

1 **Determining international spread of novel B.1.1.523 SARS-CoV-2 lineage**

2

3 **Lukas Zemaitis^{1, *+}, Gediminas Alzbutas²⁺, Dovydas Gecys¹, Andrey Komissarov³, Arnoldas**
4 **Pautienius⁴ Rasa Ugenskiene⁵ Marius Sukys⁵ and Vaiva Lesauskaite¹**

5 ¹ Lithuanian University of Health Sciences, Institute of Cardiology, Laboratory of Molecular
6 Cardiology, Kaunas, LT-50162, Lithuania

7 ² Lithuanian University of Health Sciences, Institute for Digestive Research, Laboratory of
8 Translational Bioinformatics, Kaunas LT-50162, Lithuania

9 ³ Smorodintsev Research Institute of Influenza, Saint Petersburg, 197376, Russian Federation

10 ⁴ Lithuanian University of Health Sciences, Institute of Microbiology and Virology, Kaunas LT-
11 47181, Lithuania

12 ⁵ Lithuanian University of Health Sciences Hospital Kaunas Clinics, Genetics and Molecular
13 Medicine Clinic, Kaunas LT- 50161, Lithuania

14 * lukas.zemaitis@lsmuni.lt

15 + these authors contributed equally to this work

16

17 **ABSTRACT**

18 *Here we report the emergence of variant lineage B.1.1.523 that contains a set of mutations including*
19 *156_158del, E484K and S494P in Spike protein. E484K and S494P are known to significantly reduce*
20 *SARS-CoV-2 neutralization by convalescent and vaccinee sera and are considered as mutations of*
21 *concern. Lineage B.1.1.523 has presumably originated in Russian Federation and spread across*
22 *European countries with the peak of transmission in April – May 2021. The B.1.1.523 lineage has*
23 *now been reported from 27 countries.*

24 **INTRODUCTION**

25 The emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in late
26 2019 led to the ongoing Coronavirus Disease 2019 (COVID -19), now a global pandemic with more
27 than 230 million cases of infection and about 5 million deaths worldwide ¹. On January 5, 2020, the
28 first whole genome sequence of 2019-nCoV was completed by Wuhan Institute of Virology, China
29 Centre for Disease Control and Shanghai Public Health Clinical Centre of Fudan University ². From
30 this point on, genome sequencing has played an important role in vaccine development,
31 understanding viral evolution and epidemiological characteristics. In many countries, SARS-CoV-2
32 sequencing has been implemented at the national level as a tool for epidemiological management
33 ³.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

34 A large dataset of SARS-CoV-2 genomes has been collected in the GISAID database, which now
35 contains more than 3.7 million sequenced genomes from around the world ⁴. As of May 31, 2021,
36 World Health Organisation (WHO) has proposed designations for global SARS-CoV-2 variants of
37 concern (VOC) and variants of interest (VOI) to be used alongside scientific nomenclature in
38 communications about variants to the public ⁵. This list includes the variants on the global list of
39 WHO VOC and VOI and will be updated as the list of WHO changes. There are currently three SARS-
40 CoV-2 VOCs: Beta, Gama and Delta. The variant previously classified as Alpha (B.1.1.7) has been
41 reclassified as de-escalated due to the drastic decrease in prevalence in the EU/EEA ⁶.

42 Through the routine analysis of National Lithuanian sequencing results from national
43 sequencing efforts coordinated by National Public Health Surveillance Laboratory, we have
44 identified a novel SARS-CoV-2 variant classified as B.1 by PANGO but containing multiple S protein
45 mutations associated with effects on immunity ([https://github.com/cov-lineages/pango-
46 designation/issues/69](https://github.com/cov-lineages/pango-designation/issues/69)), such as E484K; 156_158del; S494P. Preliminary phylogenetic analysis
47 indicated that this variant has a distinct viral lineage that may have originated in Russia. We reported
48 this variant to the PANGO curators and gave it the new phylum name B.1.1.523
49 (<https://github.com/cov-lineages/pango-designation/issues/69>). On 14-July-2021, WHO added this
50 variant to the list of variants under Monitoring section. By using bioinformatics tools, we performed
51 a detailed analysis of this lineage and disclosed our findings about this variant or the mutational
52 subgroups typical of this variant.

53 With this report we aim to share a detailed analysis of discovered variant, evaluate its origin
54 as well as predict potential epidemiological impact and risks.

55

56 **RESULTS**

57

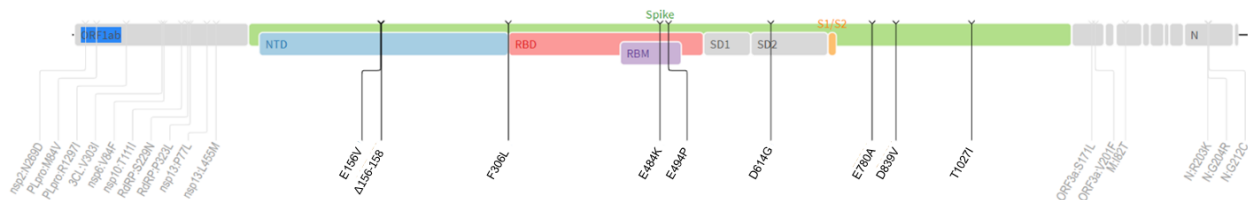
58 ***Mutation review of B.1.1.523***

59

60 Several mutations in S region have been observed in B.1.1.523 variant, from which
61 156_158del, E484K, and S494P are considered as an attribute for VOCs (Fig. 1). According to
62 previously reported data, the 156_157del and G158R mutations in the Delta variant are matching to
63 the same surface as the 144 and 241–243 deletions in the Alpha and Beta (B.1.351) variants,
64 respectively. These altered residues are found in the NTD ‘supersite’ that is targeted by most anti-
65 NTD neutralizing antibodies, thus providing a mode to dodge immune system ⁷. Moreover, E484K
66 mutation also contributes to SARS-CoV-2 immune system evasion. Several recent studies have
67 observed that E484K may significantly reduce convalescent serum neutralization ^{8,9}. Additionally, it
68 was observed that S494P mutation is related to 3-5-fold reduced SARS-CoV-2 neutralization in sera),
69 however, this mutation was not as potent at neutralization as E484K ^{8,10}. With a combination of
70 156_158del, E484K, and S494P mutations, B1.1.523 lineage should remain on epidemiologists
71 watchlist as one of the most concerning SARS-CoV-2 lineages.

72 In addition to 156_158del, E484K, and S494P more than 70% of genomes attributed as
73 B.1.1.523 lineage possess F306L, D839V and T1027I in Spike and a set of substitutions in ORF1a
74 (NSP3:M84V, R1297I; NSP2: N269D; NSP5: V303I; NSP6: V84F; NSP10:T111I), ORF1b (NSP12:S229N,

