23403G), C14408U is present upstream of 10 sub-trees,
352   which we considered to be significant (see **Figure 5** and **Materials and Methods**).
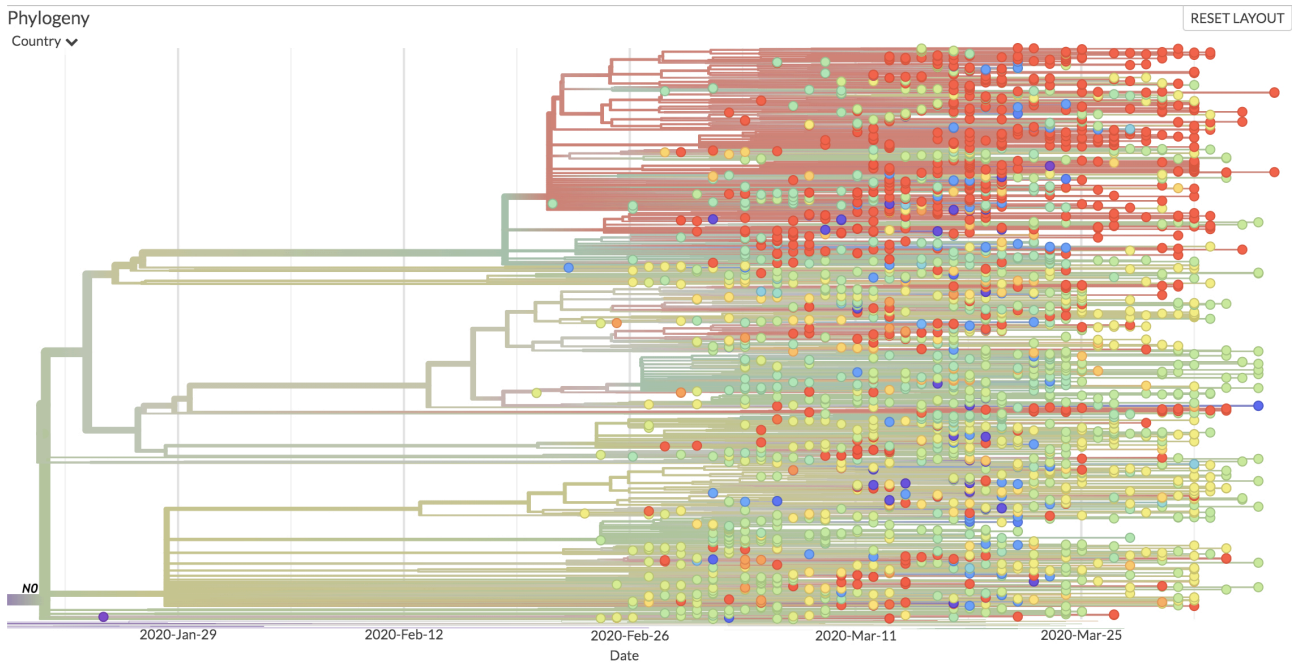
Figure 5: Example of blooms The subtree shown here contains 10 of the 20 most significant blooms in the sense of the method we used. Node N0 is the place where mutation C14408U emerges.

The synonymous mutation C3037U is located at the end of the ubiquitin-like domain 1 of protein Nsp3. This leads to a sequence (UUUUUU) that promotes changes in the reading frame and could decrease the translation efficiency of the proteins of the ORF1a region. The C241U mutation is observed very often [40]. It is found in the region that initiates replication of the virus. We can therefore assume that this may alter the frequency of replication. A23403G is a widely spread non-synonymous mutation which leads to the replacement of an aspartate by a glycine at position 614 of the spike protein, which is used by the virus to bind its host cell's receptor. For this reason, several previous analyses have suggested that this mutation has an important role in the spread of the virus [41,42]. Here, the fact that it is part of a major bloom can be considered as an additional argument favouring this interpretation. The C14408U changes an amino acid from proline to leucine (Pro314Leu) just after the end of the NiRAN domain (nidovirus RdRp-associated nucleotidyl transferase) of the protein Nsp12 ending in a "zinc finger". The NiRAN domain, essential for the replication of the virus, acts as a nucleotidyltransferase, preferring UTP as a substrate for a function which has not yet been clarified [43]. The proline modified in the mutant is part of a dipeptide diproline which plays the role of hinge of separation between the NiRAN domain and the following domain.

A second bloom, which shares several elements with the preceding one begins with the same sequence C3037U, (C241U, A23403G) and C14408U. However it continues with a series of contiguous mutations resulting in a change (G28881A, G28882A, G28883C) in nucleocapsid N, as we saw previously. It is worth noticing that this change might have a role in assembling the virus

377   genome in the capsid by phase separation [38]. This might increase the efficiency of virus

378   transmission and thus contribute to the formation of blooms. The fact that it is a cluster of

379   mutations involving G is intriguing. It may result from the fact that it spans a GGGG sequence.

380   We have previously seen that mutation G11083U (protein Nsp6, Leu37Phe) has initiated another

381   succession of mutations that led to premature translation termination of a viral protein. Here, this

382   widely distributed mutation is at the root of blooms. As discussed, it is possibly favouring

383   coronavirus infection by compromising delivery of viral components to lysosomes for degradation.

384   This would certainly favour blooms. The mutation is followed, in a first bloom-generating

385   succession, by G26144U (Gly251Val) in protein Orf3a, that forms potassium channels important for

386   innate immunity response - but the exact function of the protein still remains open to question [44].

