



**HAL**  
open science

## **SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies**

Timokratis Karamitros, Gethsimani Papadopoulou, Maria Bousali, Anastasios Mexias, Sotirios Tsiodras, Andreas Mentis

► **To cite this version:**

Timokratis Karamitros, Gethsimani Papadopoulou, Maria Bousali, Anastasios Mexias, Sotirios Tsiodras, et al.. SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. *Journal of Clinical Virology*, 2020, 131, pp.104585. 10.1016/j.jcv.2020.104585 . hal-03491185

**HAL Id: hal-03491185**

**<https://hal.science/hal-03491185v1>**

Submitted on 22 Aug 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1                   **SARS-CoV-2 exhibits intra-host genomic plasticity**  
2                   **and low-frequency polymorphic quasispecies**

3  
4  
5 Timokratis Karamitros<sup>1#</sup>, Gethsimani Papadopoulou<sup>1</sup>, Maria Bousali<sup>1</sup>, Anastasios Mexias<sup>1</sup>,  
6 Sotirios Tsiodras<sup>2</sup>, Andreas Mentis<sup>3</sup>

7  
8 1. Unit of Bioinformatics and Applied Genomics, Department of Microbiology, Hellenic  
9 Pasteur Institute, Athens, Greece

10 2. 4<sup>th</sup> Academic Department of Medicine, National and Kapodistrian University of Athens,  
11 Medical School, Athens, Greece.

12 3. Public Health Laboratories, Department of Microbiology, Hellenic Pasteur Institute,  
13 Athens, Greece

14  
15  
16 #To whom correspondence should be addressed:

17 Dr. Timokratis Karamitros

18 Researcher

19 Bioinformatics and Applied Genomics Unit

20 Department of Microbiology

21 Hellenic Pasteur Institute

22 127 Vas. Sophias avenue

23 11521, Athens, Greece

24 m: tkaram@pasteur.gr

25 t: +30 2106478-871, -874

26

27

## 28 **ABSTRACT**

29

30 In December 2019, an outbreak of atypical pneumonia (Coronavirus disease 2019 -  
31 COVID-19) associated with a novel coronavirus (SARS-CoV-2) was reported in Wuhan city,  
32 Hubei province, China. The outbreak was traced to a seafood wholesale market and human  
33 to human transmission was confirmed. The rapid spread and the death toll of the new  
34 epidemic warrants immediate intervention. The intra-host genomic variability of SARS-CoV-2  
35 plays a pivotal role in the development of effective antiviral agents and vaccines, as well as  
36 in the design of accurate diagnostics.

37 We analyzed NGS data derived from clinical samples of three Chinese patients  
38 infected with SARS-CoV-2, in order to identify small- and large-scale intra-host variations in  
39 the viral genome. We identified tens of low- or higher- frequency single nucleotide variations  
40 (SNVs) with variable density across the viral genome, affecting 7 out of 10 protein-coding  
41 viral genes. The majority of these SNVs (72/104) corresponded to missense changes. The  
42 annotation of the identified SNVs but also of all currently circulating strain variations revealed  
43 colocalization of intra-host as well as strain specific SNVs with primers and probes currently  
44 used in molecular diagnostics assays. Moreover, we de-novo assembled the viral genome,  
45 in order to isolate and validate intra-host structural variations and recombination breakpoints.  
46 The bioinformatics analysis disclosed genomic rearrangements over poly-A / poly-U regions  
47 located in ORF1ab and spike (S) gene, including a potential recombination hot-spot within S  
48 gene.

49 Our results highlight the intra-host genomic diversity and plasticity of SARS-CoV-2,  
50 pointing out genomic regions that are prone to alterations. The isolated SNVs and genomic  
51 rearrangements reflect the intra-patient capacity of the polymorphic quasispecies, which may  
52 arise rapidly during the outbreak, allowing immunological escape of the virus, offering  
53 resistance to anti-viral drugs and affecting the sensitivity of the molecular diagnostics  
54 assays.

55

56

57

58

59 **KEYWORDS**

60 SARS-CoV-2, COVID-19 epidemic, intra-host variability, quasispecies, genomic  
61 rearrangements, molecular diagnostics

62

63

64 Highlights

- 65 • SARS-CoV-2 exhibits intra-host small- and large-scale genomic variability
- 66 • SNVs are colocalized with probes and primers used in molecular diagnostic assays
- 67 • SARS-CoV-2 Spike (S) gene host a potential recombination hot-spot

68

69

## 70 INTRODUCTION

71 Coronaviruses (CoVs), considered to be the largest group of viruses, belong to the  
72 *Nidovirales* order, *Coronaviridae* family and *Coronavirinae* subfamily, which is further  
73 subdivided into four genera, the alpha- and betacoronaviruses, which infect mammalian  
74 species and gamma- and deltacoronaviruses infecting mainly birds [1,2]. Small mammals  
75 (mice, dogs, cats) serve as reservoirs for Human Coronaviruses (HCoVs), with significant  
76 diversity seen in bats, which are considered to be primordial hosts of HCoVs [3].

77 Until 2002, minor consideration was given to HCoVs, as they were associated with  
78 mild-to-severe disease phenotypes in immunocompetent people [3–5]. In 2002, the  
79 beginning of severe acute respiratory syndrome (SARS) outbreak took place [6]. In 2005,  
80 after the discovery of SARS-CoV-related viruses in horseshoe bats (*Rhinolophus*), palm  
81 civets were suggested as intermediate hosts, and bats as primordial hosts of the virus [6,7].  
82 In 2012, the emerging Middle East respiratory syndrome coronavirus (MERS-CoV) caused  
83 an outbreak in Saudi Arabia, which affected both camels and humans, with a high mortality  
84 rate of approximately 34,3% among humans [8]. MERS-CoV has zoonotic origins [9] and  
85 was transmitted to humans through direct contact with dromedary camels or indirect contact  
86 with contaminated meat or milk[10].

87 On December 31<sup>st</sup> – 2019, a novel Coronavirus (SARS-CoV-2) was first reported  
88 from the city of Wuhan, Hubei province in China, causing severe infection of the respiratory  
89 tract in humans, after the identification of a group of similar cases of patients with pneumonia  
90 of unknown etiology [11]. Similarly to SARS, epidemiological links between the majority of  
91 COVID-19 cases and Huanan South China Seafood Market, a live-animal market, have  
92 been reported. A total of 76,775 confirmed cases of “Coronavirus Disease 2019” (COVID-19)  
93 were reported up to February 21<sup>st</sup> 2020, from which 2,247 died and 18,855 recovered.  
94 Notably, 75,447 of the confirmed cases were reported in China [12].

95 The size of the ssRNA genome of SARS-CoV-2 is 29,891 nucleotides, it encodes  
96 9860 amino acids and is characterized by nucleotide identity of ~ 89% with bat SARS-

97 related (SL) CoV-ZXC21 and bat-SL-CoVZC45. However, when compared to HCoVs,  
98 SARS-CoV-2 showed genetic similarity of ~ 80% with human SARS-CoVs BJ01 2003 and  
99 Tor2 [13] and and 50% with MERS-CoV [14,15]. CoVs are enveloped positive-sense RNA  
100 viruses, characterized by a very large non-segmented genome (26 to 32kb length), ready to  
101 be translated [2,4]. The genes arrangement on the SARS-CoV-2 genome is: 5'UTR -  
102 replicase (ORF1/ab) -Spike (S) -ORF3a -Envelope (E) -Membrane (M) -ORF6 -ORF7a -  
103 ORF8 -Nucleocapsid (N) ORF10 -3'UTR [13]. SARS-CoV-2 encodes proteins that are very  
104 similar in length compared to bat-SL-CoVZC45 and bat-SL-CoVZXC21. The SARS-CoV-2 S  
105 protein however is longer compared to those encoded by SARS-CoV, and MERS-CoV [15].

106 At inter-host level, adaptive mutations are essential for the newly emerging viruses in  
107 order to increase replication and facilitate onward transmission in the new hosts [16].  
108 Particularly for MERS-CoV, SARS-CoV and SARS-CoV-2, the genetic diversity and frequent  
109 recombination events, lead to periodical emergence of new viruses capable of infecting a  
110 wide range of hosts [17]. Intra-host variability in viral infections, emerges from genomic  
111 phenomena taking place during error-prone replication, ending up to multiple circulating  
112 quasispecies of low or higher frequency [18]. These variants, in combination with the  
113 genetic profile of the host, can potentially influence the natural history of the infection, the  
114 viral phenotype, but also the sensitivity of molecular and serological diagnostics assays  
115 [19,20]. In the case of flu epidemics for example, de novo arising mutations and intra-host  
116 diversity not only forms intra-host evolution of Influenza A, but also greatly affects the  
117 pathogenesis of the virus [21–23]. Indeed, it is suggested that SARS-CoV-2 genomic  
118 variants that emerge from inter- and intra-host evolution might be associated with  
119 susceptibility to SARS-CoV-2 infection and the severity of COVID-19 [24].

120 Viruses have developed multiple adaptive strategies to counteract the host  
121 immunological response, which are subject to inter- and intra-host selection pressures;  
122 “Selfish” strategies confer a selective advantage in a particular quasispecies, impair the  
123 immune response inside the infected cell and evolve by intra-host selection, while neutral or  
124 “unselfish” defence strategies impair the immune response outside the infected cell and

125 evolve by inter-host selection, preferentially in viruses with low mutation rates [25]. SARS-  
126 CoV-2 mutation rate is moderate and similar to other RNA viruses (0.00084 per site per  
127 year) [26], but still generally higher compared to DNA viruses [27]. Moreover, most of the  
128 suggested immune escape mechanisms of SARS-CoV-2 involve intra-cellular interactions  
129 [28], thus expected to evolve by intra-host selective pressure. These observations highlight  
130 the importance of SARS-CoV-2 intra-host variability in the frame of viral evolution and host-  
131 pathogen interactions.

132 Intra-host genomic variability also leads to antigenic variability, which is of higher  
133 importance, especially for pathogens that fail to elicit long-lasting immunity in their hosts,  
134 and remains a major contributor to the complexity of vaccine design [29,30]. To date, there  
135 are no clinically approved vaccines available for protection of general population from SARS-  
136 and MERS-CoV infections as there is no effective vaccine to induce robust cell mediated  
137 and humoral immune responses [31,32].