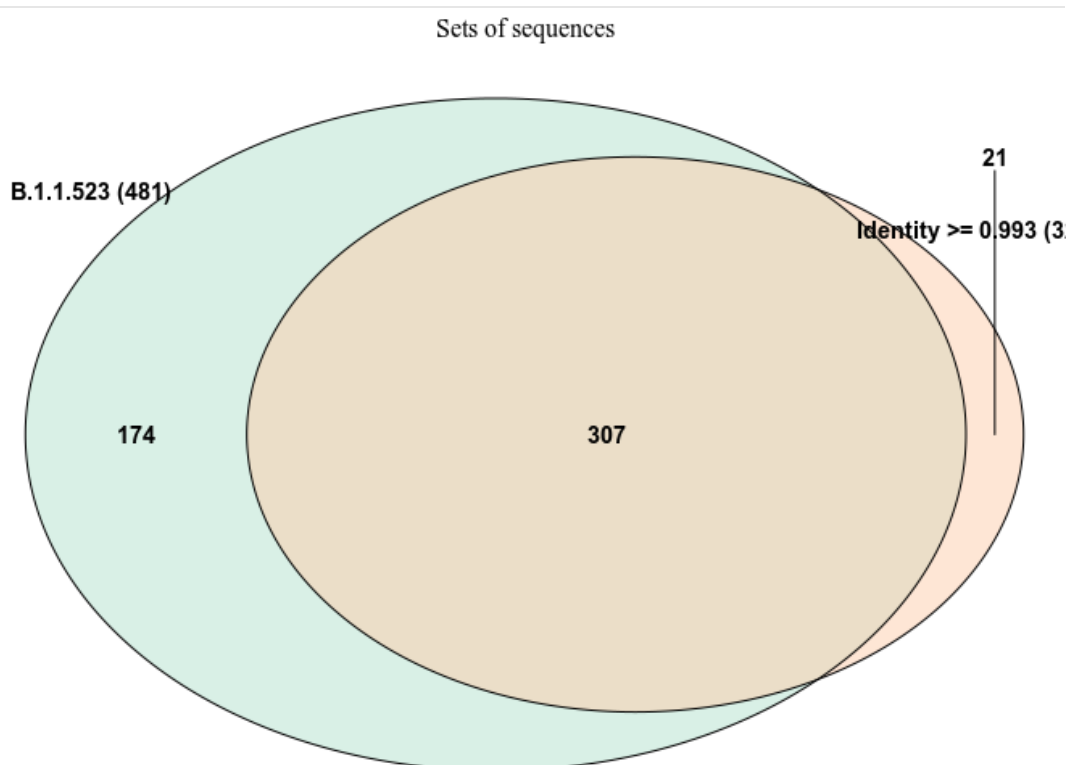
75 P323L; NSP13: P77L), ORF3a (NS3: S171L, V201F), M:I82T, N:G212C (See Fig. X) . T1027I substitution
 76 in Spike is common for VOC Gamma (P.1). Little or no information is available on these mutations.
 77
 78
 79



80
 81
 82 **Figure 1. Mutation overview in B.1.1.523 lineage.** Several other mutations have been observed in
 83 the spike-protein sequence of B1.1.523 variant, including E156V, F306L, D614G, E780A, D839V and
 84 T1027I.

85
 86 ***Origin and formation of key S protein formation of the lineage***

87
 88 One of the objectives of the analysis was to determine if the two clusters of mutations,
 89 responsible for immunity resistance, were acquired by sequential mutations or if this is a result of
 90 recombination events. Initially a list of data entries comprised of the sequences classifiable as
 91 B.1.1.523 lineage together with the closest sequences based on identity was used to construct a
 92 maximum likelihood (ML) tree. The size and overlap between the two data sets is depicted in Fig. 2.
 93



94
 95

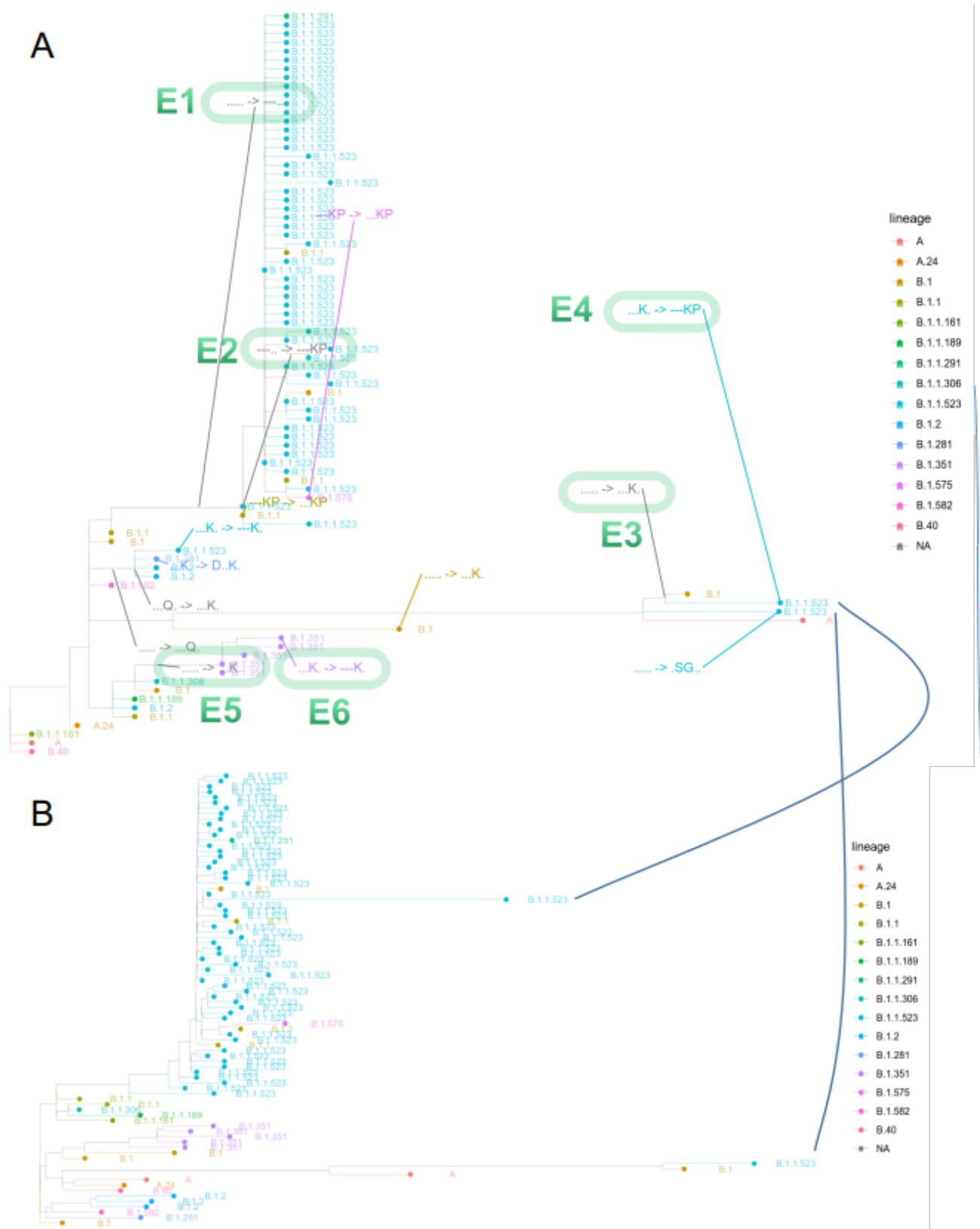
96 **Figure 2. The overlap between the two data sets used for focused ML tree.** The sequences were
97 either chosen based on Pango assignment or by the identity with a B.1.1.523 lineage sequence
98 (EPI_ISL_1590462). Most of the sequences, with high identity (> 0.993) to the Latvian B.1.1.523
99 lineage, were classified as belonging to the B.1.1.523. However, 21 (6%) sequences were not
100 assigned to the B.1.1.523.

101

102

103 Generation of ML revealed several interesting properties of B.1.1.523 (Suppl. Fig. 1). At the
104 base of the lineages leading to the B.1.1.523 sequences having a full set of expected S protein
105 mutations branches away in clusters of sequences having the triple S:156_158del deletion. The
106 sequences having the additional substitutions at S:484 and S:494 positions emerge further in the
107 evolution. However, here, we have no clear indications of that the mutations S:E484K, S:S494P are
108 introduced sequentially to form the B.1.1.523 lineage.