387   Subsequent mutations are C14805U (synonymous) and U17247C (synonymous). This succession

388   suggests that the first mutation in protein Nsp6 and perhaps the second one are the primary

389   causes of the bloom [27]. The role of the fist mutation is further substantiated by the second bloom-

390   generating succession where it is followed by a quadruplet: C6312A (Thr2016Lys) in the inter-

391   domain region that precedes domain G2M of multi-domain protease Nsp3, then associated with

392   three C $\rightarrow$ U mutations, hence expected to be more frequent: C13730U (Ala88Val) in the NiRAN

393   domain of protein Nsp12, C23929U (synonymous), and finally C28311U (in a sequence of four C,

394   Pro13Leu) at the beginning of the nucleocapsid protein, N.

395   A second succession of mutations that ends up in blooms is C8782U (synonymous), U28144C

396   (Leu84Ser) in protein Orf8, the function of which has been discussed previously and defines a

397   significant clade of the virus variants [24], ending up with C26088U (synonymous). The Leu84Ser

398   mutation co-evolves significantly with the Asp614Gly mutations of the spike protein discussed

399   above [37], which makes it another likely candidate for positive selection leading to increased

400   spreading of the virus, hence blooms.

401   *Change in frequency of transitions / transversions*

402   Among the mutations upstream of branches showing significant changes in the transition /

403   transversion flow is the mutation C17747U, which modifies a proline residue into a leucine residue

404   in the protein Nsp13 (**Figure 6** and **Materials and Methods**).
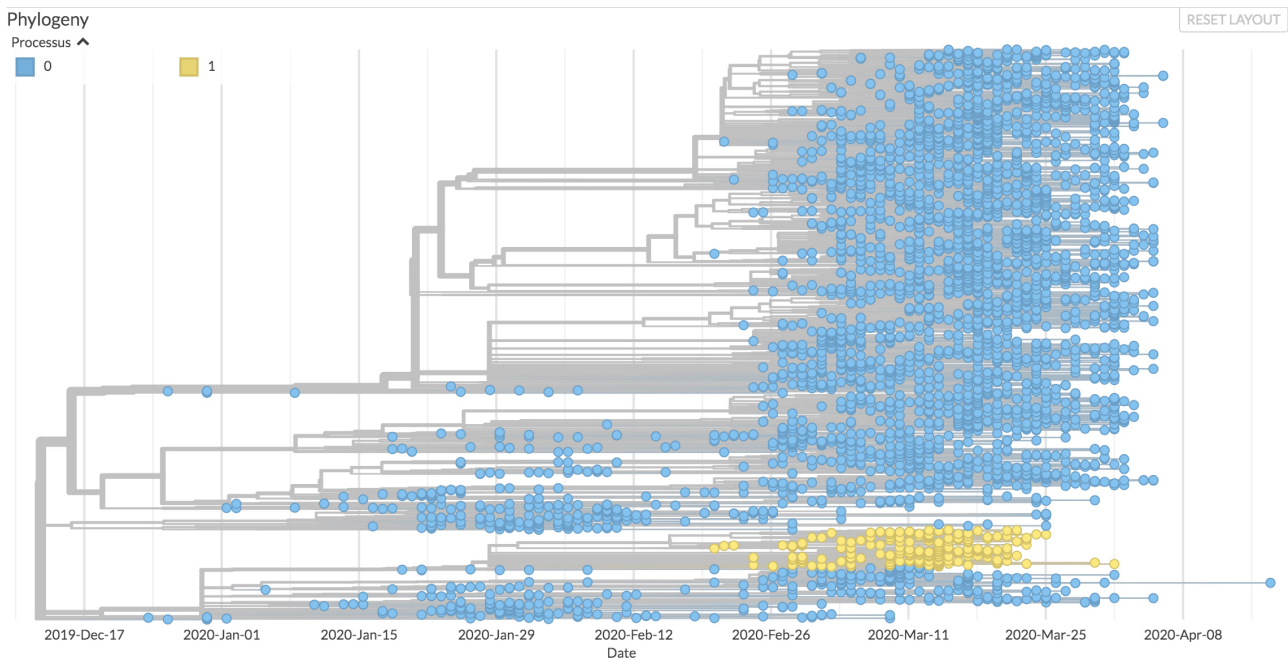
Figure 6. One of the descents considered to be significant for the change in the process of molecular evolution. The progeny resulting from the C17747U mutation is shown in yellow, and its evolutionary process is modelled by a 6-parameter TN93 model (process 1). The rest of the tree (blue leaves) is modelled by a 3-parameter TN93 model.

This mutation affects the protein domain which has nucleoside triphosphatase activity, the exact role of which is unknown but consistent with a proofreading activity [45]. We might propose that it is involved in the quality control of the product of the replication of the virus for example *via* stabilizing the "anti" form of nucleotides, thus avoiding the mismatching leading to transversions. In fact, this protein has been identified among those which lead to a significant alteration in the diversity of the viral genome [16]. The existence of a notable change in the type of mutations located downstream of the tree is therefore a strong argument for the discriminating role of the corresponding region of the protein. Furthermore, to the extent that this mutation increases the frequency of mutations in a biased manner, we can expect the ensuing descent to lead to an attenuation of the virus. However, as this changes the evolutionary landscape, this evolution could lead to "innovative" mutations modifying the pathogenicity of the virus, and this especially under conditions where recombination due to co-infections would be favoured. This is yet another argument for choosing a strong public health policy which tends to avoid the formation of clusters of infection.