138 Here, we explore intra-host genomic variants and low-frequency polymorphic  
139 quasispecies in Next Generation Sequencing (NGS) data derived from patients infected by  
140 SARS-CoV-2. Intra-host genomic variability is critical for the development of novel drugs and  
141 vaccines, which are of urgent necessity, towards the containment of the pandemic.

142

143

## 144 **MATERIALS AND METHODS**

145

146 In this study NGS data derived from three Chinese patients (oral swabs) infected by  
147 SARS-CoV-2 were analysed (SRA projects PRJNA601736 and PRJNA603194). All datasets  
148 available in SRA up to February 20<sup>th</sup>, 2020 were analysed. The two patients (SRR10903401  
149 and SRR10903402/PRJNA601736), 39- and 21-year-old respectively, experienced unusual  
150 pneumonia. Despite his anti-viral treatment, patient 1 experienced more severe symptoms  
151 The two patients were admitted to the hospital on 25<sup>th</sup> and 22<sup>th</sup> December 2019 and were

152 discharged in stable condition on 12<sup>th</sup> and 11<sup>th</sup> January 2020, respectively [33]. The third 41-  
153 year-old male patient (SRR1097138/PRJNA603194), presented acute onset of common  
154 COVID19 symptoms. A combinatory antiviral therapy was administered to the patient.  
155 However, he exhibited respiratory failure and was admitted to the intensive care unit. Six  
156 days after his admission, he was transferred to another hospital in Wuhan for further  
157 treatment [34]. Detailed clinical metadata of the patients are presented in Supplementary  
158 Material.

159 The raw read data were aligned on the complete (29,891 bp) SARS-CoV-2 reference  
160 sequence (GenBank accession no. MN975262.1, isolate 2019-nCoV\_HKU-SZ-005b\_2020)  
161 using bowtie2 v2.3.0 [35], after quality check with FastQC v0.11.5 [36]. The resulting  
162 alignments were visualized with the *Integrated Genomics Viewer* (IGV) v2.3.60 [37]. After  
163 removing PCR duplicates, SNVs were called with a Bonferroni-corrected *P*-value threshold  
164 of 0.05 using *samtools* v1.7 (htslib1.7.2) [38] and *LoFreq* v2.1.5. LoFreq is a very accurate  
165 SNV caller especially designed for viral and bacterial genomes; its performance depends on  
166 the sequencing depth and the quality of the NGS reads. For the datasets analyzed in this  
167 study (average read depth 133.5x – 598.2x ) and based on the assessed read quality >Q30  
168 = 88.2 – 92.7%, LoFreq has calling sensitivity = ~1% and PPV=100 [39]. Variants supported  
169 by absolute read concordance (>98%) were filtered-out from intra-host variant frequency  
170 calculations. Four SNVs from sample SRR10903402 and 3 SNVs from sample  
171 SRR10971381 with statistically significant strand bias (*P*-value < 0.05) were also excluded  
172 from further analyses. Variations were annotated to the reference genome using snpEff  
173 v4.3p [40], SNVs effects were further filtered with snpSift v4.3p [41] and the average  
174 mutation rate per gene across the viral genome was estimated using R scripts (v3.6.2) in  
175 RStudio v1.1.456. The colocalization of the intra-host SNVs and population level SNPs  
176 retrieved from www.GISAID.org on February 18<sup>th</sup> 2020, with primers and probes coordinates  
177 was also examined, to identify potential interferences with all currently available molecular  
178 diagnostic assays [42]. The impact of these SNVs on the binding affinity of primers and  
179 probes to their genomic targets, was predicted using FastPCR 3.3.28 [43] and DINAMelt



180 webserver [44]. To investigate intra-host genomic rearrangements, *de novo* assembly of the  
181 SARS-CoV-2 genomes was performed using Spades v3.13.1 [45]. Spades outperforms most  
182 modern *de novo* assemblers in terms of viral genome retrieval and coverage, presenting the  
183 highest sensitivity (99.48%) [46]. The resulting contigs were analyzed with BLAST v2.6.0  
184 [47] and confirmed by remapping of the raw reads, setting a threshold of 5 not replicated  
185 reads for contigs suggesting rearrangements. Smaller contigs (<200 bp) were elongated  
186 where possible, after pair-wise realignment of the corresponding mapped reads. Basic  
187 computations and visualizations were implemented in R programming language v3.6.2,  
188 using in-house scripts. The secondary structures of the genomic regions surrounding the  
189 recombination breakpoints were predicted using RNAfold webserver [48].

190

191

## 192 **RESULTS**

193 The mapping assembly of the viral genome was almost complete for all samples. The  
194 genome coverage and the average read depth across the genome was 100.0% and 133.5x  
195 for sample SRR10903401, 100.0% and 522.5x for sample SRR10903402, and 99.9%, and  
196 598.2x for sample SRR10971381, respectively (**Table 1**).

197 In all samples, the same 5 SNVs isolated with 98-100% read concordance, thus in  
198 total divergence with the reference genome (MN975262.1), were excluded from downstream  
199 analysis. For sample SRR10903401 34 lower frequency SNVs were isolated in total. Of  
200 these, 33 were present with frequencies ranking between 2 and 15%, while only one was  
201 present in 40% of the intra-host viral population. The sequencing depth, which is also  
202 evaluated during the SNV calling by the LoFreq algorithm, ranked between 39x and 290x at  
203 the corresponding SNV positions. The sequencing depth of sample SRR10903402 at the  
204 polymorphic positions was higher (103x – 1137x), allowing the isolation of 55 SNVs with  
205 frequencies distributed between 0.9% and 14%. The depth over the polymorphic positions of

206 sample SRR10971381 was between 159x – 1872x, allowing the isolation of 10 intra-host  
207 SNVs, with frequencies 1.1% - 6.8% (**Figure 1.A, Suppl.Table 1**).

208 Intra-host variants were distributed across 7 out of the 10 protein-coding genes of the  
209 viral genome, namely ORF1ab, S, ORF3a, ORF6, ORF7a, ORF8 and N. After normalising  
210 for the gene length (variants/kb-gene-length, “v/kbgl”), the density of the SNVs for each gene  
211 was estimated (**Table 2**). The majority of the SNPs corresponded to missense changes  
212 (leading to amino-acid change) compared to synonymous changes (cumulatively 72 vs. 29  
213 respectively, ratio 2.48:1) (**Table 2**), while the average number of missense changes was  
214 marginally significantly higher compared to synonymous changes (23,3 vs. 8,0 respectively,  
215 Wilcoxon rank sum test,  $p=0.054$ ). The average intra-host variant frequency did not differ  
216 significantly either between missense and synonymous polymorphisms (Wilcoxon rank sum  
217 test,  $p>0.05$ ) (**Figure 1.C**), or between their hosting genes (pairwise Wilcoxon rank sum  
218 tests,  $p>0.05$ ) (**Figure 1.D**). We did not detect any small-scale insertions or deletions in the  
219 samples (**Suppl. Table 1**).

220 The comparison of all SNVs (intra-host and population level) with the genomic  
221 targets of the molecular diagnostics assays, revealed colocalization of 3 intra-host SNVs and  
222 2 isolate-specific SNVs with primers and probes currently in use in RdRP\_SARSr, HKU-N,  
223 2019-nCoV-N1 and 2019-nCoV-N2 diagnostic reactions (**Figure 2**). The thermodynamic  
224 assessment of these SNVs revealed variable impact on the binding affinity of the  
225 corresponding primers and probes on the mutated genomic region (**Suppl. Table 2**)

226 The *de novo* assembly of the viral genomes was almost complete for samples  
227 SRR10903401 and SRR10903402 covering 99.7 % of the genome with 4 overlapping  
228 contigs and 99.5% of the genome with a single contig, respectively. The *de novo* assembly  
229 of sample SRR10971381 was complete, with one contig covering 100% of the genome.  
230 Alternative contigs revealed intra-host genomic rearrangements (**Figure 3, Table 3**). For  
231 samples SRR10903401 and SRR10903402, these large-scale structural events were  
232 systematically observed over poly-A / poly-U-rich genomic regions, located in ORF1ab and S  
233 genes. All rearrangements were validated by remapping of the raw reads on the

234 corresponding *de novo* assembled contigs, setting a threshold of at least 5 supporting reads  
235 of high mapping quality (>40) in each case. For sample SRR10903401 three  
236 inversions/misassemblies in ORF1ab (**Suppl. Figure 1**) and one inversion/misassembly in S  
237 gene (**Figure 4-A**) were isolated. Notably, we were able to validate the same inversion in S  
238 gene for sample SRR10903402 as well (**Figure 4-B**). Apart from 2 inversions in ORF1ab  
239 supported by only 2 reads each (not passing the validation threshold), there were no further  
240 large-scale intra-host events observed for sample SRR10903402. Similarly, one  
241 inversion/misassembly in sample SRR10971381 that was supported by only one read was  
242 identified. The alignment coordinates of all rearrangement-supporting contigs with respect to  
243 the reference strain are presented in **Table 3**.

244

245

246

## 247 **DISCUSSION**

248

249           The rapid spread and the death toll of the new SARS-CoV-2 epidemic warrants the  
250 immediate identification / development of effective antiviral agents and vaccines, and the  
251 design of accurate diagnostics as well. The intra- and inter- patient variability affects the  
252 compatibility of molecular diagnostics but also impairs the effectiveness of the vaccines and  
253 the serological assays by altering the antigenicity of the virus.

254           All samples analysed in this study were probably infected by the same viral strain  
255 since they shared the same set of consensus SNVs. However, apart from 3 intra-host SNVs  
256 that were common between SRR10903401 and SRR10903402, there was no other overlap  
257 observed between the low frequency variants of each sample (**Figure 1-B**). This indicates  
258 that these variations have occurred in a rather random fashion and are not subject to  
259 selective pressures, which is also supported by the fact that the missense mutations were  
260 systematically more, compared to the synonymous mutations [49]. On the other hand,  
261 missense substitutions are more common in loci involving pathogen resistance, indicating  
262 positive selection [50]. The analysed viral RNA might have originated from functional/packed  
263 virions, but also from unpacked viral genomes, unable to replicate and infect other host cells.  
264 Even if a viral genome is unable to replicate independently, its abundant presence in the  
265 pool of viral quasispecies implies some functionality regarding the intra-host evolution and  
266 adaptation. For example, defective viral genomes might affect infection dynamics such as  
267 viral persistence as well as the natural history of the infection [51,52]. At the same time,  
268 these variants may arise rapidly during an outbreak and can be used for tracking the  
269 transmission chains and the spatiotemporal characteristics of the epidemic [53–55]. More  
270 studies based on genomic datasets accompanied by clinical metadata are needed, in order  
271 to accurately define associations between intra-host SARS-CoV-2 genomic variants, the  
272 progression and the clinical outcome of COVID19.