109 We have observed that some sequences which originated from progenitors with full set of
110 expected substitutions at 484, 494, 156, 157, 158 S region positions, have undergone a reverse-type
111 mutation to wild type variants. Such events are highly unlikely and usually are caused by erroneous
112 sequence assembly or could be an indication of low-quality data. Additionally, in order to discern
113 the cases where the mutations comprising the two regions of potential enhanced resistance to
114 immune response have been combined phylogeny analysis of all S protein unique variants was
115 performed. As it is depicted in Figure 3, there are at least three distinct cases where E484K and/or
116 S494P have been combined with the 156_158del. Phylogeny analysis of S protein indicated that
117 lineage assignments by Pango sometimes can be misleading. Furthermore, relying on plain
118 assignments could hide novel developments of SARS-CoV-2. For example, some sequences from
119 Turkey that were assigned as lineage B.1.1.523 originated from distinct clusters based on S sequence
120 (Fig. 4A) and were evidently evolutionary distant from other B.1.1.523 lineage genomes.



121
 122
 123
 124
 125
 126

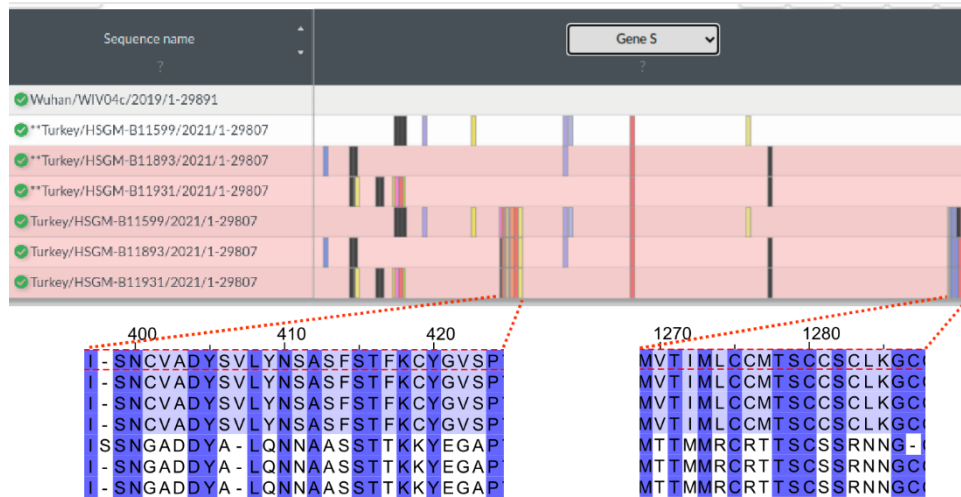
Figure 3. Phylogeny based on S protein sequence. (A) The tree represents a maximum likelihood tree based on all unique S protein sequences of the genomes deposited to the GISAID. The visible subset of the tree matches lineages that lead to branches having 156_158del and E484K or S494P

127 mutations. The arrow “->” indicates haplotype transitions detected by comparison of parental
128 sequences with their offspring variants. The five-letter haplotype strings match 156, 157, 158, 484,
129 494 positions of the S protein with “.” meaning the wild type. The green ovals and their number
130 indicate prominent transitions that are discussed in the main text. (B) The lower tree matches
131 maximum likelihood tree based on whole genomes of the cases visualised in the upper tree. The
132 black solid lines indicate a match of nodes for two B.1.1.523 sequences from Turkey in the
133 phylogenies based on S protein and whole genome.

134 Transitions E1 (introduction of the triple deletion) and E2 (introduction of E484, S494P) (Fig.
135 3) indicate a pathway where the majority of B.1.1.523 lineage sequences took. As in the case of data
136 given in (Fig. 3) we do not detect sequential acquisition of E484K or S494P, and immediately next to
137 the triple deletion we see the two aforementioned additional mutations. This could indicate a
138 potential recombination event, but it is most probably due to lack of insufficient data.

139 An interesting case is with two highly diverged Turkey sequences that are classified by Pango
140 as B.1.1.523 lineage. In this case following the most likely scenario at first the E484K was acquired
141 and only then the S494P and E156_R158del were acquired (Fig. 4). As in the case discussed before,
142 the sequential acquire steps are lacking from the data. Most probably the sequencing data was not
143 comprehensive enough to reveal the full picture on how the combination was formed. In addition,
144 these two highly diverged sequences indicate that there exist vast uncharted territories of COVID-
145 19 evolution as we get only sparse sequencing data from Central Asia regions where the COVID-19
146 infection are barely controlled. An evident third case emergent combination of immune response
147 hindering mutations from distinct S protein regions are highlighted by E5 and E6 at Figure 3A. In this
148 case within the B.1.351 lineage at first the E494K mutation was introduced and then E156_R158del
149 followed. The presented data evidently shows that the immune response hindering mutations from
150 RBD and NTD domains have been combined in one protein at least three times. In two of them
151 E484K occurred first, in one case the first one was the triple deletion. Most probably these are
152 results of independent mutational events.