Conclusions and perspectives

The COVID-19 epidemic is a life-size experiment in virus evolution. Remarkably, we neither know the real origin of the virus [46], nor where it will lead us. This explains why the vast majority of studies of the SARS-CoV-2 virus and its evolution are essentially descriptive. Here, we tried to make use of the ongoing evolution of the virus to investigate some of its related constraints using a hypothesis-driven probabilistic modelling approach to the molecular evolution of the virus. Based

429 on the assumption that the virus'metabolism is ruled by its host. Based on the metabolic set up of

430 the host cells, acting as a compulsory material framework for the multiplication of viral particles, we

431 pointed out specific changes in the evolution pattern of the virus descent, witnessed by changes in

432 the virus genome composition as time passes. Using the widely spread C to U change in this

433 genome's composition as a base line, we identified nodes where the change is shifted from this

434 direction to another one, favouring transversions rather than transitions, reversing the C to U trend

435 towards U to C enrichment or generating blooms with sudden appearance of multiple branches in

436 the evolution tree. This allowed us to point out a series of functions that are evolving towards a

437 more efficient spread of the virus (e.g. the previously identified Asp214Gly mutation of the spike

438 protein, but also the Gln57His mutation of the Orf3a potassium channel). We also noticed that Orf8

439 is the likely site of an ongoing competition for expression of two frameshift-dependent overlapping

440 proteins Orf8a and Orf8b. Similarly, the unstable region of Orf7 could promote the synthesis of the

441 very small membrane protein Orf7b, whose function remains unknown to date. Finally, the

442 reversion of the tendency to favour U over C indicates that nucleocapsid protein N may be involved

443 in the control of CTP synthesis in the host, suggesting an interesting target for future control of the

444 virus development. We hope that this combination of mathematical and biochemical knowledge will

445 help us devise further enterprises against the dire consequences of COVID-19. We noticed that

446 among the possible way for the virus to escape CTP-dependent control in cells would be to infect

447 cells that are not expected to grow, such as neurons. This may account for unexpected body sites

448 of viral development observed in the present epidemic.

Materials and Methods

Data processing

456 A total of 4,792 sequences of the SARS-CoV-2 virus were recovered from the GISAID databank

457 [47] on April 21, 2020 for the first dataset. Only the genomes of viruses from the human hosts of

458 SARS-CoV-2 of a length greater than 25,000 bp were retained. Sequences for which the sampling

459 date was insufficiently informed (absence of the harvest day, sometimes of the month) were also

460 excluded. For sequences present multiple times, only the first isolate was retained. We also reused

461 the work of the Nextstrain teams and discarded the too divergent or unstable samples that they

462 themselves had left out (github.com/nextstrain/ncov/blob/master/defaults/exclude.txt).The

463 sequence of 26 coding regions (Nsp1, Nsp2, Nsp3, Nsp4, Nsp5, Nsp6, Nsp7, Nsp8, Nsp9, Nsp10,

464 Nsp11, Nsp12, Nsp13, Nsp14, Nsp15, Nsp16, S, ORF3a, E, M, ORF6, ORF7a , ORF7b, ORF8, N

465 and ORF10) was characterized using NC_045512 as a reference. The total number of sequences

466 retained at the end of the treatment is 4,088 sequences. A second dataset of 3,246 sequences,

467 510 of which are common with the first dataset, was retrieved on July 6, 2020 using directly the

468 Nextstrain API [48].

469 We note here that, over time, data availability kept being altered, with some sequences deleted

470 from the samples, while other ones entered the database. Furthermore, it was generally difficult to

471 extract large samples of sequences so that it was extremely difficult to build up a consistent data

472 repository where correct statistical approaches could be implemented. It seems very awkward that

473 the bulk of the sequences of a virus of worldwide importance has not been made available at the

474 International Nucleotide Sequence Database despite recommendations of the major research

475 institutions [49,50].

476   *Phylogenetic reconstruction*

477 The reconstruction process begins with aligning all of the sequences to the reference sequence.

478 Insertions and deletions of genome regions were not taken into account. It was out of the question

479 here to take into account nucleotide insertions and deletions. We retained only the potential one to

480 one substitutions. We used program MAFFT [51] to generate these alignments. Some ambiguous

481 positions were highlighted during the alignment process. For example, some regions of the

482 genome may display high instability and wide variability depending on the parameters of the

483 algorithm used to perform the alignment. To overcome this problem, we used the same masks as

484 those used by the Nextstrain team. Sites 18529, 29849, 29851, 29853, as well as the first 130 and

485 last 50 sites of the genome were therefore omitted from the substitution analysis. We used a

486 General Time Reversible (GTR) model to infer the substitution process at work using the IQTREE

487 software [52]. This first tree is a fairly raw version which does not take into account the temporal

488 aspect of evolution. The Treetime software [53] allows you to refine this tree by also taking into

489 account the sampling dates of the sequences. It then reconstructs the tree with maximum

490 likelihood compared to the sampled sequences. Using maximum likelihood approaches it also

491 infers the compositions of the ancestral sequences of the samples, as well as a 90% confidence

492 interval around the most likely date of these common ancestors. Once the tree has been created,

493 we could then reconstruct the order in which the mutations in each sample appeared, in the sense

494 of maximum likelihood. For the visualization of the tree and the production of Figures 2 to 6, we

495 used the Auspice program developed by Nextstrain, to which we made some modifications to

496 display the parameters we were interested in. To this purpose, we developed a Python script to

497 modify the JSON file used as input by the Auspice program. This allowed us to enrich the

498 visualization capabilities of the software by adding quantities such as the number of C acquired or

499 lost by a sample compared to the reference and to generate original tree presentations.

500   *Identification of blooms*

501   The main pitfall we had to face when identifying blooms was the bias introduced when selecting

502   samples from the phylogenetic tree. In particular, some hospitals were likely to provide more

503   samples than others, due to the different health policies and means implemented depending on the

504   country. In order to avoid selecting nodes likely to generate a bloom due to oversampling, we

505   chose to develop a custom-made statistical method meant to cope with this difficulty.

506   A subtree is any set of nodes and leaves rooted in one of the nodes of the main tree. The idea is to

507   use the information provided by the identity of the countries represented in each sub-tree: the

508   easier a strain is spread, the higher the number of countries in which it is expected to be observed.

