Spread of endemic SARS-CoV-2 lineages in Russia

Galya V. Klink¹, Ksenia R. Safina², Sofya K. Garushyants¹, Mikhail Moldovan², Elena Nabieva², The CoRGI (Coronavirus Russian Genetic Initiative) Consortium⁵, Andrey B. Komissarov³, Dmitry Lioznov^{3,4}, Georgii A Bazykin^{2,1*}

¹A.A. Kharkevich Institute for Information Transmission Problems of the Russian Academy of Sciences, Russia

²Skolkovo Institute of Science and Technology (Skoltech), Russia

³Smorodintsev Research Institute of Influenza, Saint Petersburg, Russia

⁴ First Pavlov State Medical University, Saint Petersburg, Russia.

⁵ <https://corgi.center/en/> (see the list of consortium members in Supplementary File 1)

*Email: g.bazykin@skoltech.ru

In 2021, the COVID-19 pandemic is characterized by global spread of several lineages with evidence for increased transmissibility. Russia is among the countries with the highest number of confirmed COVID-19 cases, making it a potential hotspot for emergence of novel variants. Here, we show that among the globally significant variants of concern, B.1.1.7 (501Y.V1), B.1.351 (501Y.V2) or P.1 (501Y.V3), none have been sampled in Russia before January 2021. Instead, since summer 2020, the epidemic in Russia has been characterized by the spread of two lineages that are rare elsewhere: B.1.1.317 and a sublineage of B.1.1 including B.1.1.397 (hereafter, B.1.1.397+). In February-March 2021, these lineages reached frequencies of 26.9% (95% C.I.: 23.1%-31.1%) and 32.8% (95% C.I.: 28.6%-37.2%) respectively in Russia. Their frequency has increased in different parts of Russia. Together with the fact that these lineages carry several spike mutations of interest, this suggests that B.1.1.317 and B.1.1.397+ may be more transmissible than the previously predominant B.1.1, although there is no direct data on change in transmissibility. Comparison of frequency dynamics of lineages carrying subsets of characteristic mutations of B.1.1.317 and B.1.1.397+ suggests that, if indeed some of these mutations affect transmissibility, the transmission advantage of B.1.1.317 may be conferred by the (S:D138Y+S:S477N+S:A845S) combination; while the advantage of B.1.1.397+ may be conferred by the S:M153T change. On top of these lineages, in January 2021, B.1.1.7 emerged in Russia, reaching the frequency of 17.4% (95% C.I.: 12.0%-24.4%) in March 2021. Additionally, we identify three novel distinct lineages, AT.1, and two lineages prospectively named B.1.1.v1 and B.1.1.v2, that have started to spread, together reaching the frequency of 11.8% (95% C.I.: 7.5%-18.1%) in March 2021. These lineages carry combinations of several notable mutations, including the S:E484K mutation of concern, deletions at a recurrent deletion region of the spike glycoprotein (S:Δ140-142, S:Δ144 or S:Δ136-144), and nsp6:Δ106-108 (also known as ORF1a:Δ3675-3677). Community-based PCR testing indicates that these variants have continued to spread in April 2021, with the frequency of B.1.1.7 reaching 21.7% (95% C.I.: 12.3%-35.6%), and the joint frequency of B.1.1.v1 and B.1.1.v2, 15.2% (95% C.I.: 7.6%-28.2%). The combinations of mutations observed in B.1.1.317, B.1.1.397+, AT.1, B.1.1.v1 and B.1.1.v2 together with frequency increase of these lineages make them candidate variants of interest.

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

Introduction

Continuing evolution of SARS-CoV-2 in humans leads to emergence of new variants with novel epidemiological and/or antigenic properties. In spring 2020, the S:D614G change has spread globally due to its fitness advantage^{1,2}. Subsequently, a number of variants of concern, including B.1.1.7 (501Y.V1) first sampled in Great Britain in September, B.1.351 (501Y.V2) first sampled in South Africa in October, and P.1 (501Y.V3) first sampled in Brazil in December, were shown to be associated with increased transmissibility³⁻⁵. These variants are characterized by overlapping sets of changes in spike receptor-binding domain which affect ACE2 binding and antibody recognition, as well as other changes with demonstrated functional and antigenic effects. Emergence of SARS-CoV-2 variants with evidence for change in transmissibility, and possibly other properties, highlights the importance of continued surveillance of novel variants. In particular, locally arising variants that grow in frequency over time may suggest a transmission advantage, although such an increase may also occur by chance⁶.

Here, we show that the outbreak in Russia is characterized by a spread of two lineages, B.1.1.317 and B.1.1.397+, which are highly prevalent in Russia but rarely appear in non-Russian samples. We trace the accumulation of sequential mutations in the evolution of these lineages, and single out the spike mutations that are followed by a burst in frequency. If the frequency increase of B.1.1.317 and B.1.1.397+ has been indeed driven by changes in the intrinsic properties of the virus rather than by epidemiology, it is these mutations in spike that most likely have led to this increase, although importantly we lack direct transmission data to verify causality. We also describe three novel candidate variants of interest that are characterized by a rapid increase in frequency and combinations of important spike mutations, including the E484K mutation of concern.

