

1 **SARS-CoV-2 and ORF3a: Non-Synonymous Mutations and Polyproline Regions**

2 Elio Issa,^a Georgi Merhi,^a Balig Panossian,^a Tamara Salloum,^a Sima Tokajian^{a,#}

3

4 ^aDepartment of Natural Sciences, School of Arts and Sciences, Lebanese American University,

5 Byblos, 36, Lebanon

6

7 **Running Head:** SARS-CoV-2 Mutations and Viral Spread

8

9 #Address correspondence to Sima T. Tokajian, PhD, Department of Natural Sciences, School of

10 Arts and Sciences, Lebanese American University, Byblos, 36, Lebanon (stokjian@lau.edu.lb).

11 E.I, G.M and B.P contributed equally to this work.

12 **Abstract**

13 The effect of the rapid accumulation of non-synonymous mutations on the pathogenesis
14 of SARS-CoV-2 is not yet known. To predict the impact of non-synonymous mutations and
15 polyproline regions identified in ORF3a on the formation of B-cell epitopes and their role in
16 evading the immune response, nucleotide and protein sequences of 537 available SARS-CoV-2
17 genomes were analyzed for the presence of non-synonymous mutations and polyproline regions.
18 Mutations were correlated with changes in epitope formation. A total of 19 different non-
19 synonymous amino acids substitutions were detected in ORF3a among 537 SARS-CoV-2 strains.
20 G251V was the most common and identified in 9.9% (n=53) of the strains and was predicted to
21 lead to the loss of a B-cell like epitope in ORF3a. Polyproline regions were detected in two
22 strains (EPI_ISL_410486, France and EPI_ISL_407079, Finland) and affected epitopes
23 formation. The accumulation of non-synonymous mutations and detected polyproline regions in
24 ORF3a of SARS-CoV-2 could be driving the evasion of the host immune response thus favoring
25 viral spread. Rapid mutations accumulating in ORF3a should be closely monitored throughout
26 the COVID-19 pandemic.

28 **Importance**

29 At the surge of the COVID-19 pandemic and after three months of the identification of
30 SARS-CoV-2 as the disease-causing pathogen, nucleic acid changes due to host-pathogen
31 interactions are insightful into the evolution of this virus. In this paper, we have identified a set
32 of non-synonymous mutations in ORF3a and predicted their impact on B-cell like epitope
33 formation. The accumulation of non-synonymous mutations in ORF3a could be driving protein
34 changes that mediate immune evasion and favoring viral spread.

35 **Introduction**

36 The rapid spread of the coronavirus disease 2019 (COVID-19) caused by a novel
37 coronavirus, named SARS-CoV-2 due to its symptoms similarity to those induced by the severe
38 acute respiratory syndrome (SARS), is a major global concern (1). The epidemic started in late
39 December 2019 in Wuhan, the capital of Central China's Hubei Province and since then
40 thousands of cases have been reported in more than 46 countries

41 (<https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>).

42 Coronaviruses are enveloped non-segmented positive sense RNA viruses belonging to the family
43 Coronaviridae and the order Nidovirales and are broadly distributed in humans and other
44 mammals. The genome of SARS-CoV-2 showed 96.2% sequence similarity to a bat SARS-
45 related coronavirus (SARSr-CoV; RaTG13) collected in Yunnan province, China (1) and 79%
46 and 50% similarities to SARS-CoV and MERS-CoV, respectively (2). A transmission from wild-
47 life animals (such as pangolins) to humans has been recently suggested (3).

48 With the immediate and continuous release of sequence data, monitoring the rapid evolution of
49 the SARS-CoV-2 genome provides a strong lead towards predicting and potentially mitigating its
50 global spread. ORF3a protein (Accession # YP_009724391.1) is a hypothetical protein showing
51 a 72% sequence similarity to SARS3a protein in SARS-CoV. Here, we investigated the presence
52 of diverse non-synonymous mutations in ORF3a and their effects on the predicted protein
53 structure and its potential implication in the formation of epitopes. Moreover, polyproline
54 regions (PPRs) were detected in two strains. We used this approach to follow and understand the
55 impact of new emerging mutations in the pathogenesis and immune evasion of SARS-CoV-2.