509   To implement this heuristic, it is necessary to control two factors: the size of the tree (two trees of

510   unequal depth, that is to say rooted on different dates, naturally show diversity as different

511   countries) and the heterogeneity of sampling (countries where sampling and sequencing are

512   carried out with different intensities have different probabilities of appearing in a given sub-tree).

513   These two factors interact, because the size of a tree (the number of its leaves for example)

514   obviously varies with the sampling intensity. One way to control this interaction is to measure the

515   size of a tree by its total length, or sum of branch lengths, in time units. Indeed, this observable is

516   not very sensitive to the effects of oversampling because the presence of many sequences

517   sampled in the same place at about the same time generates a sub-tree whose length is close to

518   zero.

519   To control the effect of the length factor L on the number of countries represented, N, we seeked to

520   learn the relation $N = f(L)$ in a typical tree in order to be subsequently able to identify the sub-trees

521   whose number of countries represented, for a known length L, exceeds the expected $f(L)$. A simple

522   statistical model consists in supposing that the number of occurrences of country i in a tree of

523   length L is a Poisson distribution of parameter $\theta_i L$ and that these numbers are independent. If K is

524   the total number of countries referenced by Nextstrain, the number of countries N represented in a

525   tree of length L is therefore the sum of K Bernoulli variables independent of parameters $1 - \exp(-$

526    $\theta_i L)$. For example, if countries are divided into two groups, the k_1 'frequent' of intensity θ_1, and the

527   k_2 'rare' of intensity $\theta_2 << \theta_1$, N has the mean $K - k_1\exp(-\theta_1 L) - k_2\exp(-\theta_2 L)$,

528   which behaves when L is large like $K - k_2\exp(-\theta_2 L)$.

529   In addition, when L is large, assuming that $\theta_2 L = O(1)$, the distribution of N is approximately equal

530   to $k_1 + N_2$, where N_2 foloows a Poisson law of parameter $k_2(1-\exp(-\theta_2 L))$.

531   So, we used the parameterization:

532   N = a - bexp(-cL), interpreting the parameters as follows: a is the maximum number of countries, b

533   is the number of countries with low sampling/sequencing intensities and c is a density of presence

534   of these countries per unit length of tree. Under the null hypothesis, N is distributed as a - b + N_1,

535   where N_1 follows the Poisson's law of parameter b(1-exp(-c L)). Finally, we selected the 20 most

536   significant blooms, i.e., those whose behaviour deviated the most from that expected by our

537   estimator. This then allowed us to reconstruct the lineages and the mutations that appeared

538   successively upstream of each node at which a bloom had occurred. This allowed us to identify the

539   succession of mutations common to some of these nodes and thus those giving rise to the majority

540   of statistically significant blooms. Furthermore, we restricted the automatic selection of nodes so

541   that no selected node was present in the lineage of another one. The selected blooms are

542   therefore mutually independent, even though they may obviously have common ancestors. To

543   arbitrate the choice between two nodes present in the same lineage, we have systematically kept

544   the oldest node, and thus the most dense tree.

545   *Detection of changes in the molecular evolutionary process*

546   We investigated whether the substitution process in some sub-trees behaved differently from what

547   was observed in the rest of the tree, statistically speaking. To this aim, we used the classical TN93

548   model from Tamura and Nei [54] with 3 parameters (purine transition rate, pyrimidine transition rate

549   and transversion rate) and allowed these three rates to take, downstream of a candidate node N_i,

550   values that differed from those they take in the rest of the tree. We then used a second (6-

551   parameter) model. Since this model is nested in the first (3-parameter) model, we used as test

552   statistic the likelihood ratio 2Δl = 2(l_1 - l_0), where l_0 is the log-likelihood under assumption H_0 (3-

553   parameter TN93 model estimating all the elements of the tree) and l_1 is the log-likelihood under

554   assumption H_1 (6-parameter TN93 model with local differentiation of the parameters downstream of

555   a node of interest). We then compared the likelihood ratio to a distribution of the \chi² with 3

556   degrees of freedom, whose significance threshold at 5% is 7.81. We were then able to identify the

557   nodes from which the evolution process varied significantly and to quantify the variations of the

558   different substitution rates, i.e. the nodes for which we can reject the H_0 hypothesis that the 3-

559   parameter model TN93 produces better estimates of tree substitution rates than the 6-parameter

560   model TN93. We have chosen to implement these models ourselves in Python, in order to keep

561   this parametrization flexibility. The program allows us to determine the set of nodes and leaves

562   present downstream of a node of interest and to perform hypothesis testing by calculating the

563   likelihood ratio and the different substitution rates.

564 **TEXT BOX**

565 **Molecular evolution and Muller's Ratchet**

566 Biology rests on the laws of physics. Because it develops at approximately 300K, it is subject to the

567 universal stress of thermal noise, involving an energy that does not differ considerably from that

568 involved in the chemical bonds of biological chemistry. It follows that the reactions that come

569 through and organize living things cannot develop with strict reproducibility. Inevitable errors cause

570 the product of a reaction to differ from what it is supposed to be. Genome replication cannot

571 escape this constraint. The consequence is that, in the progeny of a virus, there is always a

572 number of variants, named mutants when they carry over alterations of the genome. In most

573 cases, these mutants correspond to the change from one of the four nucleotides to a different one.

574 This process, as a rough approximation, is random — the mutant position can be anywhere in the

575 genome, and the replacement of one nucleotide is by any of the other three. As time goes by, all

576 the nucleotides of the genome are likely to change into others. This will affect the functions

577 necessary for the multiplication of the virus, and some changes will continue to be propagated (be

578 fixated), while others will end up without a progeny. A mutation followed by a fixation is called a

579 substitution. The substitution of a purine for a pyrimidine (or vice versa) is called a transversion;

580 other substitutions are called transitions. The likelihood of a particular mutation returning to the

581 ancestral state is very low. The probability that a particular mutation will return to the ancestral

582 state is very low. This forces evolution to always go forward, without the possibility of going back.