Results

High-frequency variants in Russia

We analyzed 4,487 SARS-CoV-2 sequences with known sampling dates obtained in Russia between February 25, 2020 - March 28, 2021. 2,842 of these samples are deposited to GISAID⁷, while the remaining 1,645 (all dating to February-November 2020) will be made available through another repository. All 968 analyzed samples from December 2020 - March 2021 are available on GISAID. The vast majority of samples over this period came from several genomic surveillance programs which were not targeted towards particular variants, although representation of Russia's regions varied with time.

Throughout the pandemic, the SARS-CoV-2 diversity in Russia has been predominated by the B.1.1 Pango lineage which is frequent in Europe, as well as lineages descendant from it⁸ (Fig. 1). Three B.1.1-derived lineages with the higher prevalence in Russia in the beginning of 2021 were B.1.1.7, which was firstly introduced in Russia at the end of 2020, as well as two other lineages, B.1.1.317 and B.1.1.397, that appeared in Russia and July, 2020, respectively. B.1.1.317 was first sampled in Vietnam on March 27, 2020⁹; within Russia, it was first sampled on April 5, 2020 in Moscow, spreading across the country throughout 2020 (Fig. 2). B.1.1.397 was first sampled in the Krasnoyarsk Region of Russia on July 22, 2020. By summer 2020, both B.1.1.317 and B.1.1.397 were frequent throughout Russia (Fig. 2).

Fig. 1. Dynamics of Pango lineage frequencies in Russia (top row) and among the non-Russian samples in GISAID (bottom row). Asterisks in Pango lineage designations correspond to pooled sets of lineages of that hierarchy level, except those listed in other categories; e.g., B.1.1.* includes B.1.1 and B.1.1.6 but not B.1.1.7 or B.1.1.317. Samples are split into 1 month bins.

Fig. 2. The spatio-temporal distribution of frequent lineages in Russia.

To find the non-reference amino acid variants that gained in frequency in Russia, we selected the positions at which the mean frequency of the non-reference variant in Russian samples exceeded 5% (for the spike) or 10% (for other proteins) in February-March 2021. We found 21 such positions in spike and 21 such positions in other proteins. Among these changes, two (RdRp:P323L and S:D614G) were fixed early in the global evolution of SARS-CoV-2; other two (N:R203K and N:G204R) are the lineage-defining mutations of B.1.1.

The frequency dynamics of the derived variants at the remaining 38 positions is shown in Fig. 3. These include the mutations characterizing the B.1.1.7 variant which has been increasing in frequency in Russia since January 2021 (Fig. 1), as well as some of the other globally spreading mutations of concern or interest, including the E484K mutation in spike. However, at many of these sites, the non-reference variants were rare outside Russia (Fig. 3). Most of these variants showed similar temporal dynamics in Moscow and St. Petersburg regions, as well as in the European and Asian parts of Russia (Fig. S1), indicating that their increase in frequency is not a result of sampling bias.

We aimed to identify the high-frequency variants carrying these mutations. Many of these sites were highly homoplastic, and overall we found the resulting phylogenies not to be robust. Instead, we defined the most frequent variants composed of these mutations, independent of the alleles at other sites (Fig. 4).

We considered the allele combinations that were most frequent in Russia in February-March 2021, and noticed that they belonged to several nested sets. The most frequent combination (99 out of the 461 samples) carried the N:A211V mutation which is characteristic of the B.1.1.317 Pango lineage; the second most frequent combination carried the S:D138Y and ORF8:V62L combination of mutations which are characteristic of the B.1.1.397 lineage; the third combination carried the set of characteristic mutations of B.1.1.7.

Still, there was no one-to-one correspondence between the frequent combinations of mutations and Pango lineages. For example, while the most frequent variant carrying S:M153T was that also including N:M234I, S:D138Y and ORF8:V62L (column 2 in Fig. 4), and classified as Pango lineage B.1.1.397, the variants carrying S:M153T alone, the S:M153T+N:M234I combination and the S:M153T+N:M234I+S:N679K combination were also frequent (columns 4, 9 and 5 in Fig. 4 respectively) but were classified by PANGOLIN as other lineages (B.1.1, B.1.1.141, B.1.1.28 and others). Similarly, while B.1.1.317 is defined by the N:A211V mutation, the frequency of the variant carrying this mutation alone is currently relatively low (column 12 in Fig. 4), while most samples carrying it also carry 8 additional high-frequency mutations, including four potentially important changes in spike (Q675R+D138Y+S477N+A845S; column 1 in Fig. 4). The frequencies of such “non-canonical” combinations of mutations increased throughout 2020-2021 (Fig. 5).

Finally, we observe three high-frequency combinations of mutations, including the S:E484K mutation of concern as well as other mutations of interest (notably S:Δ140-142, S:Δ136-144 and nsp6:Δ106-108, also referred to as ORF1a:Δ3675-3677; columns 7, 8 and 10 in Fig. 4). One of these combinations (column 8 in Fig. 4) has recently got the AT.1 Pango designation. The remaining two (columns 7 and 10) currently lack Pango designations, and are hereafter referred to as B.1.1.v1 and B.1.1.v2.

Fig. 3. Frequency dynamics of SARS-CoV-2 amino acid changes. Plots represent changes in frequency over time for the non-reference amino acid mutations that reached frequencies above 10% (5% for the S protein) among the 461 Russian samples obtained in February-March 2021. The frequency of B.1.1.7 is represented by the mutation nsp3:I1412T; deletions nsp6:Δ106-108 and S:Δ144 and substitution S:P681H which are a part of B.1.1.7 as well as other lineages are shown separately; the remaining 14 mutations such that >70% of samples carrying them belonged to B.1.1.7 are not shown. Changes in frequency in Russian (red) and non-Russian (blue) samples are shown in one-month time intervals. Shaded areas show 95% confidence intervals (Wilson score intervals).