273 SNVs and quasispecies observed at low frequency could represent viral variations of  
274 low impact on the functionality of the genome. Bal et. al, suggest that development of  
275 quasispecies may promote viral evolution, however high depth of coverage is essential for  
276 the study of intra-host adaptation [56]. The abundance of low-frequency variations is largely  
277 affected by the population size and the epidemic characteristics. For example, a neutral  
278 substitution in a region that represents a primer target for a molecular diagnostic assay can  
279 drift to fixation rather quickly in a rapidly spreading virus, jeopardizing the sensitivity of the  
280 assay [57,58]. Here, we highlight three intra-host but also two fixed variants that are  
281 colocalized with primers or probes of real-time PCR diagnostics assays that are currently in  
282 use (**Figure 2**). Since the binding affinity of these oligos to their genomic targets  
283 (**Suppl. Table 2**) is directly linked to the performance of the corresponding diagnostic assays,  
284 the community should pay extra attention in the evaluation of these potentially emerging  
285 variations and be alerted, in case redesigning of these oligos is needed.

286 As it is well documented, recombination events lead to substantial changes in genetic  
287 diversity of RNA viruses [49,59]. In CoVs, discontinuous RNA synthesis is commonly  
288 observed, resulting in high frequencies of homologous recombination [60], which can be up  
289 to 25% across the entire CoV genome [61]. For pathogenic HCoV genomic rearrangements  
290 are frequently reported during the course of epidemic outbreaks, such as HCoV-OC43 [62],  
291 and HCoV-NL63 [63], SARS-CoV [64][62] and MERS-CoV [65]. We have isolated intra-host  
292 genomic rearrangements, located in poly-A and poly-U enriched palindrome regions across  
293 the SARS-CoV-2 genome (**Figure 4**). We conclude that these rearrangements do not  
294 represent artifacts derived from the NGS library preparation (e.g. PCR crosstalk artifacts),  
295 especially since all the supporting reads were not duplicated and, in some cases, differed in  
296 polymorphic positions (**Suppl. Figure 1**).

297 Recombination processes involving S gene particularly, have been reported for  
298 SARS- and SARS-like CoV but also for HCoV-OC43. In the case of sister species HCoV-  
299 NL63 and HCoV-229E, recombination breakpoints are located near 3'- and 5'-end of the  
300 gene [1][65]. S is a trimeric protein, which is cleaved into two subunits, the globular N-

301 terminal S1 and the C-terminal S2 [66]. Our analysis revealed that similarly to other genomic  
302 regions, the S1 subunit hosts many low-frequency SNVs, characterized by higher density  
303 compared to the rest of the S gene sequence (**Figure 1-E**). The S2 subunit is highly  
304 conserved [13] and contains two fusion peptides (FP, IFP) [66]. In S gene, the same  
305 rearrangement event has taken place in two samples analyzed in this study, located in  
306 nt24,000, which corresponds to the ~200nt linking region between FP and IFP (aa 812-813).  
307 This observation highlights a potential recombination hot-spot. Examining closely the  
308 secondary structure of the RNA genome around the breakpoints, we suggest a model where  
309 the palindromes 5'-UGGUUUU-3' and 5'-AAAACCAA-3', have served as donor-acceptor  
310 sequences during the recombination event, since they are both exposed in the single-  
311 stranded internal loops formed in a highly structured RNA pseudoknot (**Figure 4-C**). The RB  
312 domain of the S protein has been tested as a potential immunogen as it contains  
313 neutralization epitopes which appear to have a role in the induction of neutralizing antibodies  
314 [31]. It should be mentioned though that the S protein of SARS-CoV is the most divergent in  
315 all strains infecting humans [67], as in both C and N-terminal domains variations arise  
316 rapidly, allowing immunological escape [68]. Our findings support that apart from these  
317 variations, the N-terminal region also hosts a recombination hot-spot, which together with the  
318 rest of the observed rearrangements, indicates the genomic instability of SARS-CoV-2 over  
319 poly-A and poly-U regions.

320

321

322 **Authors' contributions:** TK: Conceptualization; Data curation; Formal analysis;  
323 Methodology; Supervision; Validation; Visualization; Writing - original draft; Writing - review  
324 & editing. GP: Data curation; Formal analysis; Writing - original draft; Writing - review &  
325 editing. MB: Visualization; Writing - review & editing. AM, ST and AM: Writing - original draft;  
326 Writing - review & editing.

327

328 **Ethics approval and consent to participate:** Not applicable.

329

330 **Declarations of interest:** None.

331

332

333 This research did not receive any specific grant from funding agencies in the public,

334 commercial, or not-for-profit sectors.

335

## 336 **FIGURE LEGENTS**

337

338 **Figure 1:** Intra – host SNVs: (A) Intra host SNV frequency vs sequencing read depth (X  
339 coverage) in the corresponding alignment position. (B) Venn diagram representing unique  
340 and common SNVs isolated from the three patients (C) Boxplot of intra-host SNVs frequency  
341 vs. SNV type – synonymous, missense, nonsense (stop gained) (low, moderate and high  
342 impact respectively). Average values are in red rhombs. (D) Intra-host SNVs frequency vs.  
343 all seven genes affected (ORF1ab, S, ORF3a, ORF6, ORF7a, ORF8, N). Average values  
344 are in red rhombs. (E) Density histogram of intra-host SNVs isolated from all patients (total  
345 number of SNVs / 100 bp - blue bars) and average sequencing read depth (X coverage –  
346 green line), across the SARS-CoV-2 genome map (genes in orange, 5' and 3' untranslated  
347 regions in light blue).

348

349 **Figure 2:** Truncated map of SARS-CoV-2 genome illustrating a subset of intra-host (blue  
350 lines) and globally collected, isolate-specific SNVs (orange lines) with respect to the genomic  
351 targets of molecular diagnostics assays (red arrows – primers, red bars - probes). Three  
352 intra-host variants (orange triangles), and two strain specific variants (Wuhan/IVD-HB-  
353 04/2020 and Chongqing/YC01/2020 - red triangles), are colocalized with the RdRP\_SARSr  
354 probe (15,474 T > G), the 2019-nCoV\_N1 forward primer (28,291 C > T), the HKU-N reverse  
355 primer (28,971 A > G) and the 2019-nCoV-N2 probe (29,188 T > C and 29,200 C > T).

356

357 **Figure 3:** Alignment of the de novo assembled contigs on the genomic map (bottom).  
358 Concordantly aligned contigs (correct or gapped) are in green, while discordantly aligned  
359 contigs are in red. Sequencing read depth (X coverage) across the genome (blue  
360 histograms) and relative % GC content (green line) is presented for each sample.

361

362 **Figure 4:** Recombination events in S gene. Samples (A) SRR10903401 and (B)  
363 SRR10903402. Alignments of the de novo assembled contigs with respect to the reference  
364 genome (MN 975262). Donor – acceptor palindrome sequences are indicated in green bars.  
365 Raw, non-duplicated NGS reads, validating the recombination event, are represented below  
366 the corresponding contig. (C): Prediction of the secondary structure of the genomic region  
367 spanning the rearrangement breakpoint (100 bases upstream and 100 bases downstream).  
368 The corresponding donor- acceptor sequences, exposed in internal loops, are indicated in  
369 green bars.

370



371 **REFERENCES**

- 372 [1] V.M. Corman, D. Muth, D. Niemeyer, C. Drosten, Hosts and Sources of Endemic  
373 Human Coronaviruses, 100 (2018) 163–188.  
374 <https://doi.org/10.1016/bs.aivir.2018.01.001>.
- 375 [2] A.R. Fehr, S. Perlman, HHS Public Access, (2016) 1–23. [https://doi.org/10.1007/978-](https://doi.org/10.1007/978-1-4939-2438-7)  
376 [1-4939-2438-7](https://doi.org/10.1007/978-1-4939-2438-7).
- 377 [3] D. Vijaykrishna, G.J.D. Smith, J.X. Zhang, J.S.M. Peiris, H. Chen, Y. Guan,  
378 Evolutionary Insights into the Ecology of Coronaviruses, *J. Virol.* 81 (2007) 4012–  
379 4020. <https://doi.org/10.1128/jvi.02605-06>.
- 380 [4] F. Li, W. Li, M. Farzan, S.C. Harrison, Structural biology: Structure of SARS  
381 coronavirus spike receptor-binding domain complexed with receptor, *Science* (80-. ).  
382 309 (2005) 1864–1868. <https://doi.org/10.1126/science.1116480>.
- 383 [5] C.I. Paules, Coronavirus Infections — More Than Just the Common Cold, 2520  
384 (2020) 3–4. <https://doi.org/10.1007/82>.
- 385 [6] J. Cui, Origin and evolution of pathogenic coronaviruses, *Nat. Rev. Microbiol.* 17  
386 (2019) 181–192. <https://doi.org/10.1038/s41579-018-0118-9>.
- 387 [7] J.H. Epstein, J. McEachern, J. Zhang, P. Daszak, H. Wang, H. Field, W. Li, B.T.  
388 Eaton, L.-F. Wang, M. Yu, Z. Hu, S. Zhang, Z. Shi, G. Crameri, H. Zhang, W. Ren, C.  
389 Smith, Bats Are Natural Reservoirs of SARS-Like Coronaviruses, *Science* (80-. ). 310  
390 (2005) 676–679. <https://doi.org/10.1109/NEMS.2006.334722>.
- 391 [8] Z.A. Memish, S. Perlman, M.D. Van Kerkhove, A. Zumla, Middle East respiratory  
392 syndrome., *Lancet* (London, England). 395 (2020) 1063–1077.  
393 [https://doi.org/10.1016/S0140-6736\(19\)33221-0](https://doi.org/10.1016/S0140-6736(19)33221-0).
- 394 [9] E. Prompetchara, C. Ketloy, T. Palaga, Immune responses in COVID-19 and potential  
395 vaccines: Lessons learned from SARS and MERS epidemic., *Asian Pacific J. Allergy*  
396 *Immunol.* 38 (2020) 1–9. <https://doi.org/10.12932/AP-200220-0772>.
- 397 [10] F.S. Aleanizy, N. Mohamed, F.Y. Alqahtani, R.A. El Hadi Mohamed, Outbreak of  
398 Middle East respiratory syndrome coronavirus in Saudi Arabia: a retrospective study,  
399 *BMC Infect. Dis.* 17 (2017) 23. <https://doi.org/10.1186/s12879-016-2137-3>.
- 400 [11] WHO | Pneumonia of unknown cause –[www.who.int/csr/don/05-january-2020-](http://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/)  
401 [pneumonia-of-unkown-cause-china/](http://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/), (2020). [https://www.who.int/csr/don/05-january-](https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/#.XyAbsvYF-rc.mendeley)  
402 [2020-pneumonia-of-unkown-cause-china/en/#.XyAbsvYF-rc.mendeley](https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/#.XyAbsvYF-rc.mendeley) (accessed July  
403 28, 2020).
- 404 [12] GISAID | [www.gisaid.org/epiflu-applications/global-cases-betacov](http://www.gisaid.org/epiflu-applications/global-cases-betacov), (n.d.).
- 405 [13] J.F. Chan, K. Kok, Z. Zhu, H. Chu, K. Kai-wang, S. Yuan, K. Yuen, Genomic  
406 characterization of the 2019 novel human-pathogenic coronavirus isolated from a