153

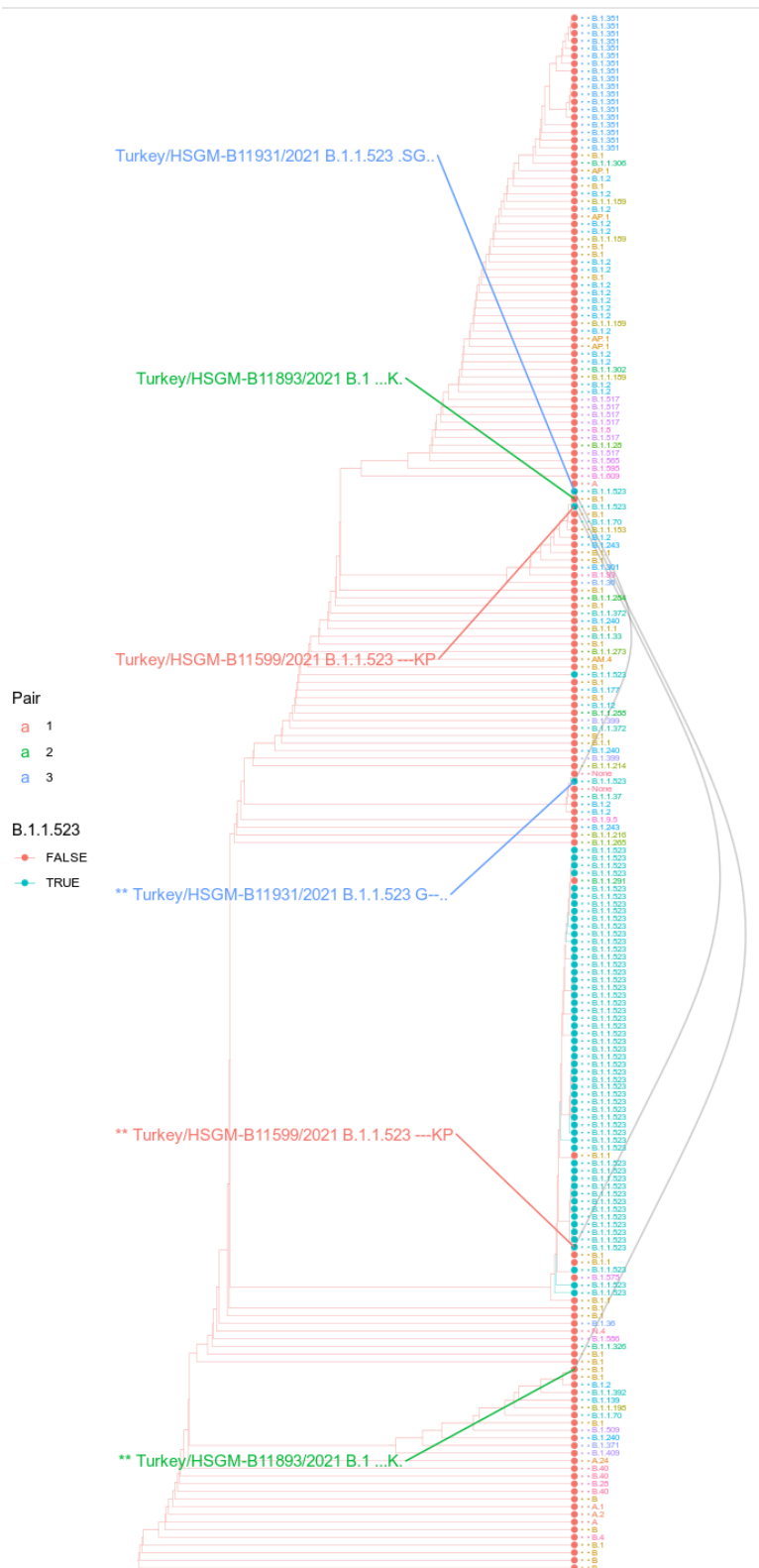


154

155

156 **Figure 4. S protein region of B.1.1.523 sequences from Turkey.** Turkey/HSGM-B11599/2021 and
 157 Turkey/HSGM-B11931/2021 were classified by Pangolin as belonging to B.1.1.523 lineage. The top
 158 figure is a snapshot from the Nextclade analysis. The top sequence is the reference sequence hCoV-
 159 19 that is used by GISAID. The asterisks ** indicate sequence from Turkey variants that have their
 160 two most variable sequences swapped with corresponding regions from the reference sequence. The
 161 two lower graphs indicate the sequence alignments from the two extremely variable fragments. The
 162 order of sequences is the same as in the upper graph: reference sequence, three Turkey sequences
 163 with swapped fragments, three original Turkey sequences.

164



165

166 **Figure 5. Phylogeny based on S protein sequence with modified sequences from Turkey.** The
167 cladogram of maximum likelihood tree that includes the three pairs of Turkey sequences with their
168 variable regions either swapped with counterparts of reference sequences or being left original.
169 Colour of the tips of leaves indicates if they are being classified as B.1.1.523. Tip labels indicates
170 their Pangolin assignment with colours indicating different lineages. The different tips connecting
171 grey lines indicates a Turkey sequence. The asterisks (“**”) in front of the sequence name labels

172 indicates the sequence variant where the variable regions were swapped with corresponding
173 regions from the reference sequences. Next to the sequence labels the haplotype at positions 156-
174 158,484,494 with “.” indicating the wild type and “-” - a gap. A Turkey sequences.

175 In Figure 5 the Turkey sequences are highly diverged from the wild type based on the region
176 variation of S protein. Based on the GISAID blast searches (performed at 2021 10 1 07) the first
177 variable region (bottom left, Fig. 5) from Turkey/HSGM-B11599/2021 was found in 7 sequences and
178 the second variable region (bottom right, Fig. 5) was found in 4 sequences. The sequences have
179 been deposited in the GISAID during several submissions. Other genomic regions of the Turkey
180 sequences showed in the Fig. 4 do not contain such large SNPs clusters as the ones showed in the S
181 protein. This indicates the changes in the placement of the three Turkey sequences upon switching
182 the two extremely deviated fragments with corresponding fragments with reference sequence.

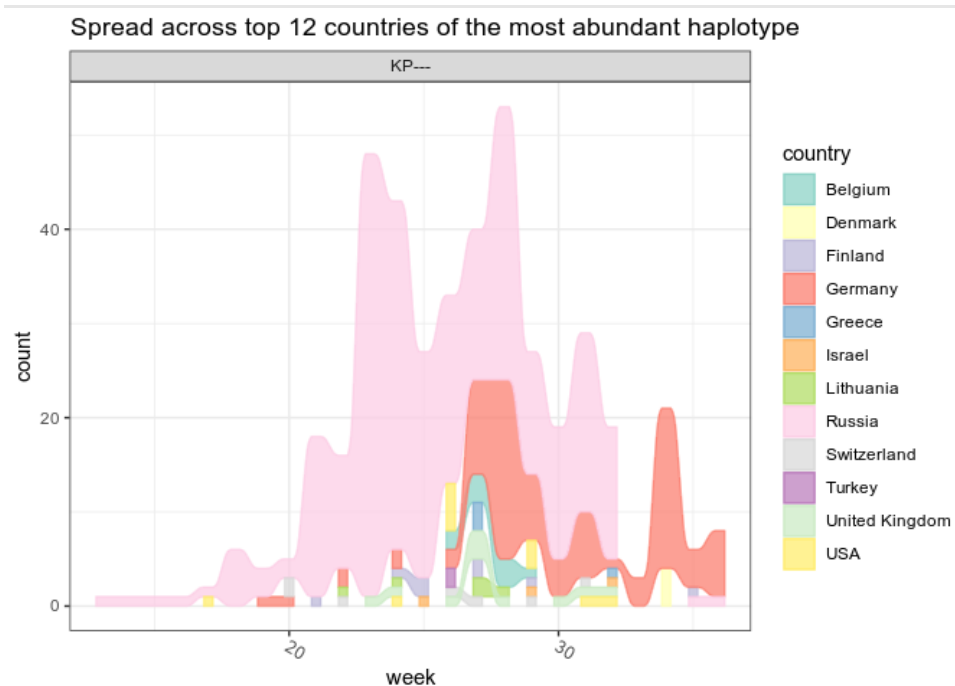
183

184 **B.1.1.523 spread worldwide**

185

186 Till August 31st, 2021, 459 B.1.1.523 sequences spread across 27 countries have been
187 published in the GISAID. According to analysis results, B1.1.523 has originated in Russian Federation
188 and spread across European countries (Fig 6). The sequenced clades peaked at week 25 and then
189 subsided. In total, 95 transmission clusters have been identified. The peak of B.1.1.523 transmission
190 intensity was around April - May 2021. The most numerous transmission clusters were detected for
191 with MRCA's originating from Germany and Russia.

192

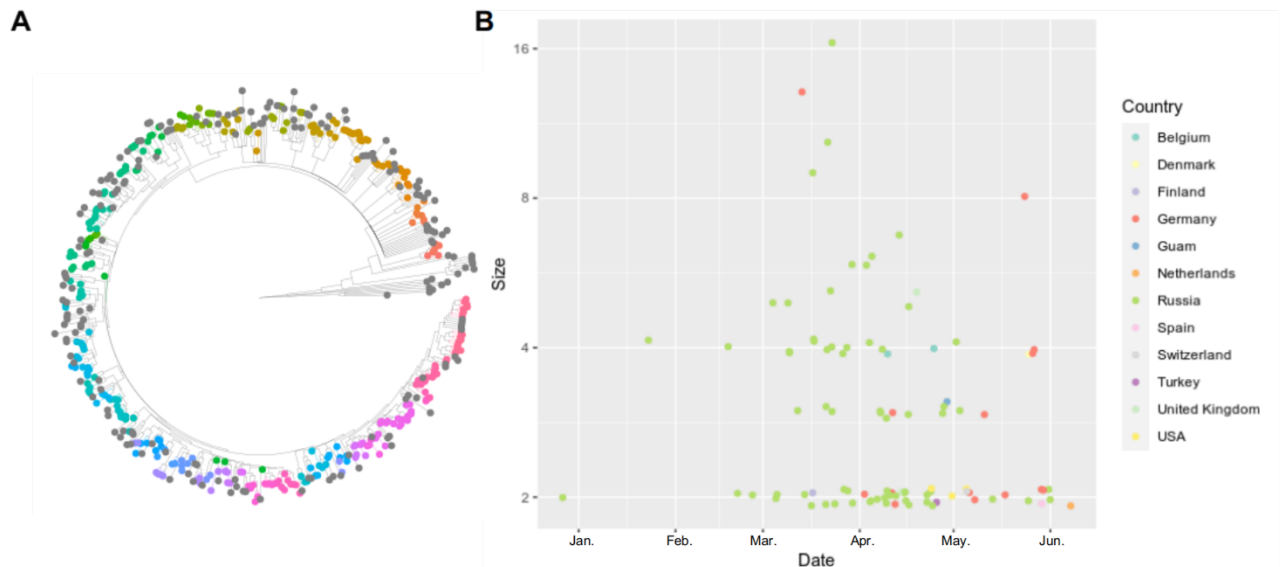


193

194 **Figure 6. The distribution of cases of the lineage B.1.1.523 across countries at different time**
195 **points.** The “0” time point indicates the date of the earliest lineage sequence uploaded on the
196 GISAID database. Only sequences that have the typical set of S mutations were considered (E484K,

197 S494P, 156_158del). Only the cases which are corresponding to top 12 countries with the most
198 abundant detection rate are included in the underlying data. The top 12 countries correspond to
199 the 93 % of all cases.