583 This process was noticed in 1932 by Hermann Muller in the special case of the effects of irradiation

584 on mutagenesis. His reflection has since been simplified and popularized. It is now known as the

585 "Muller's ratchet" [55]. It is obviously highly probable that the majority of mutations leads to the

586 partial or total loss of the functions coded by the altered regions of the genome. It follows that this

587 generally leads, in the long term - but not in the short term - to the attenuation of the functions

588 allowing multiplication and virulence of pathogenic species. This is why Louis Pasteur and his

589 successors could have the luck to isolate attenuated organisms which, in some - rare - cases,

590 could then be used for the vaccination of infected persons [56]. However, this process becomes

591 unproductive as soon as co-infection with different mutants occurs under circumstances where

592 recombination is possible. Two different mutants can recombine into the ancestral form of the

593 pathogen and erase the entire benefit of attenuation. This is all the more harmful since the old

594 forms are also, very often, those which spread most easily.

595
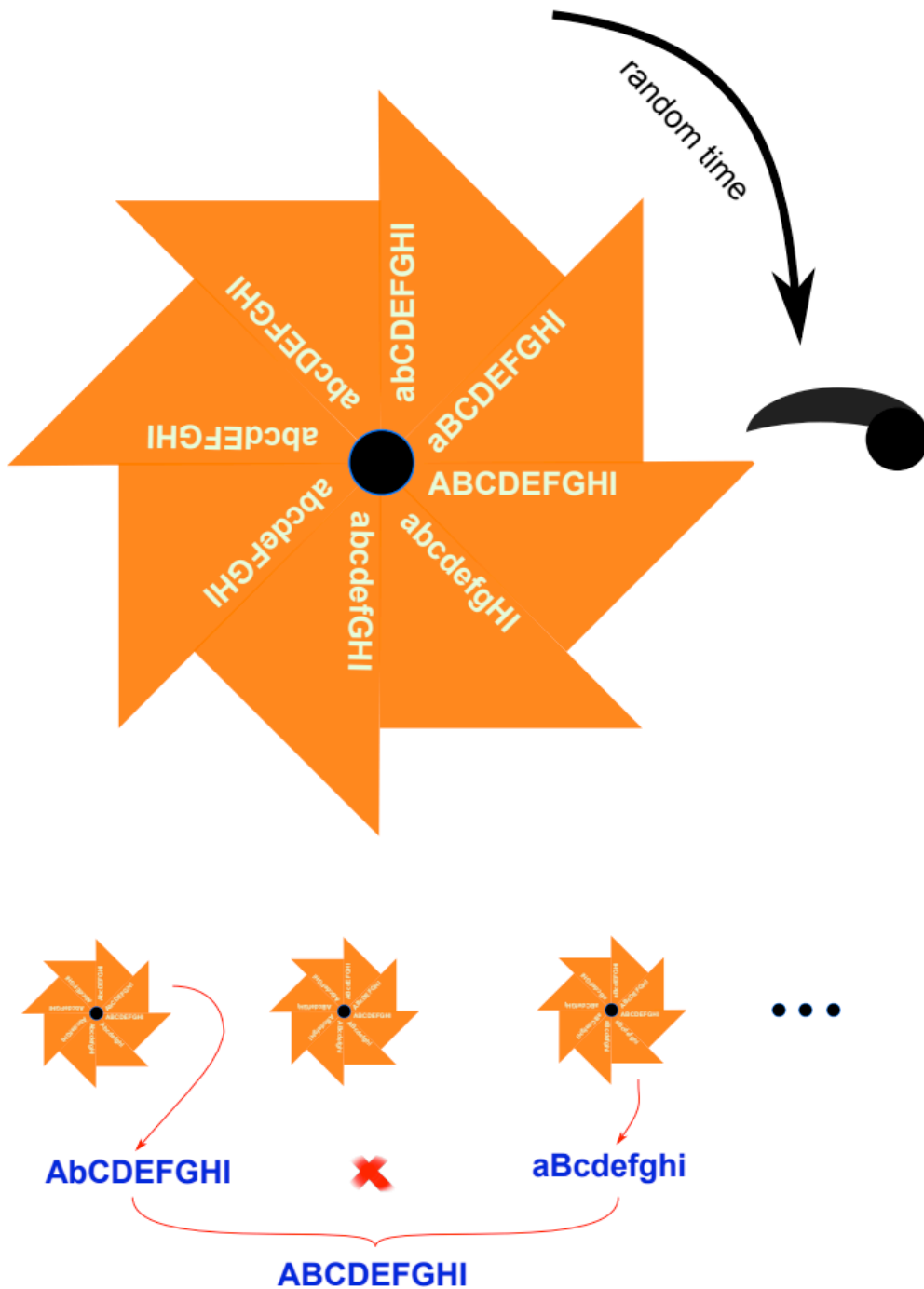
596 **Boxed text Figure. Muller's ratchet and recombination** The figure is reprinted from reference
597 [57]. Genes (capitals) are mutated at random in a different form (low case). Mutations accumulate
598 ratchet-like because the probability of reversion to the parent form is negligible. This happens
599 independently for viruses of different descents. However, if viruses from different descent happen
600 to be in the same cell, they can recombine. This allows them to recreate the ancestral form of the
601 virus.

602 **References**

[1] M. Romano, A. Ruggiero, F. Squeglia, G. Maga, R. Berisio, A structural view of SARS-CoV-2 RNA replication machinery: RNA synthesis, proofreading and final capping, Cells. 9 (2020) 1267. https://doi.org/10.3390/cells9051267.

[2] A. Lai, A. Bergna, C. Acciarri, M. Galli, G. Zehender, Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2, J. Med. Virol. 92 (2020) 675–679. https://doi.org/10.1002/jmv.25723.