- 407 patient with atypical pneumonia after visiting Wuhan, 1751 (2020).  
408 <https://doi.org/10.1080/22221751.2020.1719902>.
- 409 [14] A.A. Rabaan, S.H. Al-Ahmed, S. Haque, R. Sah, R. Tiwari, Y.S. Malik, K. Dhama, M.I.  
410 Yattoo, D.K. Bonilla-Aldana, A.J. Rodriguez-Morales, SARS-CoV-2, SARS-CoV, and  
411 MERS-COV: A comparative overview., *Le Infez. Med.* 28 (2020) 174–184.
- 412 [15] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu,  
413 Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y.  
414 Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W.J. Liu, D. Wang, W. Xu, E.C.  
415 Holmes, G.F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterisation and  
416 epidemiology of 2019 novel coronavirus: implications for virus origins and receptor  
417 binding., *Lancet (London, England)*. 395 (2020) 565–574.  
418 [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8).
- 419 [16] R. Antia, R.R. Regoes, J.C. Koella, C.T. Bergstrom, The role of evolution in the  
420 emergence of infectious diseases, *Nature*. 426 (2003) 658–661.  
421 <https://doi.org/10.1038/nature02104>.
- 422 [17] C. Taştan, B. Yurtsever, G. Sir Karakuş, D. Dilek Kançaği, S. Demir, S. Abanuz, U.  
423 Seyis, M. Yildirim, R. Kuzay, Ö. Eilbol, S. Arbak, M. Açikel Elmas, S. Bırdoğan, O.U.  
424 Sezerman, A.S. Kocagöz, K. Yalçın, E. Ovalı, SARS-CoV-2 isolation and propagation  
425 from Turkish COVID-19 patients, *Turkish J. Biol. = Turk Biyol. Derg.* 44 (2020) 192–  
426 202. <https://doi.org/10.3906/biy-2004-113>.
- 427 [18] A. Beloukas, E. Magiorkinis, G. Magiorkinis, A. Zavitsanou, T. Karamitros, A.  
428 Hatzakis, D. Paraskevis, Assessment of phylogenetic sensitivity for reconstructing  
429 HIV-1 epidemiological relationships, *Virus Res.* 166 (2012) 54–60.
- 430 [19] M. Vignuzzi, J.K. Stone, J.J. Arnold, C.E. Cameron, R. Andino, Quasispecies diversity  
431 determines pathogenesis through cooperative interactions in a viral population,  
432 *Nature*. 439 (2006) 344–348. <https://doi.org/10.1038/nature04388>.
- 433 [20] T. Karamitros, G. Papatheodoridis, E. Dimopoulou, M.-V. Papageorgiou, D.  
434 Paraskevis, G. Magiorkinis, V. Sypsa, A. Hatzakis, The interferon receptor-1 promoter  
435 polymorphisms affect the outcome of Caucasians with HB eAg-negative chronic HBV  
436 infection, *Liver Int.* 35 (2015) 2506–2513.
- 437 [21] K.S. Xue, L.H. Moncla, T. Bedford, J.D. Bloom, Within-Host Evolution of Human  
438 Influenza Virus, *Trends Microbiol.* 26 (2018) 781–793.  
439 <https://doi.org/https://doi.org/10.1016/j.tim.2018.02.007>.
- 440 [22] G. Destras, M. Pichon, B. Simon, M. Valette, V. Escuret, P.-A. Bolze, G. Dubernard,  
441 P. Gaucherand, B. Lina, L. Josset, Impact of Pregnancy on Intra-Host Genetic  
442 Diversity of Influenza A Viruses in Hospitalised Women: A Retrospective Cohort  
443 Study., *J. Clin. Med.* 9 (2019). <https://doi.org/10.3390/jcm9010058>.

- 444 [23] B. Simon, M. Pichon, M. Valette, G. Burfin, M. Richard, B. Lina, L. Josset, Whole  
445 Genome Sequencing of A(H3N2) Influenza Viruses Reveals Variants Associated with  
446 Severity during the 2016–2017 Season, *Viruses*. 11 (2019) 108.  
447 <https://doi.org/10.3390/v11020108>.
- 448 [24] Y. Toyoshima, K. Nemoto, S. Matsumoto, Y. Nakamura, K. Kiyotani, SARS-CoV-2  
449 genomic variations associated with mortality rate of COVID-19, *J. Hum. Genet.*  
450 (2020). <https://doi.org/10.1038/s10038-020-0808-9>.
- 451 [25] S. Bonhoeffer, M.A. Nowak, Intra-host versus inter-host selection: viral strategies of  
452 immune function impairment, *Proc. Natl. Acad. Sci. U. S. A.* 91 (1994) 8062–8066.  
453 <https://doi.org/10.1073/pnas.91.17.8062>.
- 454 [26] T. Day, S. Gandon, S. Lion, S.P. Otto, On the evolutionary epidemiology of SARS-  
455 CoV-2, *Curr. Biol.* (2020). <https://doi.org/https://doi.org/10.1016/j.cub.2020.06.031>.
- 456 [27] E.C. Holmes, The comparative genomics of viral emergence, *Proc. Natl. Acad. Sci.*  
457 107 (2010) 1742–1746. <https://doi.org/10.1073/PNAS.0906193106>.
- 458 [28] S. Kumar, R. Nyodu, V.K. Maurya, S.K. Saxena, Host Immune Response and  
459 Immunobiology of Human SARS-CoV-2 Infection, *Coronavirus Dis. 2019 Epidemiol.*  
460 *Pathog. Diagnosis, Ther.* (2020) 43–53. [https://doi.org/10.1007/978-981-15-4814-7\\_5](https://doi.org/10.1007/978-981-15-4814-7_5).
- 461 [29] R. Malley, K. Trzcinski, A. Srivastava, C.M. Thompson, P.W. Anderson, M. Lipsitch,  
462 CD4+ T cells mediate antibody-independent acquired immunity to pneumococcal  
463 colonization, *Proc. Natl. Acad. Sci. U. S. A.* 102 (2005) 4848–4853.  
464 <https://doi.org/10.1073/pnas.0501254102>.
- 465 [30] M. Lipsitch, J.J. O’Hagan, Patterns of antigenic diversity and the mechanisms that  
466 maintain them, *J. R. Soc. Interface.* 4 (2007) 787–802.  
467 <https://doi.org/10.1098/rsif.2007.0229>.
- 468 [31] C.Y. Yong, H.K. Ong, S.K. Yeap, K.L. Ho, W.S. Tan, Recent Advances in the Vaccine  
469 Development Against Middle East Respiratory Syndrome-Coronavirus, *Front.*  
470 *Microbiol.* 10 (2019) 1–18. <https://doi.org/10.3389/fmicb.2019.01781>.
- 471 [32] R.J. Nicolas W. Cortes-Penfield, Barbara W. Trautner, A decade after SARS:  
472 Strategies to control emerging coronaviruses, *Physiol. Behav.* 176 (2017) 139–148.  
473 <https://doi.org/10.1016/j.physbeh.2017.03.040>.
- 474 [33] L. Chen, W. Liu, Q. Zhang, K. Xu, G. Ye, W. Wu, Z. Sun, F. Liu, K. Wu, B. Zhong, Y.  
475 Mei, W. Zhang, Y. Chen, Y. Li, M. Shi, K. Lan, Y. Liu, RNA based mNGS approach  
476 identifies a novel human coronavirus from two individual pneumonia cases in 2019  
477 Wuhan outbreak, *Emerg. Microbes Infect.* 9 (2020) 313–319.  
478 <https://doi.org/10.1080/22221751.2020.1725399>.
- 479 [34] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H.  
480 Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng,

481 L. Xu, E.C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human  
482 respiratory disease in China, *Nature*. 579 (2020) 265–269.  
483 <https://doi.org/10.1038/s41586-020-2008-3>.

484 [35] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, *Nat.*  
485 *Methods*. 9 (2012) 357–359. <https://doi.org/10.1038/nmeth.1923>.

486 [36] S. Andrews, FastQC: A Quality Control Tool for High Throughput Sequence Data  
487 [Online]. Available online at:  
488 <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>, (2010).

489 [37] J.T. Robinson, H. Thorvaldsdóttir, A.M. Wenger, A. Zehir, J.P. Mesirov, Variant review  
490 with the integrative genomics viewer, *Cancer Res*. 77 (2017) e31–e34.  
491 <https://doi.org/10.1158/0008-5472.CAN-17-0337>.

492 [38] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G.  
493 Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools,  
494 *Bioinformatics*. 25 (2009) 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>.

495 [39] A. Wilm, P.P.K. Aw, D. Bertrand, G.H.T. Yeo, S.H. Ong, C.H. Wong, C.C. Khor, R.  
496 Petric, M.L. Hibberd, N. Nagarajan, LoFreq: A sequence-quality aware, ultra-sensitive  
497 variant caller for uncovering cell-population heterogeneity from high-throughput  
498 sequencing datasets, *Nucleic Acids Res*. 40 (2012) 11189–11201.  
499 <https://doi.org/10.1093/nar/gks918>.

500 [40] P. Cingolani, A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu,  
501 D.M. Ruden, A program for annotating and predicting the effects of single nucleotide  
502 polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain  
503 w1118; iso-2; iso-3, *Fly (Austin)*. 6 (2012) 80–92. <https://doi.org/10.4161/fly.19695>.