200
201



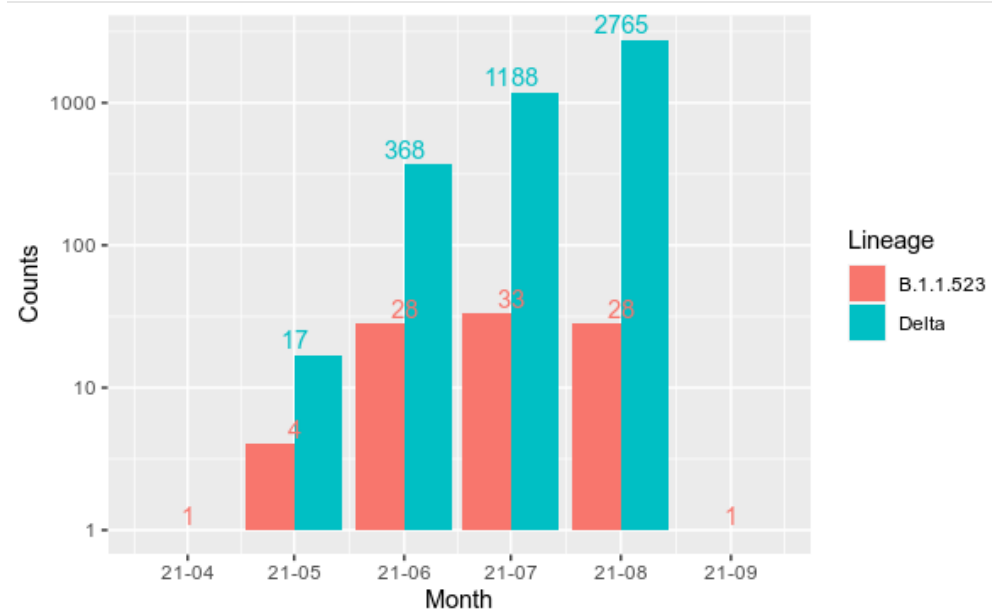
202
203

204 **Figure 7 B.1.1.523 Transmission clusters identified using the phylogeny tree.** Sequences with
205 identity larger or equal to 99.3 % to the Latvian B.1.1.523 sequence (EPI_ISL_1590462) were added
206 to the analysis. Colours indicate sequences belonging to the same transmission cluster. Grey colour
207 marks sequences that were not assigned to any transmission colour (A). Colour encodes the country
208 of the most recent common ancestor (MRCA) of all sequences that constitute a cluster. Y axis
209 represents size of a cluster and Y axis denotes the date inferred for the MRCA sequence (B).

210

211 Most of the cases transmission origin country (country inferred for a cluster MRCA sequence)
212 and the target country (country of a sequence indicated in the GISAID metadata) was Russia and as
213 we see in the Figure 6B majority of the transmission events happened within Russia. Evidently The
214 largest number of cases were transmission origin and target countries were different resembles
215 transmission from Russia to Germany. The data indicate one reverse type (Germany to Russia)
216 transmission. As of the date of writing this article, Germany can be considered as a reservoir of the
217 B.1.1.523 lineage. As for May 2021 the B.1.1.523 and Delta variants represented 0.002% and 0.054%
218 respectively, while the data of August 2021 shows a steady increase in detected cases to 0.327%
219 and 32.358% respectively. The overall number of sequenced cases in Germany is becoming
220 overwhelmed by Delta variant, however the lineage B.1.1.523 lineage seems not to be fading away
221 (Fig. 8). Data shows that B.1.1.523 is able to steadily spread without interference with the Delta
222 variants and can be considered as a predominant background lineage of SARS-CoV-2. As indicated
223 in the Figure 6. the spread of the lineage (based on GISAID submissions) in Russia diminished while
224 it remains evidently circulating in Germany. It is unclear whether B.1.1.523 is completely out of

225 circulation in Russia or its frequency dropped below the sensitivity threshold of SARS-CoV-2 genomic
226 surveillance in this country.



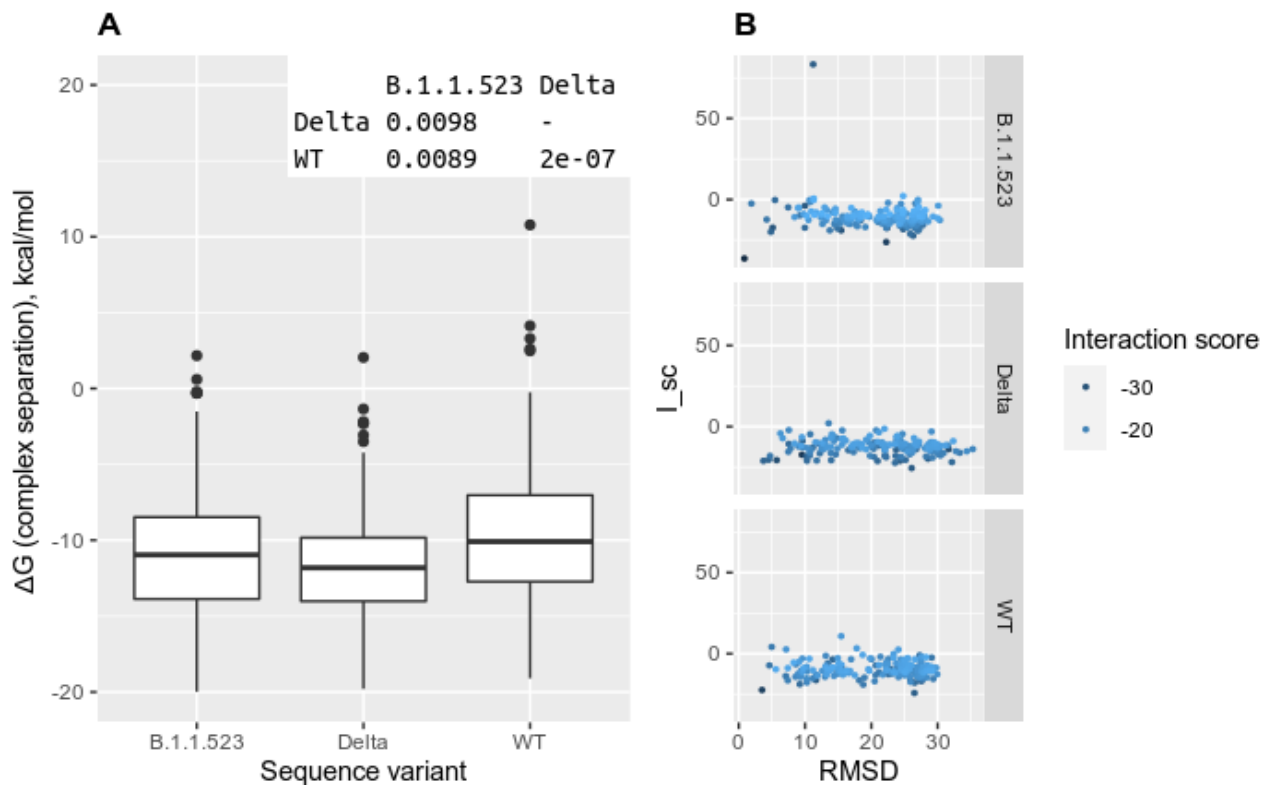
227

228 **Figure 8. Newly sequenced cases of B.1.1.523 and Delta lineages in Germany.** Data number of new
229 cases per month is based on GISAID metadata (August 31st, 2021). Numbers indicate the total
230 number of sequences deposited in GISAID.