[3] R.A. Fisher, The genetical theory of natural selection, Clarendon Press, Oxford, 1930. https://doi.org/10.5962/bhl.title.27468.

[4] A. Danchin, P. Marlière, Cytosine drives evolution of SARS-CoV-2, Environ. Microbiol. 22 (2020) 1977–1985. https://doi.org/10.1111/1462-2920.15025.

[5] P. Simmonds, Rampant C→U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short- and long-term evolutionary trajectories, MSphere. 5 (2020) e00408-20. https://doi.org/10.1128/mSphere.00408-20.

[6] P.C.Y. Woo, B.H.L. Wong, Y. Huang, S.K.P. Lau, K.-Y. Yuen, Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses, Virology. 369 (2007) 431–442. https://doi.org/10.1016/j.virol.2007.08.010.

[7] Z. Ou, C. Ouzounis, D. Wang, W. Sun, J. Li, W. Chen, P. Marlière, A. Danchin, A path towards SARS-CoV-2 attenuation: metabolic pressure on CTP synthesis rules the virus evolution, 2020. https://doi.org/10.1101/2020.06.20.162933.

[8] I. Sola, F. Almazán, S. Zúñiga, L. Enjuanes, Continuous and discontinuous RNA synthesis in coronaviruses, Annu Rev Virol. 2 (2015) 265–288. https://doi.org/10.1146/annurev-virology-100114-055218.

[9] S. Di Giorgio, F. Martignano, M.G. Torcia, G. Mattiuz, S.G. Conticello, Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2, Sci Adv. 6 (2020) eabb5813. https://doi.org/10.1126/sciadv.abb5813.

[10] J. Alonso-Carrera, C. de Miguel, B. Manzano, Economic growth and environmental degradation when preferences are non-homothetic, Environ Resource Econ. 74 (2019) 1011–1036. https://doi.org/10.1007/s10640-019-00357-4.

[11] Freeman J Dyson, Origins of life, Cambridge University Press, Cambridge, UK, 1985.

[12] K. Wellner, H. Betat, M. Mörl, A tRNA's fate is decided at its 3' end: Collaborative actions of CCA-adding enzyme and RNases involved in tRNA processing and degradation, Biochim Biophys Acta Gene Regul Mech. 1861 (2018) 433–441. https://doi.org/10.1016/j.bbagrm.2018.01.012.

[13] P. Shridas, C.J. Waechter, Human dolichol kinase, a polytopic endoplasmic reticulum membrane protein with a cytoplasmically oriented CTP-binding site, J. Biol. Chem. 281 (2006) 31696–31704. https://doi.org/10.1074/jbc.M604087200.

[14] C. Wang, Z. Liu, Z. Chen, X. Huang, M. Xu, T. He, Z. Zhang, The establishment of reference sequence for SARS-CoV-2 and variation analysis, J. Med. Virol. 92 (2020) 667–674. https://doi.org/10.1002/jmv.25762.

[15] X. Yang, N. Dong, E.W.-C. Chan, S. Chen, Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries, Emerg Microbes Infect. 9 (2020) 1287–1299. https://doi.org/10.1080/22221751.2020.1773745.

[16] L. van Dorp, M. Acman, D. Richard, L.P. Shaw, C.E. Ford, L. Ormond, C.J. Owen, J. Pang, C.C.S. Tan, F.A.T. Boshier, A.T. Ortiz, F. Balloux, Emergence of genomic

diversity and recurrent mutations in SARS-CoV-2, Infect. Genet. Evol. 83 (2020) 104351. https://doi.org/10.1016/j.meegid.2020.104351.

[17] Y. Yang, W. Yan, B. Hall, X. Jiang, Characterizing transcriptional regulatory sequences in coronaviruses and their role in recombination, BioRxiv. (2020). https://doi.org/10.1101/2020.06.21.163410.

[18] H. Grosjean, V. de Crécy-Lagard, C. Marck, Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes, FEBS Lett. 584 (2010) 252–264. https://doi.org/10.1016/j.febslet.2009.11.052.

[19] S.S. Rout, M. Singh, K.S. Shindler, J. Das Sarma, One proline deletion in the fusion peptide of neurotropic mouse hepatitis virus (MHV) restricts retrograde axonal transport and neurodegeneration, J. Biol. Chem. 295 (2020) 6926–6935. https://doi.org/10.1074/jbc.RA119.011918.

[20] E.E. Rivera-Serrano, A.S. Gizzi, J.J. Arnold, T.L. Grove, S.C. Almo, C.E. Cameron, Viperin reveals its true function, Annu. Rev. Virol. 7 (2020) annurev-virology-011720-095930. https://doi.org/10.1146/annurev-virology-011720-095930.

[21] J. Armengaud, A. Delaunay-Moisan, J.-Y. Thuret, E. van Anken, D. Acosta-Alvear, T. Aragón, C. Arias, M. Blondel, I. Braakman, J.-F. Collet, R. Courcol, A. Danchin, J.-F. Deleuze, J.-P. Lavigne, S. Lucas, T. Michiels, E.R.B. Moore, J. Nixon-Abell, R. Rosello-Mora, Z.-L. Shi, A.G. Siccardi, R. Sitia, D. Tillett, K.N. Timmis, M.B. Toledano, P. van der Sluijs, E. Vicenzi, The importance of naturally attenuated SARS-CoV-2in the fight against COVID-19, Environ. Microbiol. 22 (2020) 1997–2000. https://doi.org/10.1111/1462-2920.15039.

[22] S. Liu, J. Shen, L. Yang, C.-D. Hu, J. Wan, Distinct genetic spectrums and evolution patterns of SARS-CoV-2, Health Informatics, 2020. https://doi.org/10.1101/2020.06.16.20132902.