504 [41] P. Cingolani, V.M. Patel, M. Coon, T. Nguyen, S.J. Land, D.M. Ruden, X. Lu, Using  
505 *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a  
506 new program, SnpSift, *Front. Genet*. 3 (2012).  
507 <https://doi.org/10.3389/fgene.2012.00035>.

508 [42] V.M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D.K. Chu, T. Bleicker, S.  
509 Brünink, J. Schneider, M.L. Schmidt, D.G. Mulders, B.L. Haagmans, B. van der Veer,  
510 S. van den Brink, L. Wijsman, G. Goderski, J.L. Romette, J. Ellis, M. Zambon, M.  
511 Peiris, H. Goossens, C. Reusken, M.P. Koopmans, C. Drosten, Detection of 2019  
512 novel coronavirus (2019-nCoV) by real-time RT-PCR, *Euro Surveill*. 25 (2020).  
513 <https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045>.

514 [43] R. Kalendar, B. Khassenov, Y. Ramankulov, O. Samuilova, K.I. Ivanov, FastPCR: An  
515 in silico tool for fast primer and probe design and advanced sequence analysis,  
516 *Genomics*. 109 (2017) 312–319.  
517 <https://doi.org/https://doi.org/10.1016/j.ygeno.2017.05.005>.

- 518 [44] N.R. Markham, M. Zuker, DINAMelt web server for nucleic acid melting prediction,  
519 Nucleic Acids Res. 33 (2005) W577–W581. <https://doi.org/10.1093/nar/gki591>.
- 520 [45] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M.  
521 Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N.  
522 Vyahhi, G. Tesler, M.A. Alekseyev, P.A. Pevzner, SPAdes: A new genome assembly  
523 algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012)  
524 455–477. <https://doi.org/10.1089/cmb.2012.0021>.
- 525 [46] T.D.S. Sutton, A.G. Clooney, F.J. Ryan, R.P. Ross, C. Hill, Choice of assembly  
526 software has a critical impact on virome characterisation, Microbiome. 7 (2019) 12.  
527 <https://doi.org/10.1186/s40168-019-0626-5>.
- 528 [47] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment  
529 search tool, J. Mol. Biol. 215 (1990) 403–410. [https://doi.org/10.1016/S0022-](https://doi.org/10.1016/S0022-2836(05)80360-2)  
530 [2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- 531 [48] A.R. Gruber, R. Lorenz, S.H. Bernhart, R. Neuböck, I.L. Hofacker, The Vienna RNA  
532 websuite., Nucleic Acids Res. 36 (2008) W70. <https://doi.org/10.1093/nar/gkn188>.
- 533 [49] E.-G. Kostaki, T. Karamitros, M. Bobkova, M. Oikonomopoulou, G. Magiorkinis, F.  
534 Garcia, A. Hatzakis, D. Paraskevis, Spatiotemporal characteristics of the HIV-1  
535 CRF02\_AG/CRF63\_02A1 epidemic in Russia and Central Asia, AIDS Res. Hum.  
536 Retroviruses. 34 (2018) 415–420.
- 537 [50] D.N. Cooper, N.P. Group., Nature encyclopedia of the human genome, Nature Pub.  
538 Group, London; New York, 2003.
- 539 [51] J. Aaskov, K. Buzacott, H.M. Thu, K. Lowry, E.C. Holmes, Long-term transmission of  
540 defective RNA viruses in humans and Aedes mosquitoes, Science (80-. ). 311 (2006)  
541 236–238. <https://doi.org/10.1126/science.1115030>.
- 542 [52] E. Karamichali, H. Chihab, A. Kakkanas, A. Marchio, T. Karamitros, V. Pogka, A.  
543 Varaklioti, A. Kalliaropoulos, B. Martinez-Gonzales, P. Foka, others, HCV defective  
544 genomes promote persistent infection by modulating the viral life cycle, Front.  
545 Microbiol. 9 (2018) 2942.
- 546 [53] G. Magiorkinis, D. Paraskevis, O.-G. Pybus, T. Karamitros, T. Vasylyeva, M.  
547 Bobkova, A. Hatzakis, HIV-1 epidemic in Russia: an evolutionary epidemiology  
548 analysis, Lancet. 383 (2014) S71.
- 549 [54] D. Paraskevis, G.K. Nikolopoulos, V. Sypsa, M. Psychogiou, K. Pantavou, E. Kostaki,  
550 T. Karamitros, D. Paraskeva, J. Schneider, M. Malliori, others, Molecular investigation  
551 of HIV-1 cross-group transmissions during an outbreak among people who inject  
552 drugs (2011--2014) in Athens, Greece, Infect. Genet. Evol. 62 (2018) 11–16.
- 553 [55] G. Magiorkinis, T. Karamitros, T.I. Vasylyeva, L.D. Williams, J.L. Mbisa, A. Hatzakis,  
554 D. Paraskevis, S.R. Friedman, An innovative study design to assess the community

555 effect of interventions to mitigate HIV epidemics using transmission-chain  
556 phylodynamics, *Am. J. Epidemiol.* 187 (2018) 2615–2622.

557 [56] A. Bal, G. Destras, A. Gaymard, M. Bouscambert-Duchamp, M. Valette, V. Escuret, E.  
558 Frobert, G. Billaud, S. Trouillet-Assant, V. Cheynet, K. Brengel-Pesce, F. Morfin, B.  
559 Lina, L. Josset, Molecular characterization of SARS-CoV-2 in the first COVID-19  
560 cluster in France reveals an amino acid deletion in nsp2 (Asp268del), *Clin. Microbiol.*  
561 *Infect.* 26 (2020) 960–962. <https://doi.org/https://doi.org/10.1016/j.cmi.2020.03.020>.

562 [57] J.Y. Noh, S.-W. Yoon, D.-J. Kim, M.-S. Lee, J.-H. Kim, W. Na, D. Song, D.G. Jeong,  
563 H.K. Kim, Simultaneous detection of severe acute respiratory syndrome, Middle East  
564 respiratory syndrome, and related bat coronaviruses by real-time reverse transcription  
565 PCR., *Arch. Virol.* 162 (2017) 1617–1623. <https://doi.org/10.1007/s00705-017-3281-9>.

566 [58] T. Karamitros, D. Paraskevis, A. Hatzakis, M. Psychogiou, I. Elefsiniotis, T. Hurst, A.M.  
567 Geretti, A. Beloukas, J. Frater, P. Klenerman, others, A contaminant-free assessment  
568 of Endogenous Retroviral RNA in human plasma., *Nat. Sci. Reports.* 6 (2016) 33598.

569 [59] E.C. Holmes, *The evolution and emergence of RNA viruses*, Oxford University Press,  
570 2009.

571 [60] M.M.C. Lai, RNA recombination in animal and plant viruses, *Microbiol. Rev.* 56 (1992)  
572 61–79. <https://doi.org/10.1128/membr.56.1.61-79.1992>.

573 [61] R.S. Baric, K. Fu, M.C. Schaad, S.A. Stohlman, Establishing a genetic recombination  
574 map for murine coronavirus strain A59 complementation groups, *Virology.* 177 (1990)  
575 646–656. [https://doi.org/10.1016/0042-6822\(90\)90530-5](https://doi.org/10.1016/0042-6822(90)90530-5).

576 [62] S.K.P. Lau, P. Lee, A.K.L. Tsang, C.C.Y. Yip, H. Tse, R.A. Lee, L.-Y. So, Y.-L. Lau,  
577 K.-H. Chan, P.C.Y. Woo, K.-Y. Yuen, Molecular Epidemiology of Human Coronavirus  
578 OC43 Reveals Evolution of Different Genotypes over Time and Recent Emergence of  
579 a Novel Genotype due to Natural Recombination, *J. Virol.* 85 (2011) 11325–11337.  
580 <https://doi.org/10.1128/jvi.05512-11>.

581 [63] K. Pyrc, R. Dijkman, L. Deng, M.F. Jebbink, H.A. Ross, B. Berkhout, L. van der Hoek,  
582 Mosaic Structure of Human Coronavirus NL63, One Thousand Years of Evolution, *J.*  
583 *Mol. Biol.* 364 (2006) 964–973. <https://doi.org/10.1016/j.jmb.2006.09.074>.

584 [64] C.-C. Hon, T.-Y. Lam, Z.-L. Shi, A.J. Drummond, C.-W. Yip, F. Zeng, P.-Y. Lam, F.C.-  
585 C. Leung, Evidence of the Recombinant Origin of a Bat Severe Acute Respiratory  
586 Syndrome (SARS)-Like Coronavirus and Its Implications on the Direct Ancestor of  
587 SARS Coronavirus, *J. Virol.* 82 (2008) 1819–1826. [https://doi.org/10.1128/jvi.01926-](https://doi.org/10.1128/jvi.01926-07)  
588 07.