231

232 ***B.1.1.523 antibody escape***

233 In Figure 9A the modelled free energy of complex separation is given as calculated for the
234 three sequence variants considering the NTD-antibody complex. As Rosetta documentation states¹¹
235 ΔG of complex separation ($dG_{\text{separated}}$) shows the change in Rosetta energy when the interface
236 forming chains are separated, versus when they are complexed - therefore the lower the value the
237 more energetically favourable is complex separation and less favourable is complex formation. In
238 other words - the higher the value - the less likely should be the neutralization by the antibody.
239 Pairwise comparisons using Wilcoxon rank sum test with Benjamini-Hochberg correction for
240 multiple comparisons indicated that all three cases significantly differ from each other (p values are
241 given in Fig. 9A upper right corner).

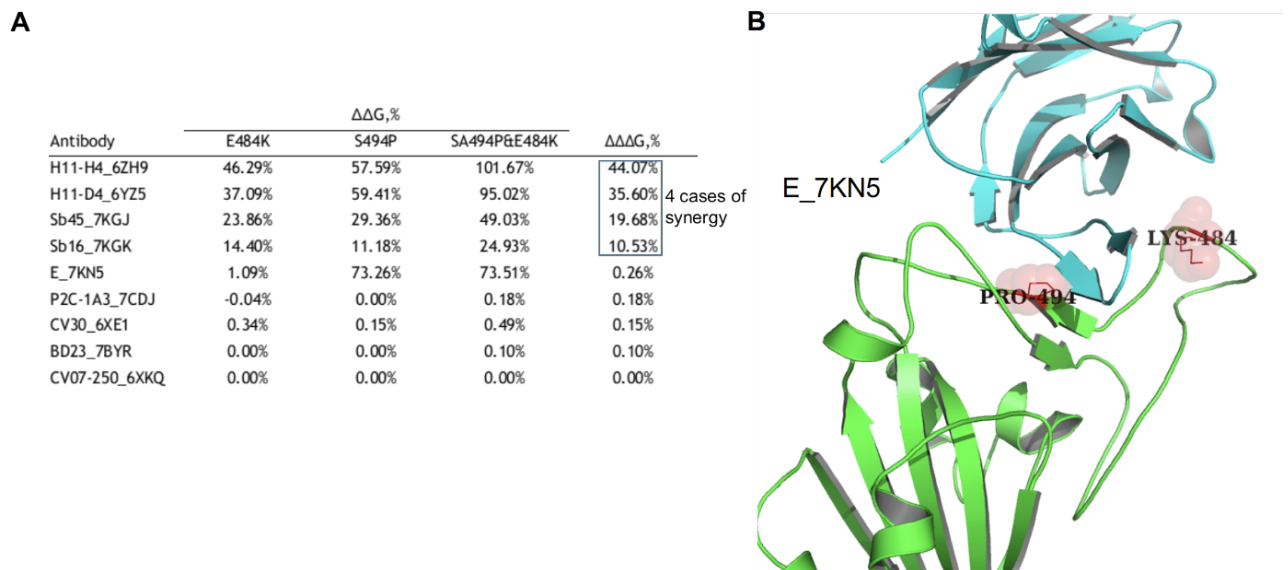


242

243

244 **Figure 9. Escape effects of the d156_158 mutation (B.1.1.523) and del156_157&R158G mutations**
245 **(delta variant) based on NTD-directed neutralizing antibody 4-8 Fab (PDBID: 7LQV).** Predicted
246 complex separation ΔG values for the mutants and the wild type complexes (A). The relationship
247 between structure deviation from the starting structure used for the docking with SnugDock and
248 the I_{sc} score (B).

249 All calculated FoldX binding energies are given in the supplementary file FileS1.xlsx. At least
250 in four cases a significant synergy in the effect of the E484 and S494P mutations were observed:
251 notably for the antibodies H11-H4_6ZH9, H11-D4_6YZ5, Sb45_7KGJ, Sb16_7KGK (Fig 10A). For the
252 most part the most significant antibody escape effect was noticed by the E484K mutation; however,
253 in some cases the S494P effect was also prominent. The largest effect by the S494P mutation was
254 observed for the E_7KN5¹² antibody structure (Fig. 10B): the S494P mutation increased the binding
255 $\Delta \Delta G$ by 74 % and the additional effect of E484 mutation was negligible (Fig. 10A).



256

257 **Figure 10. Escape effects of the E484, S494P mutations and their combination.** The $\Delta\Delta G$ values
 258 indicates relative increase in binding energy compared to the wild type structure as inferred from
 259 the FoldX calculations. The $\Delta\Delta\Delta G$ indicates the minimum difference between the $\Delta\Delta G$ of the double
 260 mutation and any of the two single point mutations. The large the value - the large the synergy (A).
 261 The structure of the antibody and receptor binding domain of the S protein complex which was
 262 affected by the S494P mutation most significantly (PDB ID: 7KN5) (B).

263 DISCUSSION

264 We have identified a new SARS-CoV-2 virus lineage with multiple mutations associated with
 265 immune escape and reported this to Pango at 5'th of May [[https://github.com/cov-lineages/pango-](https://github.com/cov-lineages/pango-designation/issues/69)
 266 [designation/issues/69](https://github.com/cov-lineages/pango-designation/issues/69)], which mandates the new lineage designation B.1.1.523. This Lineage was
 267 first determined in March 2021 and at the time of writing this article, the total amount of cases has
 268 reached 598 over 32 countries 12. It is likely that the rapid increase in circulation of Delta variant
 269 could have diminished the rise of B.1.1.523 lineage, however, the spread of the novel SARS-CoV-2
 270 lineage not only has not ceased, but even has started to rise.

271 Currently, a vast growth of B.1.1.523 can now be observed in Germany. Interestingly, the
 272 transmission of this lineage has diminished in Russia, where it was most expected to rise. This can
 273 be explained by different diagnostic strategy approaches in Russian Federation, where the testing
 274 is performed on non-randomly selected sources in the country. Alternatively, this could be explained
 275 by the steep rise of Delta variant in Russia, which started a month earlier than in Europe. E.g. in mid
 276 June we had >80% delta, while in Germany the same frequency of Delta was observed only in mid
 277 July.

278 The B.1.1.523 lineage possesses three or more mutations that characterizes SARS-CoV-2
 279 VOCs, including S:D156-158 deletion, S:E484K and S:S494P. D156-158 deletion at β -hairpin antigenic
 280 supersite, that is located at the same region typical for the Delta variant (E156G and 157-158del)¹³.
 281 E484K mutation has been detected in Beta variant (B.1.351) and VUM Zeta (B.1.1.28). The mutation
 282 is in the genomic region coding SARS-CoV-2 spike protein, and it appears to have a significant impact
 283 on the body's immune response and possibly, vaccine efficacy. On February 1st, Public Health

284 England (PHE) announced that the Covid-19 Genomics (COG-UK) consortium had identified this
285 same E484K mutation in 11 samples carrying the UK variant B.1.1.7 (sometimes called the Kent
286 variant), after analysing 214 159 sequences¹⁴.

287 An in vitro study for SARS-CoV-2 spike protein mutations that are responsible for antibody
288 evasiveness has identified that S494P mutation 15 reduce SARS-CoV-2 neutralization by 3-5-fold in
289 some convalescent sera. However, this mutation was not as potent at neutralization as E484K 16.
290 The results show that S494P mutation increases the spike protein stability. Also, applying docking
291 by HADDOCK displayed higher binding affinity to hACE2 for mutant spike than wild type possibly
292 due to the increased β -strand and turn secondary structures which increases surface accessibly
293 surface area (SASA) and chance of interaction. Currently deposited sequences in GISAID do not
294 support hypothesis that S:156_158del were combined with S:E484K and S:S494P during a
295 recombination event; rather it looks like that initially the triple deletion was introduced and then
296 followed addition of S:E484K and S:S494P.