Fig. 4. Variants with high frequencies in Russia in February-March. Horizontal rows represent all positions with non-reference alleles at frequency above 5% (for spike protein) or 10% (for other proteins) in Russia in February-March. Columns represent all observed combinations of these variants that included 2 or more samples, with black or colored dots indicating the presence of the non-reference variant. Colors dots represent the variants that are discussed in the text, with the same color coding as in Figs. 1, 2 and 5; blue color corresponds to variants B.1.1.v1 (column 7), AT.1 (column 8) and B.1.1.v2 (column 10).

Fig. 5. Mutational composition and frequency dynamics of the B.1.1.317 and B.1.1.397+ lineages. A, B: schematic representation of the B.1.1.317 and B.1.1.397+ lineages. Pango lineage designations are approximate. C: Muller plots representing the frequency dynamics of the corresponding combinations of mutations in Russia.

Frequency dynamics of the variants prevalent in Russia

To study potential effects of individual mutations composing a variant on the frequency dynamics of this variant, we fit the logistic growth model for the 10 most-frequent combinations of mutations and for N:A211V (which is the 12th most-frequent combination), and compared the dynamics of nested combinations with each other (Figs. 6-8).

The variant carrying just the N:A211V change (largely coincident with the B.1.1.317 Pango lineage) has increased in frequency since the start of the epidemic in Russia. However, since fall 2020, it is being displaced by the variant with 8 additional mutations, including four in spike: Q675R+D138Y+S477N+A845S. When the logistic growth model is fit to the N:A211V variant alone, it demonstrates modest growth (Fig. 6A); however, its combination with S:Q675R+D138Y+S477N+A845S demonstrates a much more rapid frequency increase, with the estimated daily growth rate of 1.93% (95% CI: 1.8%–2.06%; Fig. 6B). While this leads to a frequency increase of the N:A211V mutation independent of the background (Fig. 6C), this suggests that the frequency increase is more likely driven by the S:Q675R+D138Y+S477N+A845S combination than the N:A211V change defining the B.1.1.317 Pango lineage.

183 **Fig. 6. Logistic growth model for nested variants defined by amino acid changes in the**
 184 **N:A211V context.** Red dots, sliding window 14-day average frequency; shaded area, 95%
 185 confidence interval. Variants are identified according to the presence of the mutations in S and
 186 N proteins; see Fig. 5 for complete lists of mutations in the corresponding variants.

187 By contrast, the frequency of the S:M153T mutation grows independently of the presence of
 188 other mutations from our list (Fig. 7). While subsequent mutations may add to the estimated
 189 growth rates, these rates are comparable when the S:M153T mutation occurs alone or in
 190 combination with N:M234I, N:M234I+S:N679K, or N:M234I+S:D138Y, and all these
 191 combinations are still frequent many months after they all originated (Fig. 5, 7).

192

Fig. 7. Logistic growth model for variants defined by amino acid changes in the S:M153T context. Notations are the same as in Fig. 6.

Finally, the five remaining variants which also reached high frequency in February-March carry unnested, although partially overlapping sets of mutations. These include B.1.1.7 (Fig. 8A); a variant carrying the S:P681H mutation of interest in the absence of other high-frequency mutations (Fig. 8B); as well as three novel variants carrying the following combinations of

200 mutations: (i) nsp6:Δ106-108+S:P9L+S:Δ140-142 (nine of these 14 samples are currently
 201 classified by PANGOLIN as B.1.1.74, four as B.1.1, and one as B.1.1.354; this variant is
 202 referred to here as B.1.1.v1; Fig.8C); (ii) S:P9L+S:Δ136-144+S:E484K (which recently got the
 203 AT.1 Pango designation; Fig.8D); and (iii) nsp6:Δ106-108+S:Δ144+S:E484K (currently
 204 classified by PANGOLIN as B.1.1; this variant is referred to here as B.1.1.v2; Fig.8E). Variants
 205 B.1.1.v1, AT.1 and B.1.1.v2 were only observed in 13-14 samples each, but are of interest
 206 because this constitutes an appreciable fraction of samples obtained in February-March (3.0%,
 207 2.8% and 2.6% respectively), and also because they are composed of known mutations of
 208 interest or concern. The daily growth rate estimated for these variants by the logistic growth
 209 model is in the range of 2.44% to 7.18% (Fig. 8C-E).

Fig. 8. Logistic growth model for the five remaining amino acid variants with high frequencies in Russia in February-March. Notations are the same as in Fig. 6. S:P681H has been observed both independently and as part of B.1.1.7; in panel B, the cases of B.1.1.7 are not shown, by excluding the S:P681H+nsp3:I1412T combination.