[23] C.A. Nelson, A. Pekosz, C.A. Lee, M.S. Diamond, D.H. Fremont, Structure and intracellular targeting of the SARS-Coronavirus Orf7a accessory protein, Structure. 13 (2005) 75–85. https://doi.org/10.1016/j.str.2004.10.010.

[24] J.-S. Kim, J.-H. Jang, J.-M. Kim, Y.-S. Chung, C.-K. Yoo, M.-G. Han, Genome-wide identification and characterization of point mutations in the SARS-CoV-2 genome, Osong Public Health Res Perspect. 11 (2020) 101–111. https://doi.org/10.24171/j.phrp.2020.11.3.05.

[25] A. Addetia, H. Xie, P. Roychoudhury, L. Shrestha, M. Loprieno, M.-L. Huang, K.R. Jerome, A.L. Greninger, Identification of multiple large deletions in ORF7a resulting in in-frame gene fusions in clinical SARS-CoV-2 isolates, J. Clin. Virol. 129 (2020) 104523. https://doi.org/10.1016/j.jcv.2020.104523.

[26] S.R. Schaecher, J.M. Mackenzie, A. Pekosz, The ORF7b protein of severe acute respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and incorporated into SARS-CoV particles, J. Virol. 81 (2007) 718–731. https://doi.org/10.1128/JVI.01691-06.

[27] D. Benvenuto, S. Angeletti, M. Giovanetti, M. Bianchi, S. Pascarella, R. Cauda, M. Ciccozzi, A. Cassone, Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy, J. Infect. 81 (2020) e24–e27. https://doi.org/10.1016/j.jinf.2020.03.058.

[28] S. Angeletti, D. Benvenuto, M. Bianchi, M. Giovanetti, S. Pascarella, M. Ciccozzi, COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis, J. Med. Virol. 92 (2020) 584–588. https://doi.org/10.1002/jmv.25719.

[29] M.C. Hagemeijer, I. Monastyrska, J. Griffith, P. van der Sluijs, J. Voortman, P.M. van Bergen en Henegouwen, A.M. Vonk, P.J.M. Rottier, F. Reggiori, C.A.M. de Haan,

Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4, Virology. 458–459 (2014) 125–135. https://doi.org/10.1016/j.virol.2014.04.027.

[30] B. Short, A call for oxygen in the ER, The Journal of Cell Biology. 203 (2013) 552–552. https://doi.org/10.1083/jcb.2034iti3.

[31] C. Selvaraj, D.C. Dinesh, U. Panwar, R. Abhirami, E. Boura, S.K. Singh, Structure-based virtual screening and molecular dynamics simulation of SARS-CoV-2 Guanine-N7 methyltransferase (nsp14) for identifying antiviral inhibitors against COVID-19, J. Biomol. Struct. Dyn. (2020) 1–12. https://doi.org/10.1080/07391102.2020.1778535.

[32] S. Chen, X. Zheng, J. Zhu, R. Ding, Y. Jin, W. Zhang, H. Yang, Y. Zheng, X. Li, G. Duan, Extended ORF8 Gene Region Is Valuable in the Epidemiological Investigation of Severe Acute Respiratory Syndrome-Similar Coronavirus, J. Infect. Dis. 222 (2020) 223–233. https://doi.org/10.1093/infdis/jiaa278.

[33] E. Issa, G. Merhi, B. Panossian, T. Salloum, S. Tokajian, SARS-CoV-2 and ORF3a: nonsynonymous mutations, functional domains, and viral pathogenesis, MSystems. 5 (2020) e00266-20. https://doi.org/10.1128/mSystems.00266-20.

[34] M. Bolles, E. Donaldson, R. Baric, SARS-CoV and emergent coronaviruses: viral determinants of interspecies transmission, Current Opinion in Virology. 1 (2011) 624–634. https://doi.org/10.1016/j.coviro.2011.10.012.

[35] V.M. Corman, H.J. Baldwin, A.F. Tateno, R.M. Zerbinati, A. Annan, M. Owusu, E.E. Nkrumah, G.D. Maganga, S. Oppong, Y. Adu-Sarkodie, P. Vallo, L.V.R.F. da Silva Filho, E.M. Leroy, V. Thiel, L. van der Hoek, L.L.M. Poon, M. Tschapka, C. Drosten, J.F. Drexler, Evidence for an ancestral association of human coronavirus 229E with bats, J. Virol. 89 (2015) 11858–11870. https://doi.org/10.1128/JVI.01755-15.

[36] M. Thoms, R. Buschauer, M. Ameismeier, L. Koepke, T. Denk, M. Hirschenberger, H. Kratzat, M. Hayn, T. Mackens-Kiani, J. Cheng, C.M. Stürzel, T. Fröhlich, O. Berninghausen, T. Becker, F. Kirchhoff, K.M.J. Sparrer, R. Beckmann, Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2, Molecular Biology, 2020. https://doi.org/10.1101/2020.05.18.102467.

[37] S. Laha, J. Chakraborty, S. Das, S.K. Manna, S. Biswas, R. Chatterjee, Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission, Infect. Genet. Evol. 85 (2020) 104445. https://doi.org/10.1016/j.meegid.2020.104445.

[38] C. Iserman, C. Roden, M. Boerneke, R. Sealfon, G. McLaughlin, I. Jungreis, C. Park, A. Boppana, E. Fritch, Y.J. Hou, C. Theesfeld, O.G. Troyanskaya, R.S. Baric, T.P. Sheahan, K. Weeks, A.S. Gladfelter, Specific viral RNA drives the SARS CoV-2 nucleocapsid to phase separate, Biochemistry, 2020. https://doi.org/10.1101/2020.06.11.147199.