56 **Results**

57 **Micro-clonality within ORF3a**

58 The clonal diversity of SARS-CoV-2 core genomes was highly similar in tree topology to
59 the gene tree of ORF3a (Figure 1). Signature mutations within SARS-CoV-2 genomes cluster
60 them into defined phylogenetic clades. Similarly, we observed micro-clonality within the ORF3a
61 gene tree defined by highlighted non-synonymous mutations G251V (green) and Q57H (pink)
62 that are found in conserved phylogenetic micro-clades representing sub-populations of mutants.

63

64 **Non-synonymous Mutations in ORF3a**

65 ORF3a, encoding a hypothetical protein, showed a 97.82% sequence similarity (100%
66 coverage) to a nonstructural protein NS3 of Bat coronavirus RaTG13 (Accession #
67 QHR63301.1). Moreover, ORF3a has a pro-apoptosis inducing APA3_viroporin conserved
68 domain, also found in SARS-CoV.

69 Sequence alignment of 537 ORF3a protein sequences revealed a total of 19 non-synonymous
70 amino acids substitutions, of which 52.6% (n=10) had a predicted deleterious functional outcome
71 and 47.4% (n=9) had a neutral functional outcome (Figure 2.A).

72 G251V was the most frequently detected substitution found in 9.9% (n=53) of the strains
73 followed by Q57H found in 3.9% (n=21) of the strains. Both G251V and Q57H were predicted
74 to be deleterious (Table 1).

75

76 **G251V linked to an Epitope Loss**

77 The G251V mutations were further investigated. Motif scanning demonstrated that G251V
78 resulted in the loss of a phosphatidylinositol-specific phospholipase X-box domain

79 (PIPLC_X_DOMAIN; 203-275 aa). The G251V substitution created serine protease cleavage
80 site. IEDB analysis revealed the presence of six putative epitopes in the non-mutant ORF3a
81 compared to five epitopes in the mutant ORF3a (Figure 2.B). The G251V substitution in ORF3a
82 was linked to the loss of a putative epitope the impact of which on viral spread and pathogenesis
83 requires further experimental studies. Other T176I and G254R substitutions resulted in a
84 decreased intensity of epitopes number two (blue) and epitope number five (yellow) (Figure
85 2.D).

86

87 **Detection of PPRs**

88 Notably, we detected PPRs in two SARS-CoV-2 genomes (EPI_ISL_407079, Finland and
89 EPI_ISL_410486, France). PPRs resulted in the joining of epitopes number four (purple) and
90 five (yellow) into one larger epitope (red) of 22 amino acids in size (start:235; end:256; sequence:
91 KIPPPPPPPPLHTIDGSSGVV) in EPI_ISL_410486 (France) and led the appearance of a new
92 epitope 23 amino acids in size (start 135; end 157; sequence: SKNPPPPPPPPPPPPPPHYC) in
93 EPI_ISL_407079 (Finland) (Figure 2.C). Blastn search of a 23 bp DNA stretch from non-mutant
94 strains showed a 100% identity to RaTG13 (Accession # MN996532.1).

95

96 **Discussion**

97 These combined results suggest that the non-synonymous G251V mutation introduced into
98 ORF3a protein in SARS-CoV-2 could be linked to immune evasion and thus viral spread and
99 pathogenesis. ORF3a is a transmembrane protein that localizes to the plasma membrane
100 especially in the ER-Golgi region and activates the PKR-like ER kinase (PERK) signaling
101 pathway which protects viral proteins against ER-associated degradation. The activation of this

102 pathway leads to apoptosis(11). A pro-apoptosis inducing APA3_viroporin conserved domain
103 detected in ORF3a of SARS-CoV-2 is also found in SARS-CoV 3A protein (11).

104 The G251V was detected in ORF3a in 9.9% of the strains (n=53). G251V led to the loss a B cell-
105 like epitope and a PIPLC_X_DOMAIN the eukaryotic homologue of which is involved in signal
106 transduction processes (8). The accumulation of non-synonymous mutations could be driven by
107 the humoral immunity as reported previously in the mucin-like domain of the Ebola virus
108 glycoprotein (12).

109 Of paramount importance is the emergence of PPRs in ORF3a detected in two of the SARS-
110 CoV-2 sequenced genomes (in EPI_ISL_410486, France and EPI_ISL_407079, Finland). PPRs
111 are an open field for recombination that viruses use to adapt based on selective pressure (13).
112 PPRs were previously shown to be indispensable for the activity of the Coxsackievirus B 3A
113 protein which blocks ER-to-Golgi transport affecting protein synthesis (14). Studies on Hepatitis
114 E virus also highlighted the role of PPRs in host-range adaptation and viral replication (15).