The continued spread of some of these variants between February-April 2021 is confirmed by community-based PCR testing. To obtain independent frequency estimates, we made use of a

PCR system sensitive to the presence of nsp6:Δ106-108 and S:Δ69-70 deletions (see Methods) to detect the B.1.1.7, B.1.1.v1 and B.1.1.v2 variants. Specifically, S:Δ69-70⁺ nsp6:Δ106-108⁺ samples correspond to B.1.1.7, while S:Δ69-70⁻ nsp6:Δ106-108⁺ samples correspond to either the B.1.1.v1 or the B.1.1.v2 variant (Fig. 4). While the frequency estimates were highly uncertain (Table 1), they indicate that B.1.1.7, and one or both of variants B.1.1.v1 and B.1.1.v2, were wide-spread in April (Fig. 9, Table 1). A considerable fraction (59.6%) of PCR samples from February and March were included in our main analysis, as their sequences were in GISAID. However, the frequency increase was also observed in the 136 PCR samples for which no sequencing data was available (Fig. S2), providing independent validation of the NGS results. Similarly, it was observed when the PCR tests only for St. Petersburg were analysed (Fig. S3, S4), indicating that the prevalence of these variants increases at least in this city as opposed to being an artefact of changing sampling between regions.

Figure 9. Frequencies of S:Δ69-70, nsp6:Δ106-108, and their combination in Russia in Feb-Apr 2021 based on PCR data. S:Δ69-70⁺ nsp6:Δ106-108⁺ samples correspond to B.1.1.7, and S:Δ69-70⁻ nsp6:Δ106-108⁺ samples correspond to B.1.1.v1 or B.1.1.v2. The rare instances of S:Δ69-70⁺ nsp6:Δ106-108⁻ probably correspond to false positive S:Δ69-70 detection. Notations for logistic curves are the same as in Fig. 6.

Table 1. Frequencies of (B.1.1.v1 or B.1.1.v2) and B.1.1.7 estimated from PCR data. The point estimate and the 95% confidence intervals (Wilson score intervals) are shown.

	2021-Feb	2021-Mar	2021-Apr
nsp6:Δ106-108 ⁺ S:Δ69-70 ⁻ (B.1.1.v1 or B.1.1.v2)	10.5% (2.3%-31.4%)	6.9% (4.1%-11.2%)	15.2% (7.6%-28.2%)
nsp6:Δ106-108 ⁺ S:Δ69-70 ⁺ (B.1.1.7)	10.5% (2.9%-31.4%)	13.7% (9.7%-19.1%)	21.7% (12.3%-35.6%)

Mutational composition of the high-frequency variants

In this section, we discuss the mutations that constitute the variants spreading in Russia.

B.1.1.317. This lineage is defined by the presence of the N:A211V mutation. Changes at nucleocapsid position 211 experience both persistent (according to the FEL model of HyPhy¹⁰) and episodic (according to the MEME model¹¹) positive selection both in the global¹² and in the Russian dataset ($p=0.0396$ for the MEME model and $p=0.0268$ for the FEL model, the likelihood-ratio test), as well as a rapid increase in frequency of non-reference variants in the

global dataset¹². While the global frequency of 211V has remained low (<0.4%), in Russia, it has reached 26.9% in February-March 2021. According to immunoinformatic analysis, site N:211 is included in one of the four regions of the nucleocapsid protein with the highest affinity to multiple MHC-I alleles¹³. Nevertheless, the frequency of the variant carrying the N:A211V mutation alone has declined since October 2020 (Fig. 5), suggesting that it is unlikely to confer transmission advantage against the background of other currently frequent variants (Fig. 6).

A rapidly spreading subclade within B.1.1.317 carries the (Q675R+D138Y+S477N+A845S) combination of changes in spike. Two of these mutations are of interest. S:D138Y, first described as one of the lineage-defining mutations of the P.1 lineage¹⁴, is a change in the N-terminal domain (NTD) of spike. Site 138 is adjacent to the NTD antigenic supersite, and together with other NTD mutations of P.1, S:D138Y was suggested to be the cause of disruption of binding with mAb159¹⁵ which is one of the most potent inhibitory antibodies¹⁶. S:S477N is positioned in the receptor-binding motif (RBM) of the S-protein near the antibody binding site (Fig. 10) and was reported to promote resistance to multiple antibodies and plasma from convalescent patients^{17,18}. Additionally, S477N is thought to increase ACE2 binding¹⁹. It is one of the lineage-defining mutations of the B.1.160 (20A.EU2) variant of interest prevalent in Europe in Autumn 2020²⁰, and the only one among them to occur in the S-protein; it also defines one of the two subclades of the B.1.526 variant of interest currently spreading in the USA^{21,22}. S:Q675R is located at the central part of S1 (Fig. 10), and S:A845S, in S2; the significance of these two mutations is unknown.

B.1.1.397+. S:M153T is a characteristic mutation of B.1.1.397, which also comprises several other mutations. However, the frequency of S:M153T in Russia also increases in the absence of these other mutations (Fig. 7). This increase has been ongoing since late spring 2020 (Fig. 5), and has been noticed in Russia^{23,24}. S:M153T, however, has remained rare outside Russia. S:153 is the first position of the 6-amino acid insertion specific to SARS-CoV-2 and some closely related bat betacoronaviruses that was absent in SARS-CoV²⁵. While the effect of S:M153T on antigenic properties is unknown, S:153 is a part of the N3 loop of the NTD. This loop is a part of the NTD antigenic supersite (Fig. 10; ²⁶), and nearby residues, including S:152, were recently shown to bind highly neutralizing 4A8 antibody from a convalescent patient²⁷. Besides Russia, S:M153T is prevalent in Kazakhstan²⁸ which has a long border with Russia, suggesting common ancestry of this change in these countries.