589 [65] J.S.M. Sabir, T.T.Y. Lam, M.M.M. Ahmed, L. Li, Y. Shen, S.E.M. Abo-Aba, M.I.  
590 Qureshi, M. Abu-Zeid, Y. Zhang, M.A. Khiyami, N.S. Alharbi, N.H. Hajrah, M.J. Sabir,  
591 M.H.Z. Mutwakil, S.A. Kabli, F.A.S. Alsulaimany, A.Y. Obaid, B. Zhou, D.K. Smith,

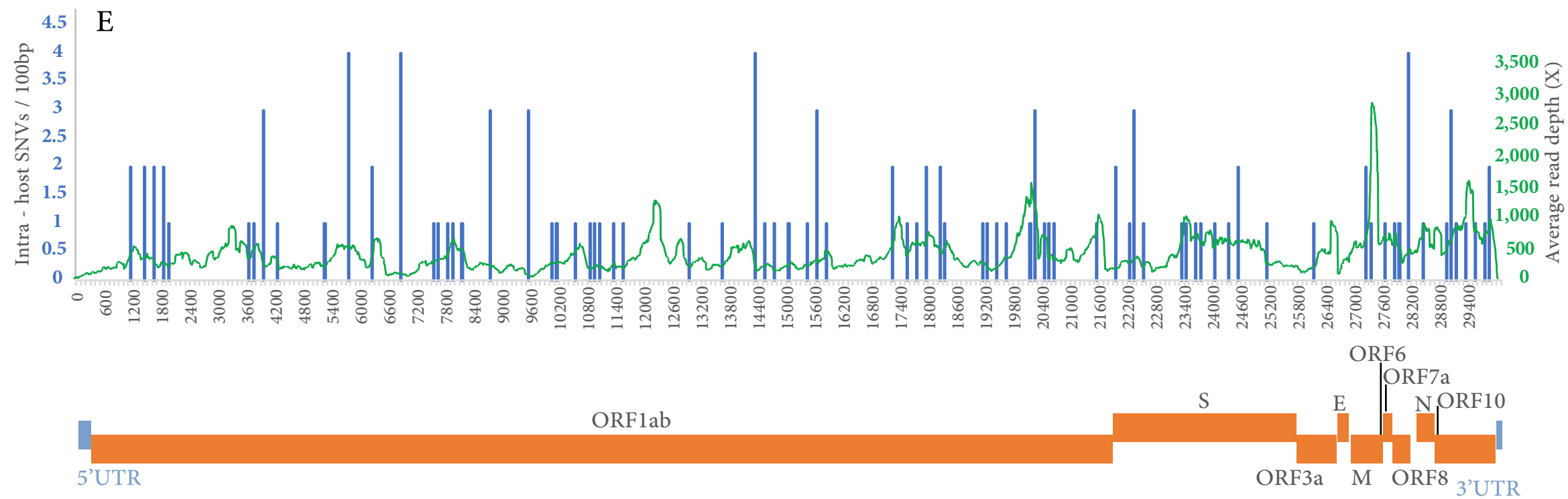
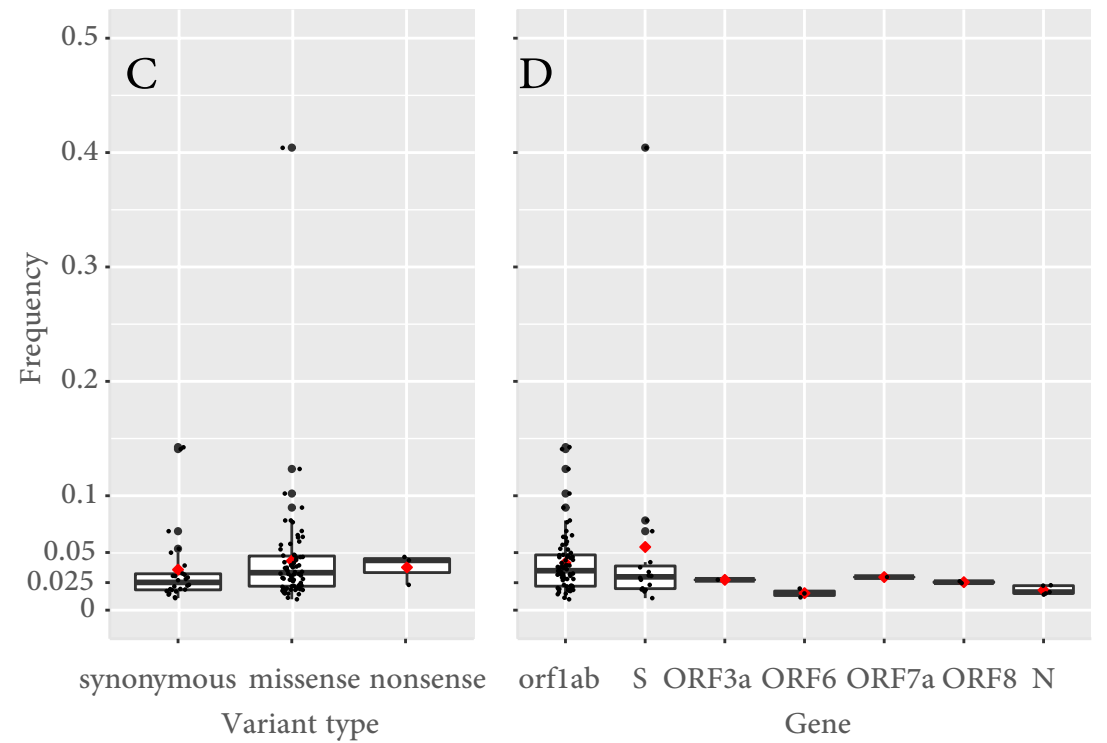
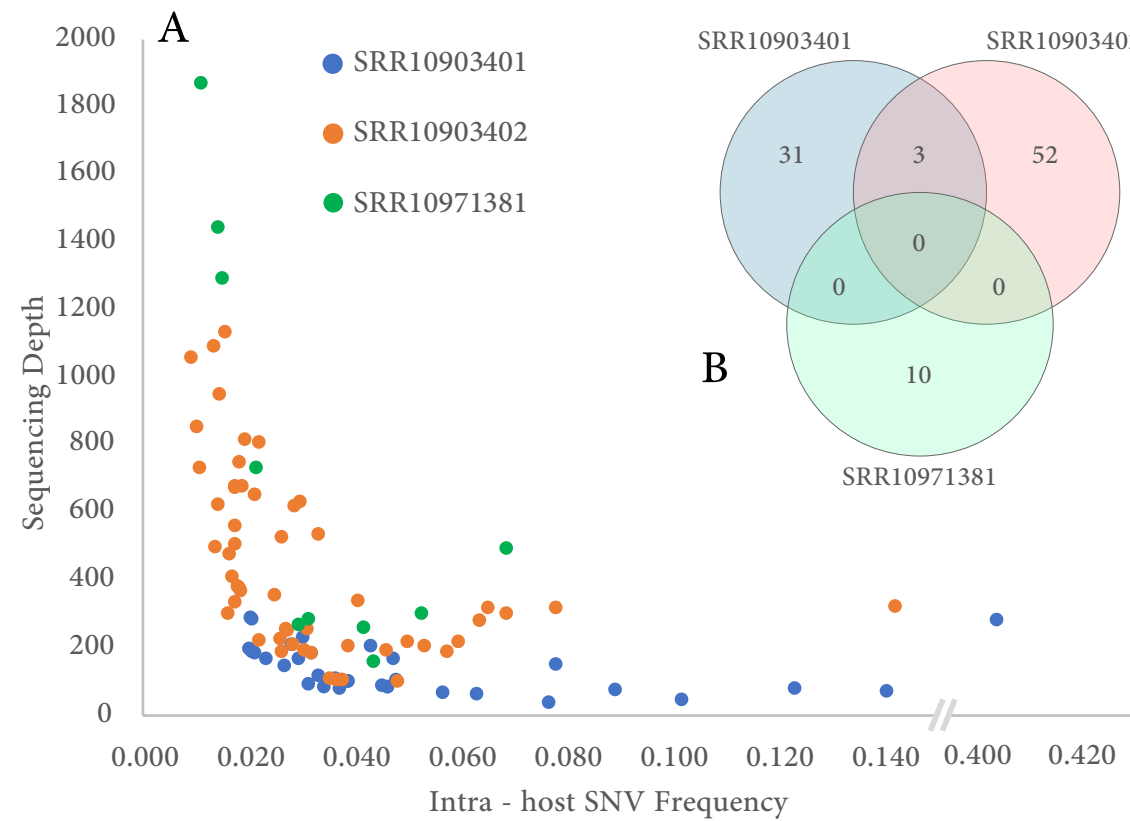
592 E.C. Holmes, H. Zhu, Y. Guan, Co-circulation of three camel coronavirus species and  
593 recombination of MERS-CoVs in Saudi Arabia, *Science* (80-. ). 351 (2016) 81–84.  
594 <https://doi.org/10.1126/science.aac8608>.

595 [66] G. Lu, Q. Wang, G.F. Gao, Bat-to-human: Spike features determining “host jump” of  
596 coronaviruses SARS-CoV, MERS-CoV, and beyond, *Trends Microbiol.* 23 (2015)  
597 468–478. <https://doi.org/10.1016/j.tim.2015.06.003>.

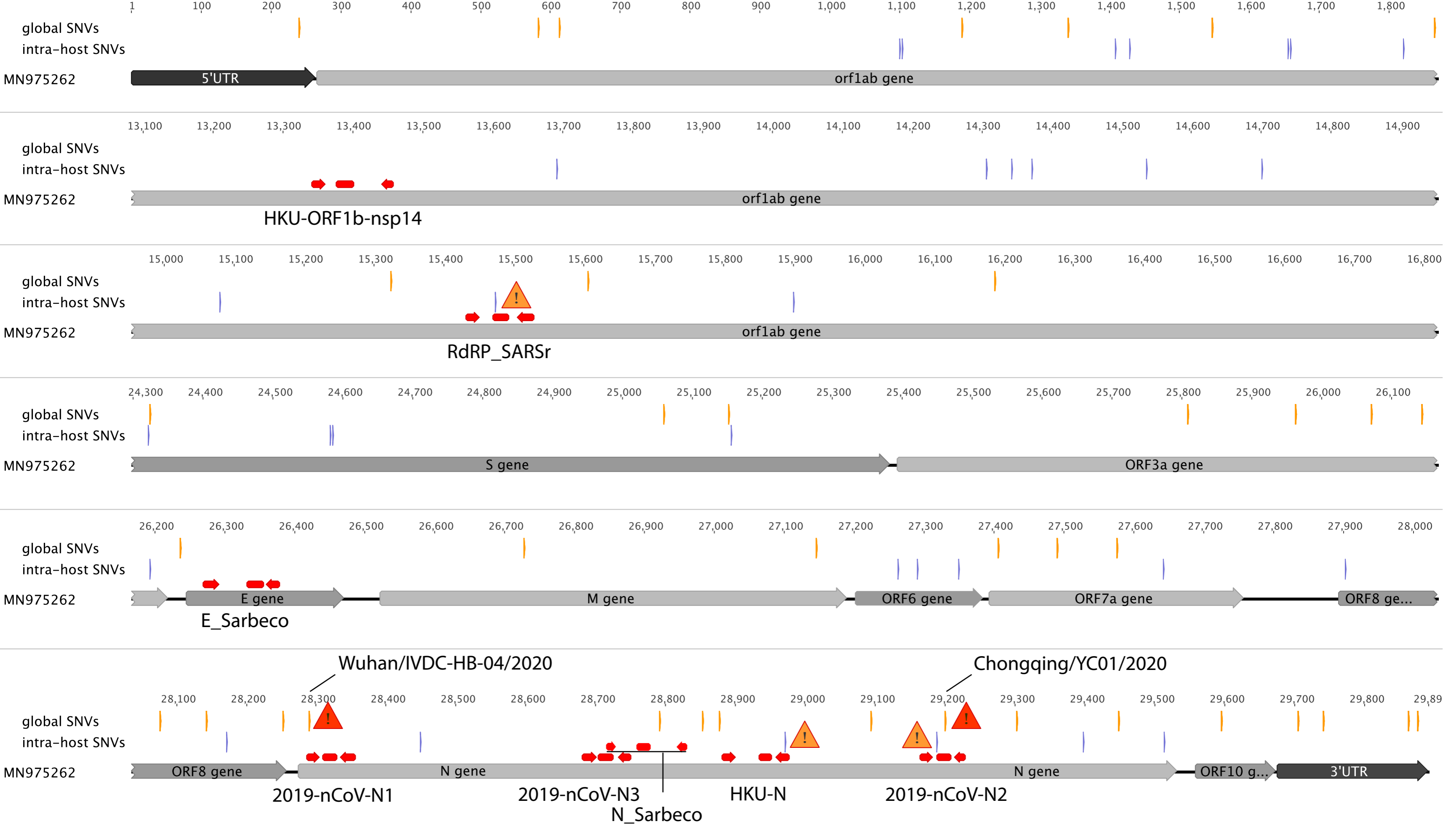
598 [67] L. Enjuanes, M.L. DeDiego, E. Álvarez, D. Deming, T. Sheahan, R. Baric, Vaccines to  
599 prevent severe acute respiratory syndrome coronavirus-induced disease, *Virus Res.*  
600 133 (2008) 45–62. <https://doi.org/10.1016/j.virusres.2007.01.021>.