115 In conclusion, our study reveals and for the first time a common non-synonymous G251V
116 substitution and PPRs in ORF3a which could be respectively linked to the loss of a putative
117 epitope and viral spread and pathogenesis.

118 **Materials and Methods**

119 **Pan-genome analysis**

120 A total of 537 SARS-CoV-2 complete genomes with high quality sequencing downloaded from
121 GISAID were utilized for genome and ORF3a alignments.

122 All sequences were uniformly annotated using Prokka v 1.1.3 (4). The annotated Genbank files
123 were edited to have more concise locus tag identifiers. The Genbank annotations of the genomes
124 were used as input in the PanX (5) pipeline for pan genome analysis. A core genome threshold of
125 0.99, MCL inflation parameter of 1.5, and a modified core diversity cutoff for branch lengths
126 above 0.001 were used alongside the default parameters.

127 **Protein Structure prediction**

128 Sequences were aligned using MUSCLE v3.8.31 (6). PROVEAN was used to predict the
129 functional effects of amino acid substitutions (7). ExPASy and PROSPER were used for motif
130 scanning and protease site prediction, respectively (8, 9). The Immune epitope database analysis
131 resource (IEDB-AR) was used for epitopes prediction using a 0.5 threshold and default settings
132 (10).

133 **Acknowledgements**

134 We thankfully acknowledge the authors, generating and submitting laboratories of the sequences
135 from GISAID's EpiCoV™ database. We also acknowledge the authors of all Coronaviridae
136 genome sequences deposited in GenBank. This study does not claim ownership of these
137 sequences, which were used within the analysis workflow to further our understanding of the on-
138 going pandemic of SARS-CoV-2 and the underlying molecular changes that govern the virus'
139 transmission and infectivity patterns. The authors wish to declare that they do not have any
140 conflict of interests.

141 **Author Contributions:** *Concept and design:* S.T. *Acquisition, analysis, or interpretation of*
142 *data:* All authors. *Drafting of the manuscript:* All authors. *Critical revision of the manuscript for*
143 *important intellectual content:* Tokajian. *Administrative, technical, or material support:* E.I, B.P,
144 G.M. *Supervision:* S.T.

145 **Conflicts of interest:** The authors wish to declare that they do not have any conflict of interests.

146 **Funding/support:** This work was partially financed by the School of Arts and Sciences
147 Research and Development Council at the Lebanese American University.

148 **References**

- 149 1. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-
150 L, Chen H-D, Chen J, Luo Y, Guo H, Jiang R-D, Liu M-Q, Chen Y, Shen X-R, Wang X,
151 Zheng X-S, Zhao K, Chen Q-J, Deng F, Liu L-L, Yan B, Zhan F-X, Wang Y-Y, Xiao G-F,
152 Shi Z-L. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat
153 origin. *Nature* 1–4.
- 154 2. Gralinski LE, Menachery VD. 2020. Return of the Coronavirus: 2019-nCoV. 2. *Viruses*
155 12:135.
- 156 3. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of
157 SARS-CoV-2. *Nat Med* 1–3.
- 158 4. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinforma Oxf Engl*
159 30:2068–2069.
- 160 5. Ding W, Baumdicker F, Neher RA. 2018. panX: pan-genome analysis and exploration.
161 *Nucleic Acids Res* 46:e5.
- 162 6. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
163 throughput. *Nucleic Acids Res* 32:1792–1797.
- 164 7. Choi Y, Chan AP. 2015. PROVEAN web server: a tool to predict the functional effect of
165 amino acid substitutions and indels. *Bioinformatics* 31:2745–2747.
- 166 8. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel
167 V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti

- 168 R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, Stockinger H. 2012.
169 ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res* 40:W597–W603.
- 170 9. Song J, Li F, Leier A, Marquez-Lago TT, Akutsu T, Haffari G, Chou K-C, Webb GI, Pike
171 RN. 2018. PROSPERous: high-throughput prediction of substrate cleavage sites for 90
172 proteases with improved accuracy. *Bioinformatics* 34:684–687.
- 173 10. Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, Bourne PE, Bui H-H, Buus S,
174 Frankild S, Greenbaum J, Lund O, Lundegaard C, Nielsen M, Ponomarenko J, Sette A, Zhu
175 Z, Peters B. 2008. Immune epitope database analysis resource (IEDB-AR). *Nucleic Acids*
176 *Res* 36:W513-518.
- 177 11. Minakshi R, Padhan K, Rani M, Khan N, Ahmad F, Jameel S. 2009. The SARS
178 Coronavirus 3a protein causes endoplasmic reticulum stress and induces ligand-
179 independent downregulation of the type 1 interferon receptor. *PloS One* 4:e8342.
- 180 12. Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG, Sealfon RS, Ladner JT,
181 Kugelman JR, Matranga CB, Winnicki SM, Qu J, Gire SK, Gladden-Young A, Jalloh S,
182 Nosamiefan D, Yozwiak NL, Moses LM, Jiang P-P, Lin AE, Schaffner SF, Bird B, Towner
183 J, Mamoh M, Gbakie M, Kanneh L, Kargbo D, Massally JLB, Kamara FK, Konuwa E,
184 Sellu J, Jalloh AA, Mustapha I, Foday M, Yillah M, Erickson BR, Sealy T, Blau D,
185 Paddock C, Brault A, Amman B, Basile J, Bearden S, Belser J, Bergeron E, Campbell S,
186 Chakrabarti A, Dodd K, Flint M, Gibbons A, Goodman C, Klens J, McMullan L, Morgan
187 L, Russell B, Salzer J, Sanchez A, Wang D, Jungreis I, Tomkins-Tinch C, Kislyuk A, Lin
188 MF, Chapman S, MacInnis B, Matthews A, Bochicchio J, Hensley LE, Kuhn JH, Nusbaum
189 C, Schieffelin JS, Birren BW, Forget M, Nichol ST, Palacios GF, Ndiaye D, Happi C,

- 190 Gevao SM, Vandi MA, Kargbo B, Holmes EC, Bedford T, Gnirke A, Ströher U, Rambaut
191 A, Garry RF, Sabeti PC. 2015. Ebola Virus Epidemiology, Transmission, and Evolution
192 during Seven Months in Sierra Leone. *Cell* 161:1516–1526.
- 193 13. Lhomme S, Abravanel F, Dubois M, Sandres-Saune K, Mansuy J-M, Rostaing L, Kamar N,
194 Izopet J. 2014. Characterization of the polyproline region of the hepatitis E virus in
195 immunocompromised patients. *J Virol* 88:12017–12025.
- 196 14. Wessels E, Duijsings D, Notebaart RA, Melchers WJG, Kuppeveld FJM van. 2005. A
197 Proline-Rich Region in the Coxsackievirus 3A Protein Is Required for the Protein To
198 Inhibit Endoplasmic Reticulum-to-Golgi Transport. *J Virol* 79:5163–5173.
- 199 15. Purdy MA, Lara J, Khudyakov YE. 2012. The Hepatitis E Virus Polyproline Region Is
200 Involved in Viral Adaptation. *PLOS ONE* 7:e35974.

201

202 **Tables**

203 **Table 1. List of 19 non-synonymous amino acids substitutions in ORF3a among 537 strains.**