The most frequent subclade within B.1.1.397+, and the one with some evidence for an independent increase in frequency (Fig. 7), is defined by the presence of two additional mutations of interest: S:D138Y discussed above in the context of the B.1.1.317 lineage (but acquired in the B.1.1.397 lineage independently); and N:M234I. Position N:234 is a part of a disordered linker domain of the nucleocapsid protein²⁹. Outside B.1.1.397, the N:M234I change has also occurred independently in several lineages that attracted attention. It is among the lineage-defining mutations of the B.1.160 (20A.EU2) lineage as well as the B.1.526 lineage that increases in frequency in the USA at rates comparable to those of B.1.1.7²¹. It is also one of the changes defining a newly detected lineage (preliminarily identified as B.1.x) which also contains S:N501Y and S:P681H and seems to spread rapidly in the USA³⁰. Independent emergence of N:M234I in several variants of interest may reflect its impact on at least one of multiple functions of the N protein³¹.

Other notable variants. The five other combinations of mutations observed at high frequencies in Russia in February-March 2021 are B.1.1.7, the best-known variant with increased transmissibility; the variant carrying the S:P681H mutation alone; and three novel variants.

S:P681H is one of the nine spike changes that characterize the rapidly spreading B.1.1.7 lineage³; however, it is absent from the two other lineages of concern, B.1.351 and P.1, indicating that it is not essential for increased transmissibility. The 681 position is adjacent to the furin cleavage site; this site is absent in non-human CoV, and is assumed to have contributed to pathogenicity in humans³². Changes at this position experience both persistent and episodic positive selection¹². P681H appears to increase in frequency globally³³, although it is hard to

299 disentangle this increase from that of the other changes constituting the rapidly spreading
300 B.1.1.7 lineage. We find that the frequency of this mutation in Russia in the absence of other
301 B.1.1.7 mutations does not increase (Fig. 8), indicating that it does not increase transmissibility
302 by itself.

303 The three remaining high-frequency variants with evidence for rapid frequency increase carry
304 combinations of the following high-frequency mutations: S:P9L, S:Δ140-142 (or S:Δ136-144),
305 S:E484K, and nsp6:Δ106-108. The sets of mutations in these variants are in conflict (i.e., not
306 nested within each other; Fig. 4), indicating that at least some of these mutations emerged in
307 them independently. These mutations are of interest or concern. Specifically, S:E484K (present
308 in AT.1 and B.1.1.v2) is involved in multiple variants of concern including the B.1.351
309 (501Y.V2)⁴, P.1 (501Y.V3)^{5,34} and P.2 (S.484K)^{5,34} lineages, and has been shown by several
310 groups to cause escape from neutralizing antibodies³⁵⁻³⁷. nsp6:Δ106-108 (also referred to as
311 ORF1a:Δ3675-3677, and present in B.1.1.v1 and B.1.1.v2) is a part of all three variants of
312 concern. S:Δ140-142 (present in B.1.1.v1), S:Δ144 (present in B.1.1.v2) and S:Δ136-144
313 (present in AT.1) are distinct deletions at a recurrent deletion region of the spike glycoprotein
314 which confer resistance to neutralizing antibodies³⁸.

315
316

Fig. 10. Position of residues 138, 153, 477 and 675 in the spatial structure of the S protein bound with 4A8 and 4-59 antibodies (PDB IDs: 7c2l and 7czx). Each of these residues is colored in its own color. Yellow, receptor binding domain; green, receptor binding motif; ocean blue, heavy chain of the 4A8 antibody; blue, heavy chain of the 4-59 antibody; olive, antibody binding epitope.

Discussion

Russia has been relatively well isolated in the course of COVID-19 pandemic: both the first cases of COVID-19 and the arrival of variants of concern, notably the B.1.1.7, have happened here later than in many European countries ^{8,39}. Together with the large size of the outbreak in Russia, such isolation could have created conditions for emergence of novel important domestic variants.

A steady increase in frequency of lineages B.1.1.317 and B.1.1.397+, as well as the presence of multiple mutations with potential effect on antigenic properties, notably S:D138Y, merit classification of these two lineages as variants of interest^{40,41}. Nevertheless, the rate of spread of B.1.1.317 and B.1.1.397+ has been lower than that of VOCs (e.g., ~7% for B.1.1.7⁴²). In particular, while B.1.1.317 has been observed in Russia since April 2020, the subclade of B.1.1.317 carrying the three spike mutations, since July 2020, and B.1.1.397+, since April 2020, all these lineages have remained at frequencies below 30% in Russia, and the logistic growth rates estimated by our model are all below 2% (Fig. 6-8). Besides, these variants are currently missing spike changes L452R, E484K or N501Y which occur in other VOCs⁴¹.

The combinations of mutations seen in the three variants that emerged in 2021, AT.1, B.1.1.v1 and B.1.1.v2 (Fig. 8C-E), look more suspicious, because their estimated rate of frequency increase is higher and because they include mutations with known effects and occurring in other variants of interest or concern. While little can be told about their frequency dynamics on the basis of the currently available data, they require careful monitoring.