601 [68] L. Jiaming, Y. Yanfeng, D. Yao, H. Yawei, B. Linlin, H. Baoying, Y. Jinghua, G.F. Gao,  
602 Q. Chuan, T. Wenjie, The recombinant N-terminal domain of spike proteins is a  
603 potential vaccine against Middle East respiratory syndrome coronavirus (MERS-CoV)  
604 infection, *Vaccine.* 35 (2017) 10–18. <https://doi.org/10.1016/j.vaccine.2016.11.064>.

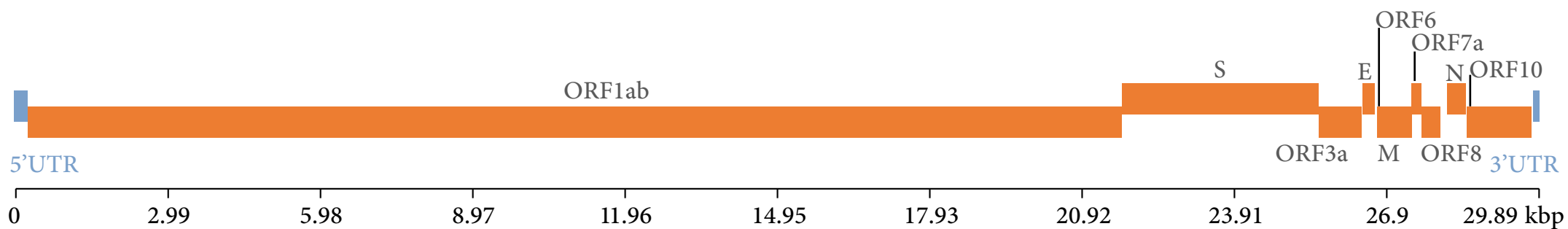
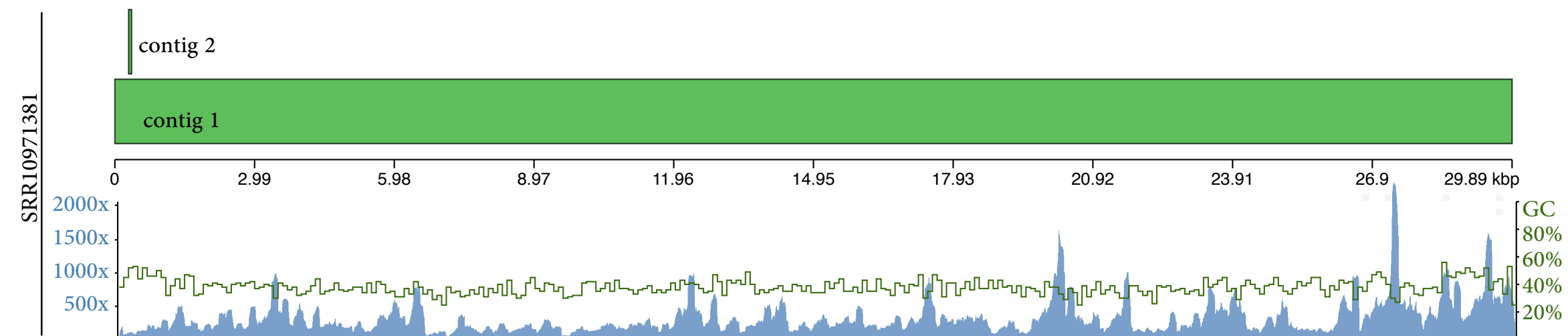
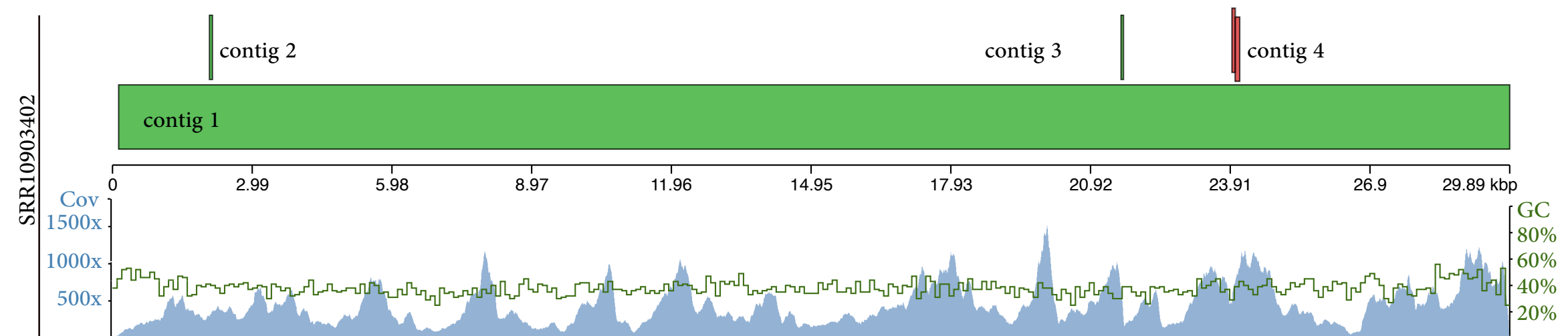
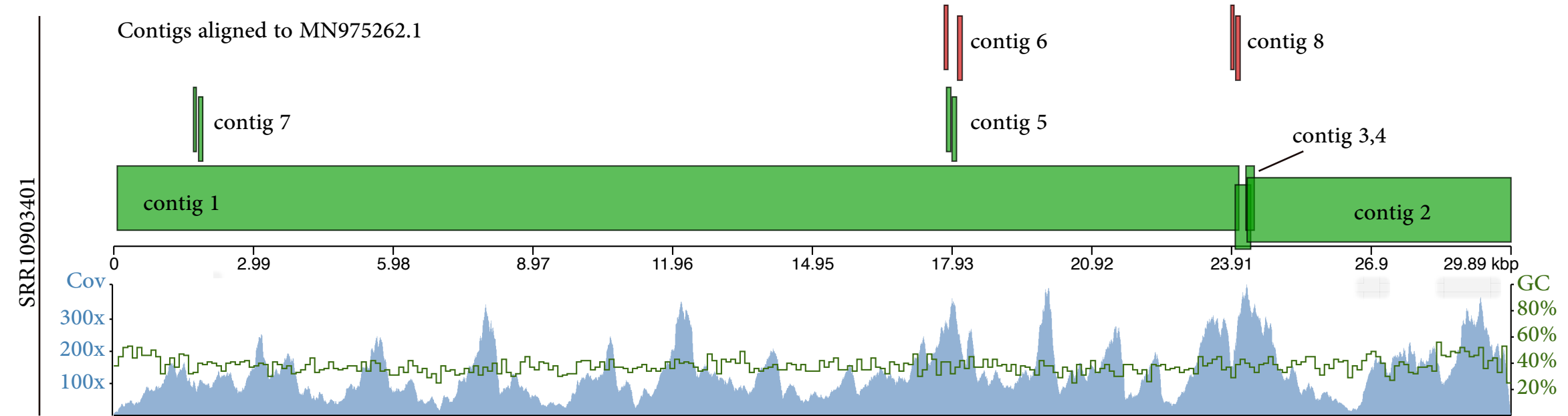
605  
606  
607  
608







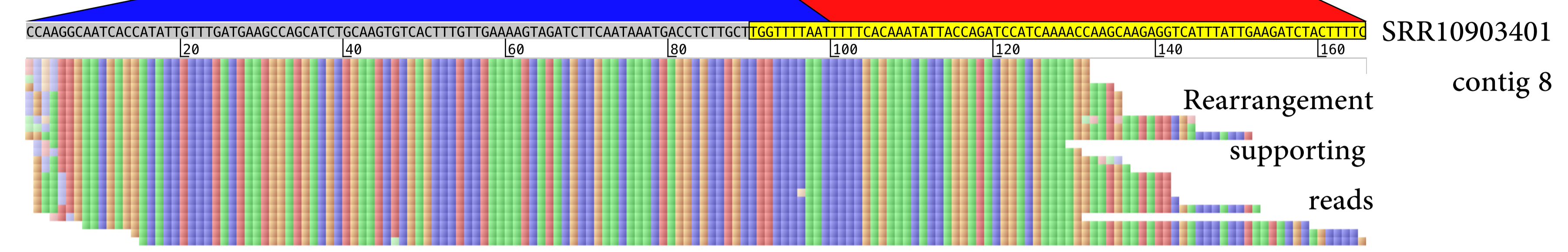
Contigs aligned to MN975262.1



CAAACAAATTTACAAAACACCACCAATTAAGATTTGGTGGTTTTAATTTTCACAAATATTACCAGATCCATCAAACCAAGCAAGAGGTCATTTATTGAAGATCTACTTTTCAAAAGTGACACTTGCAGATGCTGGCTTCATCAACAATATGGTGATTGCCTTGGTGATATTGCTGCTAGAGACCTCATTGTGCACAAAAGT  
 23920 23940 23960 23980 24000 24020 24040 24060 24080 24100 24120  
 GTTTGTTTAAATGTTTTGTGGTGGTTAATTTCTAAAACACCAAAAATAAAAAGTGTTTATAATGGTCTAGGTAGTTTTGGTTCGTTCTCCAGTAAATAACTTCTAGATGAAAAGTTGTTTCACGTGGAACGCTACGACCGAAGTAGTTTGTATACCACTAACGGAACCACTATAACGACGATCTCTGGAGTAAACACGTGTTTTCA/