297 We have showed by molecular modelling that in at least one case of antibody the triple
298 deletion del156-158 could decrease interaction. The combination with other immune escape
299 enhancing mutation at RBD this could result in highly resistant variant to immunity that was formed
300 by the initial virus variants. Delta variant also poses sequence changes at the S protein residues 156-
301 158 that can induce immune escape and recombination with the B.1.1.523 variant or de novo
302 introduction of the N484K and S494P mutations could make the Delta variant even more dangerous.
303 The case with Turkey sequences is controversial. The sequences at the S region significantly deviated
304 fragments with many “private” mutations. These sequences have been seen across several
305 submissions; hence, these might be not artificial. If these sequences are not artifacts, then the
306 Turkey sequences classified by Pangolin as belonging to B.1.1.523 lineage, resulted after a
307 recombination event between a highly diverged Turkey variant and a typical B.1.1.523 lineage
308 sequence. Turkey has been an extensive place of the virus spread with limited control and
309 reportability¹⁷; therefore, highly diverged variants could have evolved.

310 The results indicate that Pangolin classification should not be taken with granted. Out of the
311 two Turkey sequences that were assigned to the B.1.1.523 lineage, only one had characteristic to
312 the lineage SNP's at the S protein, the other one (Turkey/HSGM-B11931/2021) has mutations
313 characteristic for the Delta variant (double deletion and G residue at 156-158 region).

314

315 **CONCLUSIONS**

316 Presence and spread of SARS-CoV-2 B.1.1.523 lineage is evident regardless of the rapid
317 spread of the delta variant. This variant needs to be carefully observed and studied to keep a look
318 out for new mutations, that may cause even more harm in the Covid-19 pandemic. It is also
319 important to monitor other SARS-CoV-2 variants to keep track if this or similar mutations occur
320 spontaneously or by recombination.

321

322

323 METHODS

324 *Collection of SARS-CoV-2 sequences and initial data processing*

325 Sequences used for the analyses were downloaded from GISAID¹⁸ as for date of August 31st,
326 2021. Fasta files and metadata were extracted using ncov-ingest tool
327 [<https://github.com/nextstrain/ncov-ingest> (cloned at 2021 04 19)]. Lineages for all downloaded
328 sequences were assigned using pangolin 3.1.11 (pangoLEARN 2021-08-24 and pango-designation
329 v1.2.66)¹⁹ [<https://github.com/cov-lineages/pangolin>].

330 General sequence quality evaluation, extraction and alignment of S protein sequences,
331 variant calling was performed using Nextclade 1.3.0²⁰ [<https://github.com/nextstrain/nextclade>].

332

333 *Transmission cluster analysis*

334 In order to elucidate potential origin of the lineage and transmission cluster, a phylogeny
335 analysis of full genomes representing a small subset of GISAID has been done. The chosen sequences
336 for analysis composed from the union of two sets of sequences: (i) sequences that were assigned
337 B.1.1.523 lineage by pangolin, (ii) sequences that were at least 99.3 % identical to the Latvian
338 B.1.1.523 sequence EPI_ISL_1590462 and number of matched residues makes up equal or more
339 than 95% of the reference sequence. The reference sequence was chosen as it was closest to the
340 one of the first this lineage sequences sequenced at Lithuania but with smaller gaps regions. The
341 alignment against the GISAID sequences was conducted using minimap2 2.20-r1061
342 [<https://github.com/lh3/minimap2/>]. The limits were chosen arbitrary after several tries looking for
343 cut-offs resulting in a set of sequences that includes majority of sequences from the lineage and
344 some more diverged ones that are classified as belonging to other lineages by pangolin. The
345 sequences with quality control overall status being bad or having more than 1000 bps missing (as
346 indicated by Nextclade analysis) were discarded. The maximum likelihood tree was calculated using
347 a modified version of Nextstrain workflow [<https://github.com/nextstrain/zika> (cloned at 2021 05
348 01)]. The tree was build using IQ-TREE with 2.1.2 General time reversible model with unequal rates
349 and unequal base frequencies were used²¹ allowing for a proportion of invariable sites together with
350 discrete Gamma model²². Ultrafast bootstrap with 1000 replicates was used. The maximum
351 likelihood emergence time and origin of country for inner nodes were calculated by treetime 0.8.1
352 [<https://libraries.io/pypi/phylo-treetime>] as described by the aforementioned workflow. The set of
353 sequences collected as described above were clustered into transmission clusters using Phydentity
354 v2.0 [<https://github.com/alvinxhan/Phydentity>]²³.

355 *S protein phylogeny analysis*

356 The S protein-based phylogeny was based on the S protein sequences extracted from GISAID
357 by the Nextclade and aligned to the reference COVID-19 sequence. Sequences shorter than 1175
358 residues or having more than one stop codon or having any number of undetermined residues were
359 discarded. Sequences were further clustered into identical sequence clusters using CD-HIT v4.8.1
360 (command line option “-c 1.0 ”. The sequences representing all high-quality S protein variants were
361 used for maximum likelihood tree calculation by VeryFastTree 3.0.1 [DOI:

362 10.1093/bioinformatics/btaa582] using LG substitution model. The sequence alignment that was
363 used to construct the tree was composed from the alignment produced by the Nextclade leaving
364 only the representative sequences of the clusters and CAT approximation with 20 rate categories.
365 lg -gamma". Also command line flags that should increase calculations accuracy were added: "-spr
366 4 -mlacc 2 -slownni -double-precision ". The tree was re-rooted using sequence matching to the
367 EPI_ISL_402124 as an out-group using gotree 0.4.1 [<https://anaconda.org/bioconda/gotree/files>].

368 Ancestral states for inner nodes for 156, 157, 158, 484, 494 positions of protein S were
369 inferred using command line version GRASP-suite²⁵. This tool for inference was chosen due to high
370 speed and, most importantly, ability to handle insertion and deletions. The same substitution model
371 that used for phylogeny tree was also used in this case (LG). The resulting tree was analysed
372 detected potential changes in the haplotypes was done using a custom script written in julia 1.6
373 exploiting capabilities of the NewickTree library [<https://github.com/arzwa/NewickTree.jl>]. The tree
374 was trimmed to keep only those inner nodes that leads to leaves containing the triple deletion at
375 146-148 positions and either E484K, S494P and visualised using ggtree 3.0.426. Additionally, a set
376 of full genome sequences has been composed matching the leaves of the aforementioned trimmed
377 tree and corresponding maximum likelihood tree was calculated using the aforementioned
378 Nextstrain workflow.

379

380 ***Analysis of potential recombination events***

381 The focused set of sequences was composed based on the sequences used for the S protein
382 phylogeny. The set further narrowed to the sequences containing either the triple deletion at 156-
383 158 positions or a combination of E484K and S494P. The detection of recombination events at DNA
384 level was done using PoSeiDon workflow²⁷ [<https://github.com/hoelzer/poseidon> cloned at 2021
385 09 28] that runs GARD program to identify recombination events²⁸. The genomic regions matching
386 to the S protein were used. As the workflow is limited to 100 sequences the initial set 110 was
387 clustered to 100 sequences using cd-hit-est program from the CD-HIT package 4.8.1 using identity
388 cut-off equal to 0.9998658. The detection of recombination effects at protein level was done using
389 detREC tool²⁹ [https://github.com/qianfeng2/detREC_program (cloned at 2021-10-01)]. The
390 recombination detection was conducted using either genomic or protein sequence corresponding
391 to the S protein. The full genome scale recombination analysis was done for the same set of 110
392 sequences using 3SEQ program²⁸.