Methods

Dataset preparation. 1,060,545 sequences of SARS-CoV-2 were downloaded from GISAID on April 15, 2021 (Supplementary File 2) and aligned with MAFFT⁴³ v7.45324 against the reference genome Wuhan-Hu-1/2019 (NCBI ID: MN908947.3) with --addfragments --keeplength options. 100 nucleotides from the beginning and from the end of the alignment were trimmed. After that, we excluded sequences (1) shorter than 29,000 bp, (2) with more than 3,000 (for Russian sequences) or 300 (for all other countries) positions of missing data (Ns), (3) excluded by Nextstrain⁴⁴, (4) in non-human animals, (5) with a genetic distance to the reference genome more than four standard deviations from the epi-week mean genetic distance to the reference, or (6) with incomplete collection dates. As our focus was on the spread of lineages in Russia, and since Russia is relatively poorly sampled, we chose a less stringent threshold at step (2) for Russian compared to non-Russian sequences in order to keep more Russian data in the dataset. To this dataset, we added the 1,645 Russian sequences described in our earlier study⁴⁵ and 344 samples produced by the CoRGI consortium which had not been yet available in GISAID on April 15 (but have been deposited to GISAID since then). The final dataset consisting of 830,249 sequences, including 4,487 Russian samples, was then annotated by the PANGOLIN package (v2.3.8). For the categories in Fig. 1, we selected Pango lineages represented by at least 100 Russian samples (there were six such lineages: B.1, B.1.1, B.1.1.397, B.1.1.28, B.1.1.317 and B.1.1.294), added the two earliest lineages A and B and the VOC lineage B.1.1.7, and aggregated A, B, B.1 and B.1.1 with lineages nested in them except those included in other categories into A.*, B.*, B.1.* and B.1.1.*, respectively.

Data analysis. To trace the frequency dynamics of mutations, for each non-reference nucleotide at each position in each month, we calculated the fraction of samples with this nucleotide among all samples where this position contained an ambiguous nucleotide. We did this separately for the Russian and non-Russian samples, and selected changes with frequency more than 5% in S protein and more than 10% in other proteins in February-March 2021 (a total of 461 sequences) in Russian samples for consideration. Calculations were performed with custom Perl and R scripts. Wilson score intervals were estimated using Hmisc R package⁴⁶, and results were visualised with the ggplot2 package⁴⁷ of R language⁴⁸. Upset plots were built with the UpSetR package⁴⁹ for R. Logistic growth models were fit to frequencies of each variant averaged across 14-days sliding windows with nls() function of R language⁴⁸. Only windows with total number of samples > 20 were taken into account. Confidence intervals for estimated parameters were obtained with confint2 function from nlstools R package⁵⁰, similar to ⁴². Results were visualised with R packages ggplot2⁴⁷, tidyverse⁵¹ and gridExtra⁵². To estimate positive selection, we employed MEME and FEL models implemented in the HyPhy package^{10,11}. For this analysis, we added Russian sequences with incomplete collection dates to

the main dataset, which resulted in 5006 Russian sequences. The tree for the selection analysis was built upon the whole-genome alignment of Russian sequences with the RAXML package v.8.0.26 (model GTRCAT)⁵³.

Mapping residues onto the S-protein structure. To visualize spike mutations, we utilized two different spike protein structures, corresponding to S-protein in complex bound with 4A8 (PDB ID: 7c2l) or with P5A-1B9 (PDB ID: 7czx) antibody. The NTD antibody binding epitope is defined as in²⁶. The two structures were structurally aligned and visualized with Open-Source PyMOL⁵⁴.

PCR data. Community-based PCR tests aimed at detection of S:Δ69-70 and nsp6:Δ106-108 deletions were performed for 739 samples. We further analysed only those samples for which both tests were performed and produced unambiguous results. There were 269 such samples from 22 regions (including 170 from Saint Petersburg, 43 from Sverdlovsk Oblast, and 12 from Leningrad Oblast) obtained between February-April, 2021. For S:Δ69-70 detection⁵⁵, we used the Yale69/70del RT-PCR assay described elsewhere. For nsp6:Δ106-108 detection, we used a newly designed RT-PCR assay. 133 of these 269 samples were also sequenced (sequence data made available through GISAID); for 126 of them (94.7%), the results on the presence of S:Δ69-70 and nsp6:Δ106-108 were consistent between the NGS and PCR data, indicating that our PCR tests are highly specific.

Acknowledgements

We are grateful to all GISAID submitting and originating labs (Supplementary File 1) for rapid open release of SARS-CoV-2 sequencing data. We thank Sergei L Kosakovsky Pond for help with HyPhy analyses, and Evgeniya Alekseeva and members of the Bazykin lab for fruitful discussions. This work was supported by the RFBR grant 20-54-80014 to G.A.B.

References

1. Korber, B. et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e19 (2020).
2. Volz, E. et al. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184**, 64-75.e11 (2021).
3. ECDPC. Rapid increase of a SARS-CoV-2 variant with multiple spike protein mutations observed in the United Kingdom. (2020).
4. Tegally, H. et al. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv* 2020.12.21.20248640 (2020) doi:10.1101/2020.12.21.20248640.
5. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. *Virological* <https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586> (2021).