204 The G251V substitution is shown in bold.

205

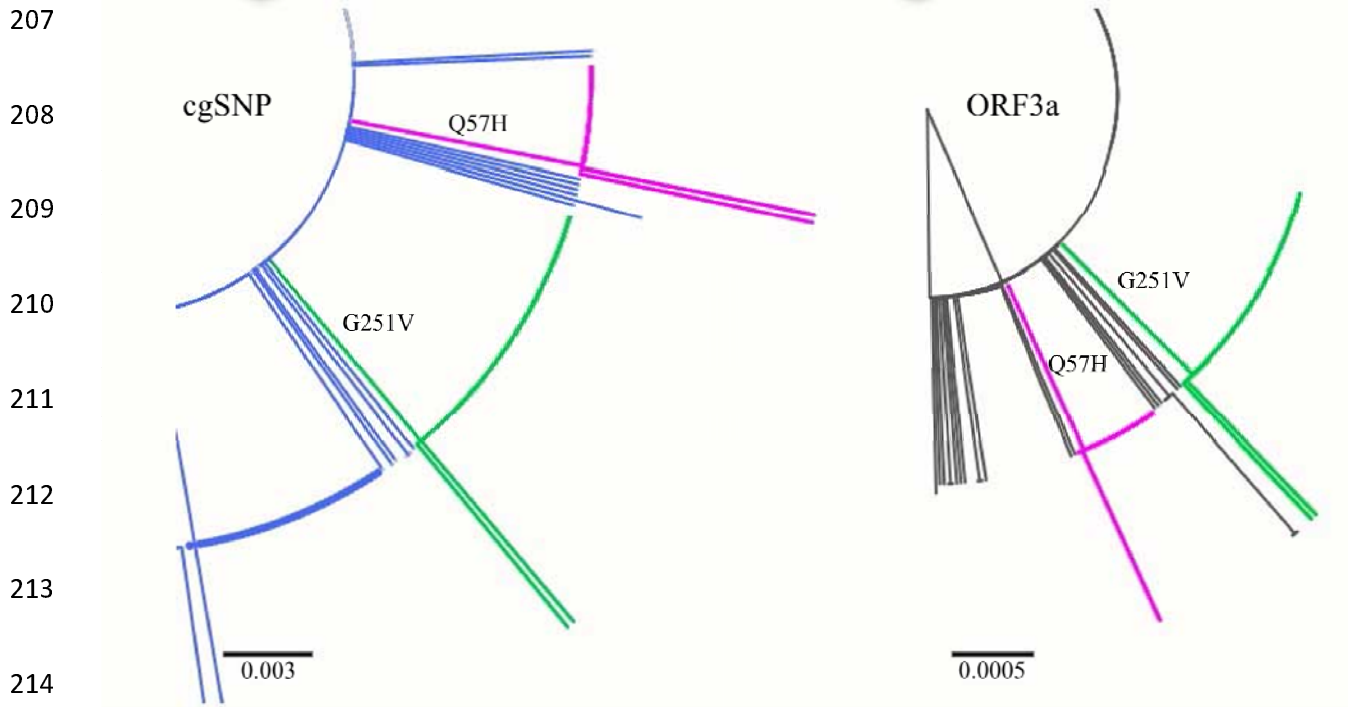
Amino acids substitutions in ORF3a	Incidence^a	Variation Effect on Protein(7)
F8L	0.02% (n =1)	Deleterious
A54V	0.04% (n =2)	Neutral
Q57H	3.90% (n =21)	Deleterious
K61N	0.02% (n =1)	Deleterious
G76S	0.02% (n =1)	Neutral
V88L	0.02% (n =1)	Neutral
W128L	0.02% (n =1)	Deleterious
L140V	0.04% (n =2)	Neutral
D155Y	0.02% (n =1)	Deleterious
T176I	0.02% (n =1)	Deleterious
E191G	0.02% (n =1)	Deleterious
G196V	0.07% (n =4)	Deleterious
H227R	0.02% (n =1)	Neutral
E239V ^b	0.02% (n =1)	Neutral
D250V ^b	0.02% (n =1)	Neutral
G251V	9.90% (n =53)	Deleterious
G254R	0.02% (n =1)	Deleterious
V259L	0.02% (n =1)	Neutral
T269M	0.02% (n =1)	Neutral

^a Percentage values in this column do not add to 100% as mutations only cover a fraction of the total sample size;