393 ***Antibody escape effect estimation***

394 Two S protein sequence variants 156_158del (B.1.1.523 lineage) and 156_157del & R158G
395 (delta variant) were modelled using Rosetta package (2021.16.61629_bundle)³¹. The escape effect
396 was evaluated based on the structure PDBID: 7LQV containing a NTD domain targeting binding
397 antibody. The S protein structure from the "A" chain with matching antibody was used. The S protein
398 residues from 283 position to the "N" terminus were discarded. The PDB structures for fragment
399 library generation and other required databases for Rosetta were downloaded at 2021 06 02. The S
400 structure models were created using a comparative modelling approach. The antibody structure
401 including side chains was kept constrained using CoordinateConstraintGenerator. For the relaxation

402 part “InterfaceRelax2019” script was used. The lowest energy model was chosen from 200 models.
403 The antibody docking was done using the SnugDock program³² from the Rosetta package, using
404 these non-default flags: “-auto_generate_h3_kink_constraint - h3_loop_csts_hr -h3_filter false -
405 docking_centroid_inner_cycles=20 -docking_centroid_outer_cycles=2 “. for each model docking
406 was run 300 times and the top 175 structures based on the interface score were chosen for analysis.
407 Antibody escape effect of E484K and S494P mutations were analysed using FoldX 5
408 program³³. The antibody structures targeting the receptor binding domain of the protein S were
409 downloaded from the CoV3D³⁴ as available at 2021 05 01. The mutations into the complexes were
410 introduced by consecutively applying FoldX command: RepairPDB, BuildModel, AnalyseComplex.
411 The residues at 484 and 494 positions were modelled either to the residue matching the mutation
412 or to the wild type residue and changes in free energy upon binding were evaluated.

413

414 **COMPETING INTERESTS**

415 The authors declare no competing interests.

416 REFERENCES

- 417 1. World Health Organisation. WHO Coronavirus (COVID-19) Dashboard. WHO Coronavirus
418 (COVID-19) Dashboard (2021).
- 419 2. F, W. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature*
420 **579**, 265–269 (2020).
- 421 3. European Centre for Disease Prevention and Control. How ECDC collects and processes
422 COVID-19 data. <https://www.ecdc.europa.eu/en/covid-19/data-collection> (2021).
- 423 4. S, E. & G, B.-M. Data, disease and diplomacy: GISAID’s innovative contribution to global
424 health. *Global challenges (Hoboken, NJ)* **1**, 33–46 (2017).
- 425 5. World Health Organisation. Tracking SARS-CoV-2 variants.
426 <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/> (2021).
- 427 6. European Centre for Disease Prevention and Control. SARS-CoV-2 variants of concern as of
428 30 September 2021. (2021).
- 429 7. Planas, D. *et al.* Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization.
430 *Nature* **2021 596:7871 596**, 276–280 (2021).
- 431 8. Collier, D. A. *et al.* SARS-CoV-2 B.1.1.7 escape from mRNA vaccine-elicited neutralizing
432 antibodies. *medRxiv* 2021.01.19.21249840 (2021) doi:10.1101/2021.01.19.21249840.
- 433 9. AJ, G. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding
434 domain that affect recognition by polyclonal human plasma antibodies. *Cell host & microbe* **29**, 463-
435 476.e6 (2021).
- 436 10. Lasek-Nesselquist, E., Pata, J., Schneider, E. & George, K. St. A tale of three SARS-CoV-2
437 variants with independently acquired P681H mutations in New York State. *medRxiv*
438 2021.03.10.21253285 (2021) doi:10.1101/2021.03.10.21253285.
- 439 11. Lewis, S., Stranges, B. P. & Adolf-Bryfogle, J. Interface Analyzer Documentation.
440 [https://www.rosettacommons.org/docs/latest/application_documentation/analysis/interface-](https://www.rosettacommons.org/docs/latest/application_documentation/analysis/interface-analyzer)
441 [analyzer](https://www.rosettacommons.org/docs/latest/application_documentation/analysis/interface-analyzer) (2016).
- 442 12. Koenig, P. A. *et al.* Structure-guided multivalent nanobodies block SARS-CoV-2 infection and
443 suppress mutational escape. *Science* **371**, (2021).
- 444 13. Latif, A. A. *et al.* B.1.1.523 Lineage Report. [https://outbreak.info/situation-](https://outbreak.info/situation-reports?pango=B.1.1.523)
445 [reports?pango=B.1.1.523](https://outbreak.info/situation-reports?pango=B.1.1.523) (2021).
- 446 14. McCallum, M. *et al.* Molecular basis of immune evasion by the delta and kappa SARS-CoV-2
447 variants. *bioRxiv* 2021.08.11.455956 (2021) doi:10.1101/2021.08.11.455956.
- 448 15. Wise, J. Covid-19: The E484K mutation and the risks it poses. *BMJ* **372**, n359 (2021).
- 449 16. Z, L. *et al.* Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum
450 antibody neutralization. *Cell host & microbe* **29**, 477-488.e4 (2021).
- 451 17. AJ, G. *et al.* Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding
452 Domain that Escape Antibody Recognition. *Cell host & microbe* **29**, 44-57.e9 (2021).
- 453 18. S, K. & A, K. Under-reporting of COVID-19 cases in Turkey. *The International journal of health*
454 *planning and management* **35**, 1009–1013 (2020).
- 455 19. Y, S. & J, M. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro*
456 *surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease*
457 *bulletin* **22**, (2017).
- 458 20. O’Toole, Á. *et al.* Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and
459 B.1.351/501Y-V2. *Wellcome Open Research* **2021 6:121 6**, 121 (2021).
- 460 21. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**,
461 4121–4123 (2018).
- 462 22. Tavaré, S. *Some probabilistic and statistical problems in the analysis of DNA sequences*. (The
463 American Mathematical Society, 1986).
- 464 23. X, G., YX, F. & WH, L. Maximum likelihood estimation of the heterogeneity of substitution
465 rate among nucleotide sites. *Molecular biology and evolution* **12**, 546–557 (1995).

- 466 24. AX, H., E, P., S, M.-S. & CA, R. Inferring putative transmission clusters with Phydelity. *Virus*
467 *evolution* **5**, (2019).
- 468 25. SQ, L. & O, G. An improved general amino acid replacement matrix. *Molecular biology and*
469 *evolution* **25**, 1307–1320 (2008).
- 470 26. Foley, G. *et al.* Identifying and engineering ancient variants of enzymes using Graphical
471 Representation of Ancestral Sequence Predictions (GRASP). *bioRxiv* 2019.12.30.891457 (2020)
472 doi:10.1101/2019.12.30.891457.
- 473 27. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in*
474 *Bioinformatics* **69**, e96 (2020).
- 475 28. M, H. & M, M. PoSeiDon: a Nextflow pipeline for the detection of evolutionary recombination
476 events and positive selection. *Bioinformatics (Oxford, England)* **37**, 1018–1020 (2021).
- 477 29. SL, K. P., D, P., MB, G., CH, W. & SD, F. GARD: a genetic algorithm for recombination detection.
478 *Bioinformatics (Oxford, England)* **22**, 3096–3098 (2006).
- 479 30. Feng, Q. *et al.* A scalable method for identifying recombinants from unaligned sequences.
480 *bioRxiv* 2020.11.18.389262 (2020) doi:10.1101/2020.11.18.389262.
- 481 31. HM, L., O, R. & MF, B. Improved Algorithmic Complexity for the 3SEQ Recombination
482 Detection Algorithm. *Molecular biology and evolution* **35**, 247–251 (2018).
- 483 32. Leman, J. K. *et al.* Macromolecular modeling and design in Rosetta: recent methods and
484 frameworks. *Nature Methods* 2020 17:7 **17**, 665–680 (2020).
- 485 33. Sircar, A. & Gray, J. J. SnugDock: Paratope Structural Optimization during Antibody-Antigen
486 Docking Compensates for Errors in Antibody Homology Models. *PLOS Computational Biology* **6**,
487 e1000644 (2010).
- 488 34. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic acids research* **33**,
489 (2005).
- 490 35. Gowthaman, R. *et al.* CoV3D: a database of high resolution coronavirus protein structures.
491 *Nucleic acids research* **49**, D282–D287 (2021).
- 492