4196. Hodcroft, E. B. et al. Emergence and spread of a SARS-CoV-2 variant through Europe in the
420 summer of 2020. *medRxiv* 2020.10.25.20219063 (2021) doi:10.1101/2020.10.25.20219063.
4217. Eurosurveillance | GISAID: Global initiative on sharing all influenza data – from vision to reality.
422 <https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2017.22.13.30494>.
4238. Komissarov, A. B. et al. Genomic epidemiology of the early stages of the SARS-CoV-2 outbreak
424 in Russia. *Nat. Commun.* **12**, 649 (2021).
4259. CoVizu. Near real-time visualization of SARS-CoV-2 (hCoV-19) genomic variation
426 <https://filogeneti.ca/CoVizu/> (2021).
42710. Kosakovsky Pond, S. L. & Frost, S. D. W. Not So Different After All: A Comparison of Methods
428 for Detecting Amino Acid Sites Under Selection. *Mol. Biol. Evol.* **22**, 1208–1222 (2005).
42911. Murrell, B. et al. Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLOS*
430 *Genet.* **8**, e1002764 (2012).
43112. Selection history of genes in SARS-CoV-2/COVID-19 genomes enabled by data from.
432 <https://observablehq.com/@spond/sc2-genes> (2021).
43313. Oliveira, S. C., de Magalhães, M. T. Q. & Homan, E. J. Immunoinformatic Analysis of SARS-
434 CoV-2 Nucleocapsid Protein and Identification of COVID-19 Vaccine Targets. *Front. Immunol.*
435 **11**, (2020).
43614. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil | *Science*.
437 <https://science.sciencemag.org/content/early/2021/04/13/science.abh2644>.
43815. Dejnirattisai, W. et al. Antibody evasion by the P.1 strain of SARS-CoV-2. *Cell* (2021)
439 doi:10.1016/j.cell.2021.03.055.
44016. Dejnirattisai, W. et al. The antigenic anatomy of SARS-CoV-2 receptor binding domain. *Cell*
441 **184**, 2183-2200.e22 (2021).
44217. Gaebler, C. et al. Evolution of Antibody Immunity to SARS-CoV-2. *bioRxiv* 2020.11.03.367391
443 (2021) doi:10.1101/2020.11.03.367391.
44418. Liu, Z. et al. Landscape analysis of escape variants identifies SARS-CoV-2 spike mutations that
445 attenuate monoclonal and serum antibody neutralization. *bioRxiv* 2020.11.06.372037 (2021)
446 doi:10.1101/2020.11.06.372037.

44719. Serine 477 plays a crucial role in the interaction of the SARS-CoV-2 spike protein with the
448 human receptor ACE2 | Scientific Reports. [https://www.nature.com/articles/s41598-021-83761-](https://www.nature.com/articles/s41598-021-83761-5)
449 5.
45020. CoVariants. <https://covariants.org/variants/20A.EU2>.
45121. A Novel and Expanding SARS-CoV-2 Variant, B.1.526, Identified in New York | medRxiv.
452 <https://www.medrxiv.org/content/10.1101/2021.02.23.21252259v2.full>.
45322. West, A. P. *et al.* Detection and characterization of the SARS-CoV-2 lineage B.1.526 in New
454 York. *bioRxiv* 2021.02.14.431043 (2021) doi:10.1101/2021.02.14.431043.
45523. The expert reported on the mutation of the coronavirus in 13 regions of the Russian Federation
456 (in Russian). <https://www.interfax.ru/russia/737621>.
45724. Dangerous COVID-19 mutations, which Popova warned about, were found in the Urals (in
458 Russian). *vesti.ru* <https://www.vesti.ru/article/2486564>.
45925. Guruprasad, L. Evolutionary relationships and sequence-structure determinants in human
460 SARS coronavirus-2 spike proteins for host receptor recognition. *Proteins Struct. Funct.*
461 *Bioinforma.* **88**, 1387–1393 (2020).
46226. Cerutti, G. *et al.* Potent SARS-CoV-2 Neutralizing Antibodies Directed Against Spike N-Terminal
463 Domain Target a Single Supersite. *bioRxiv* 2021.01.10.426120 (2021)
464 doi:10.1101/2021.01.10.426120.
46527. Chi, X. *et al.* A neutralizing human antibody binds to the N-terminal domain of the Spike protein
466 of SARS-CoV-2. *Science* **369**, 650–655 (2020).
46728. outbreak.info. *outbreak.info* <https://outbreak.info/>.
46829. Cubuk, J. *et al.* The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase
469 separates with RNA. *Nat. Commun.* **12**, 1936 (2021).
47030. Thornlow, B. *et al.* A new SARS-CoV-2 lineage that shares mutations with known Variants of
471 Concern is rejected by automated sequence repository quality control. *bioRxiv*
472 2021.04.05.438352 (2021) doi:10.1101/2021.04.05.438352.
47331. Gao, T. *et al.* Identification and functional analysis of the SARS-COV-2 nucleocapsid protein.
474 *BMC Microbiol.* **21**, 58 (2021).