Ref. Strain MN975262.1

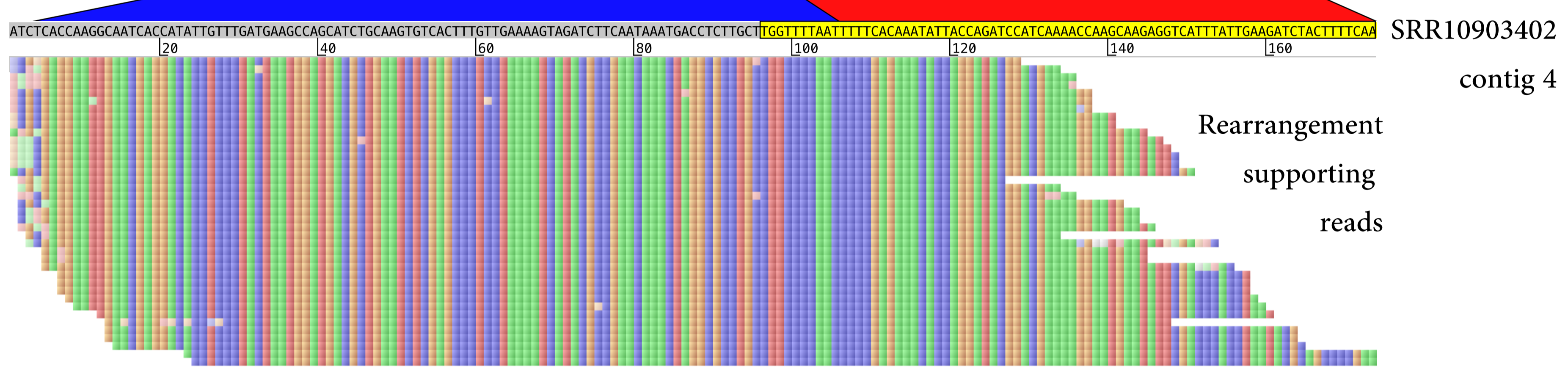
A



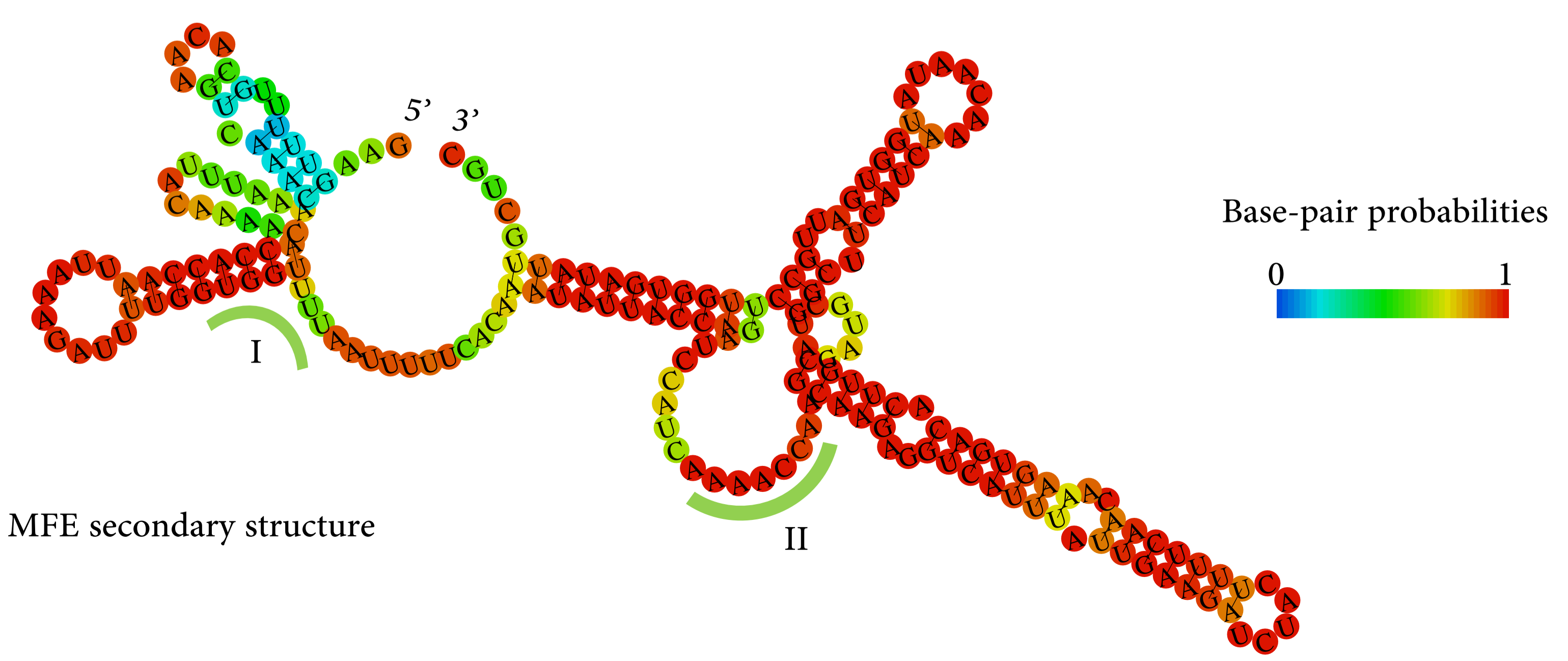
CAAACAAATTTACAAAACACCACCAATTAAGATTTGGTGGTTTTAATTTTCACAAATATTACCAGATCCATCAAACCAAGCAAGAGGTCATTTATTGAAGATCTACTTTTCAAAGTGACACTTGCAGATGCTGGCTTCATCAACAATATGGTGATTGCCTTGGTGATATTGCTGCTAGAGACCTCATTGTGCACAAAAGT  
 23920 23940 23960 23980 24000 24020 24040 24060 24080 24100 24120  
 GTTTGTTTAAATGTTTTGTGGTGGTTAATTTCTAAAACACCAAAAATAAAAAGTGTTTATAATGGTCTAGGTAGTTTTGGTTCGTTCTCCAGTAAATAACTTCTAGATGAAAAGTTGTTTCACGTGGAACGCTACGACCGAAGTAGTTTGTATACCACTAACGGAACCACTATAACGACGATCTCTGGAGTAAACACGTGTTTTCA/

Ref. Strain MN975262.1

B



C



MFE secondary structure

## TABLES

**Table 1: NGS read alignment and genome coverage metrics**

	Sample		
	SRR10903401	SRR10903402	SRR10971381
Paired Reads, N (%)			
Total Number	476632 (100)	676694 (100)	28282964 (100)
Aligned	13913 (2.94)	54723 (8.18)	62,288 (0.22)
Concordantly Aligned	11469 (2.40)	44176 (6.52)	59261(0.21)
Discordantly Aligned	2444 (0.53)	10547 (1.67)	3027 (0.01)
Single Mates, N (%)			
Aligned	244 (0.03)	1308 (0.11)	294(0.001)
Overall Alignment Rate (%)	2.94	8.18	0.22
Quality score >Q30 (%)	92.7	92.1	88.2
Genome Coverage (%)	100.0	100.0	99.9
Average read depth (X)	133.5	522.2	598.2

**Table 2: Impact of Intra-host SNVs on viral genes**

Gene	Intra-host Variants Impact, N			Total, N (v/kbgl)*
	Low (synonymous)	Moderate (missense)	High (stop gained)	
ORF1ab	19	53	2	74 (3.47)
S	6	9	1	16 (4.18)
ORF3a	0	1	0	1 (1.20)
E	0	0	0	0 (0)
M	0	0	0	0 (0)
ORF6	2	1	0	3 (16.21)
ORF7a	0	1	0	1 (2.73)
ORF8	0	3	0	3 (8.21)
N	2	4	0	6 (4.76)
ORF10	0	0	0	0 (0)
Total, N	29	72	3	

\* normalised variants per 1 kb gene length (variants / gene-length \*1000)

**Table 3: Alignment characteristics of de novo assembled contigs**

Contig Name	Contig Length	Reference*		Contig Coordinates		Alignment Identity (%)	Alignment Type	Average Read Depth (x)	QC Pass#
		start	end	start	end				
<b>SRR10903401</b> (99.7% coverage)									
Contig 1	23994	75	24068	23994	1	99.99	Correct	57.01	+
Contig 2	5681	24246	29891	1	5646	99.96	Correct	71.40	+
Contig 3	331	23992	24322	331	1	100	Correct	164.39	+
Contig 4	179	24221	24399	179	1	100	Correct	97.56	+
Contig 5	192	17816	17909	94	1	100	Inversion	7.22	+
		17933	18030	95	192	100	Correct		
Contig 6	181	18052	18152	101	1	100	Relocation, Inconsistency	8.12	+
		17766	17845	102	181	100	Misassembly		
Contig 7	169	1707	1765	62	4	100	Inversion	7.62	+
		1815	1903	63	151	97.75	Correct		
Contig 8	165	23992	24087	96	1	100	Inversion	18.04	+
		23963	24031	97	165	100	misassembly		
<b>SRR10903402</b> (99.5% coverage)									
Contig 1	29842	133	29891	29842	84	99.98	Correct	234.32	+
Contig 2	242	2075	2139	178	242	100	Partial	1.09	-
Contig 3	242	21577	21629	242	190	100	Partial	1.06	-
Contig 4	173	23992	24090	102	4	100	Inversion	39.30	+
		23963	24033	103	173	100	Misassembly		
<b>SRR10971381</b> (100.0% coverage)									
Contig 1	29902	1	29891	29897	7	99.98	Correct	267.59	+
Contig 2	241	516	559	163	120	100	Inversion	1.00	-
		472	501	119	90	100	Misassembly		

\* Corresponding to reference MN975262 coordinates

# contig supported by at least 5 non duplicated reads of mapping quality >40