[39] Y. Cong, M. Ulasli, H. Schepers, M. Mauthe, P. V'kovski, F. Kriegenburg, V. Thiel, C.A.M. de Haan, F. Reggiori, Nucleocapsid protein recruitment to replication-transcription complexes plays a crucial role in coronaviral life cycle, J Virol. 94 (2019) e01925-19, /jvi/94/4/JVI.01925-19.atom. https://doi.org/10.1128/JVI.01925-19.

[40] O.M. Ugurel, O. Ata, D. Turgut-Balik, An updated analysis of variations in SARS-CoV-2 genome, Turk. J. Biol. 44 (2020) 157–167. https://doi.org/10.3906/biy-2005-111.

[41] Z. Daniloski, X. Guo, N.E. Sanjana, The D614G mutation in SARS-CoV-2 Spike increases transduction of multiple human cell types, 2020. https://doi.org/10.1101/2020.06.14.151357.

[42] R. Lorenzo-Redondo, H.H. Nam, S.C. Roberts, L.M. Simons, L.J. Jennings, C. Qi, C.J. Achenbach, A.R. Hauser, M.G. Ison, J.F. Hultquist, E.A. Ozer, A unique clade of SARS-CoV-2 viruses is associated with lower viral loads in patient upper airways, 2020. https://doi.org/10.1101/2020.05.19.20107144.

[43] C.C. Posthuma, A.J.W. Te Velthuis, E.J. Snijder, Nidovirus RNA polymerases: Complex enzymes handling exceptional RNA genomes, Virus Res. 234 (2017) 58–73. https://doi.org/10.1016/j.virusres.2017.01.023.

[44] S.-Y. Fung, K.-S. Yuen, Z.-W. Ye, C.-P. Chan, D.-Y. Jin, A tug-of-war between severe acute respiratory syndrome coronavirus 2 and host antiviral defence: lessons from other pathogenic viruses, Emerg Microbes Infect. 9 (2020) 558–570. https://doi.org/10.1080/22221751.2020.1736644.

[45] K.A. Ivanov, J. Ziebuhr, Human coronavirus 229E nonstructural protein 13: characterization of duplex-unwinding, nucleoside triphosphatase, and RNA 5'-triphosphatase activities, J. Virol. 78 (2004) 7833–7838. https://doi.org/10.1128/JVI.78.14.7833-7838.2004.

[46] M. Letko, S.N. Seifert, K.J. Olival, R.K. Plowright, V.J. Munster, Bat-borne virus diversity, spillover and emergence, Nat. Rev. Microbiol. 18 (2020) 461–471. https://doi.org/10.1038/s41579-020-0394-z.

[47] S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health: Data, Disease and Diplomacy, Global Challenges. 1 (2017) 33–46. https://doi.org/10.1002/gch2.1018.

[48] J. Hadfield, C. Megill, S.M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R.A. Neher, Nextstrain: real-time tracking of pathogen evolution, Bioinformatics. 34 (2018) 4121–4123. https://doi.org/10.1093/bioinformatics/bty407.

[49] R.I. Amann, S. Baichoo, B.J. Blencowe, P. Bork, M. Borodovsky, C. Brooksbank, P.S.G. Chain, R.R. Colwell, D.G. Daffonchio, A. Danchin, V. de Lorenzo, P.C. Dorrestein, R.D. Finn, C.M. Fraser, J.A. Gilbert, S.J. Hallam, P. Hugenholtz, J.P.A. Ioannidis, J.K. Jansson, J.F. Kim, H.-P. Klenk, M.G. Klotz, R. Knight, K.T. Konstantinidis, N.C. Kyrpides, C.E. Mason, A.C. McHardy, F. Meyer, C.A. Ouzounis, A.A.N. Patrinos, M. Podar, K.S. Pollard, J. Ravel, A.R. Muñoz, R.J. Roberts, R. Rosselló-Móra, S.-A. Sansone, P.D. Schloss, L.M. Schriml, J.C. Setubal, R. Sorek, R.L. Stevens, J.M. Tiedje, A. Turjanski, G.W. Tyson, D.W. Ussery, G.M. Weinstock, O. White, W.B. Whitman, I. Xenarios, Toward unrestricted use of public genomic data, Science. 363 (2019) 350–352. https://doi.org/10.1126/science.aaw1280.

[50] I. Karsch-Mizrachi, T. Takagi, G. Cochrane, International Nucleotide Sequence Database Collaboration, The international nucleotide sequence database collaboration, Nucleic Acids Res. 46 (2018) D48–D51. https://doi.org/10.1093/nar/gkx1097.

[51] K.D. Yamada, K. Tomii, K. Katoh, Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees, Bioinformatics. 32 (2016) 3246–3251. https://doi.org/10.1093/bioinformatics/btw412.

[52] B.Q. Minh, H.A. Schmidt, O. Chernomor, D. Schrempf, M.D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era, Mol. Biol. Evol. 37 (2020) 1530–1534. https://doi.org/10.1093/molbev/msaa015.

[53] P. Sagulenko, V. Puller, R.A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis, Virus Evol. 4 (2018) vex042. https://doi.org/10.1093/ve/vex042.

[54] Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees., Molecular Biology and Evolution. (1993). https://doi.org/10.1093/oxfordjournals.molbev.a040023.

[55] H.J. Muller, Some genetic aspects of sex, The American Naturalist. 66 (1932) 118–138. https://doi.org/10.1086/280418.

[56] K.A. Smith, Louis Pasteur, the father of immunology?, Front Immunol. 3 (2012) 68. https://doi.org/10.3389/fimmu.2012.00068.

[57] A. Danchin, K. Timmis, SARS-CoV-2 variants: Relevance for symptom granularity, epidemiology, immunity (herd, vaccines), virus origin and containment?, Environ. Microbiol. 22 (2020) 2001–2006. https://doi.org/10.1111/1462-2920.15053.

603
604