47532. The sequence at Spike S1/S2 site enables cleavage by furin and phospho-regulation in SARS-
476 CoV2 but not in SARS-CoV1 or MERS-CoV | Scientific Reports.
477 <https://www.nature.com/articles/s41598-020-74101-0>.
47833. Maison, D. P., Ching, L. L., Shikuma, C. M. & Nerurkar, V. R. Genetic Characteristics and
479 Phylogeny of 969-bp S Gene Sequence of SARS-CoV-2 from Hawaii Reveals the Worldwide
480 Emerging P681H Mutation. *bioRxiv* 2021.01.06.425497 (2021) doi:10.1101/2021.01.06.425497.
48134. Voloch, C. M. *et al.* Genomic characterization of a novel SARS-CoV-2 lineage from Rio de
482 Janeiro, Brazil. *medRxiv* 2020.12.23.20248598 (2020) doi:10.1101/2020.12.23.20248598.
48335. Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody
484 neutralization: Cell Host & Microbe. [https://www.cell.com/cell-host-microbe/fulltext/S1931-](https://www.cell.com/cell-host-microbe/fulltext/S1931-3128(21)00044-5)
485 [3128\(21\)00044-5](https://www.cell.com/cell-host-microbe/fulltext/S1931-3128(21)00044-5).
48636. Andreano, E. *et al.* SARS-CoV-2 escape in vitro from a highly neutralizing COVID-19
487 convalescent plasma. *bioRxiv* 2020.12.28.424451 (2020) doi:10.1101/2020.12.28.424451.
48837. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding
489 domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **0**,
490 (2021).
49138. McCarthy, K. R. *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody
492 escape. *Science* **371**, 1139–1142 (2021).
49339. The first case of the ‘British variant’ of the coronavirus is detected in Russia (in Russian).
494 *Kommersant* <https://www.kommersant.ru/doc/4639704> (2021).
49540. Emma Griffiths *et al.* CanCOGeN Interim Recommendations for Naming, Identifying, and
496 Reporting SARS-CoV-2 Variants of Concern.
49741. CDC. SARS-CoV-2 Variant Classifications and Definitions. *Centers for Disease Control and*
498 *Prevention* [https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-](https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html)
499 [surveillance/variant-info.html](https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html) (2021).
50042. Washington, N. L. *et al.* Genomic epidemiology identifies emergence and rapid transmission of
501 SARS-CoV-2 B.1.1.7 in the United States. *medRxiv* 2021.02.06.21251159 (2021)
502 doi:10.1101/2021.02.06.21251159.

50343. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
504 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
50544. <https://github.com/nextstrain/ncov/blob/master/defaults/exclude.txt>. *GitHub*
506 <https://github.com/nextstrain/ncov/blob/master/defaults/exclude.txt>.
50745. Matsvay, A. *et al.* Genomic epidemiology of SARS-CoV-2 in Russia reveals recurring cross-
508 border transmission throughout 2020. *medRxiv* 2021.03.31.21254115 (2021)
509 doi:10.1101/2021.03.31.21254115.
51046. Harrel, F. E., Jr. & others, with contributions from C. D. and many. *Hmisc: Harrell*
511 *Miscellaneous*. (2021).
51247. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag, 2009).
513 doi:10.1007/978-0-387-98141-3.
51448. R Core Team. *R: A language and environment for statistical computing*. *R Foundation for*
515 *Statistical Computing*. (2013).
51649. Conway, J. & Gehlenborg, N. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams*
517 *for Visualizing Intersecting Sets*. (2019).
51850. Baty, F. *et al.* A Toolbox for Nonlinear Regression in R: The Package nlstools. *J. Stat. Softw.*
519 **66**, 1–21 (2015).
52051. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
52152. Auguie, B. & Antonov, A. *gridExtra: Miscellaneous Functions for ‘Grid’ Graphics*. (2017).
52253. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
523 thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
52454. [schrodinger/pymol-open-source](https://www.schrodinger.com/pymol-open-source). (Schrödinger, Inc., 2021).
52555. Multiplexed RT-qPCR to screen for SARS-COV-2 B.1.1.7 variants: Preliminary results - SARS-
526 CoV-2 coronavirus / nCoV-2019 Diagnostics and Vaccines. *Virological*
527 [https://virological.org/t/multiplexed-rt-qpcr-to-screen-for-sars-cov-2-b-1-1-7-variants-preliminary-](https://virological.org/t/multiplexed-rt-qpcr-to-screen-for-sars-cov-2-b-1-1-7-variants-preliminary-results/588)
528 [results/588](https://virological.org/t/multiplexed-rt-qpcr-to-screen-for-sars-cov-2-b-1-1-7-variants-preliminary-results/588) (2021).
52956. Takada, K., Ueda, M. T., Watanabe, T. & Nakagawa, S. Genomic diversity of SARS-CoV-2 can
530 be accelerated by a mutation in the nsp14 gene. *bioRxiv* 2020.12.23.424231 (2020)
531 doi:10.1101/2020.12.23.424231.

532 57. Variants: distribution of cases data. GOV.UK [https://www.gov.uk/government/publications/covid-](https://www.gov.uk/government/publications/covid-19-variants-genomically-confirmed-case-numbers/variants-distribution-of-cases-data)
533 [19-variants-genomically-confirmed-case-numbers/variants-distribution-of-cases-data.](https://www.gov.uk/government/publications/covid-19-variants-genomically-confirmed-case-numbers/variants-distribution-of-cases-data)
534 58. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by
535 a novel set of spike mutations - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology.
536 *Virological* [https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-](https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563)
537 [cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563](https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563) (2020).

538

539 **Supplementary Information**

Fig. S1. Frequency dynamics of SARS-CoV-2 amino acid changes in different regions of Russia. Notations are the same as in Fig.2.

Fig. S2. Logistic growth model for the S:Δ69-70- nsp6:Δ106-108- and S:Δ69-70- nsp6:Δ106-108+ samples in Feb-Apr 2021 based on the 136 samples for which no NGS data was available. Notations are the same as in Fig. 6.

Figure S3. Frequencies of S:Δ69-70, nsp6:Δ106-108, and their combination in Saint Petersburg in Feb-Apr 2021 based on the PCR data.

Fig. S4. Logistic growth model for nsp6:Δ106-108 with and without S:Δ69-70 in Saint Petersburg in Feb-Apr 2021 based on PCR data. Notations are the same as in Fig. 5.