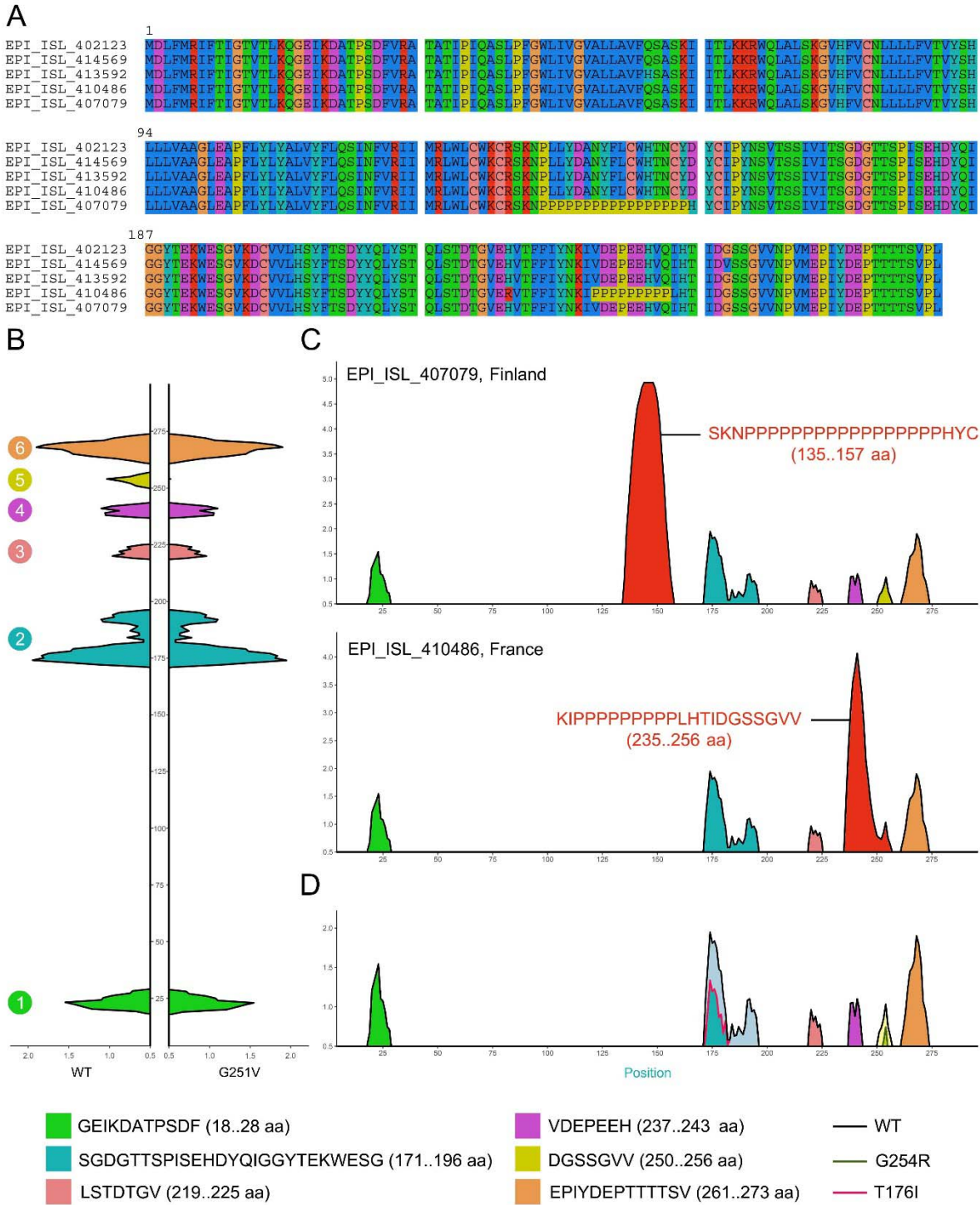
Total number of sequences= 537.

^b Both mutations were detected in the same isolate ESL_ISL_406592

206 **Figures**



215 **Figure 1: Phylogenetic trees of SARS-CoV-2 core genomes and ORF3a.** Magnified
216 maximum-likelihood phylogenetic trees of (A) SARS-CoV-2 genomes based on core genome
217 SNP differences in all concatenated ORFs and (B) ORF3a gene tree highlighting G251V mutant
218 clade in green and Q57H mutant clade in pink.



219 **Figure 2: Mutations analysis of ORF3a.** (A) Multiple sequence alignment between ORF3a
 220 protein of G251V (EPI_ISL_414569, Hong Kong), G254R (EPI_ISL_415627, USA), T176I
 221 (EPI_ISL_411950, Jiangsu), PPR-containing proteins (EPI_ISL_410486, France and

222 EPI_ISL_407079, Finland) mutants compared to non-mutant (EPI_ISL_402123, Wuhan) **(B)** B-
223 cell like epitopes of the non-mutated ORF3a protein (left) and G251V mutant (right). Only
224 values above the threshold (0.5) are included. The mutation lead to the loss of one B cell epitope.
225 **(C)** B-cell like epitopes of PPR-containing isolates. Additional epitopes are indicated in red. **(D)**
226 B-cell like epitopes of T176I and G254R mutants with decreased intensity as compared to non-
227 mutant.