VIRUS EVOLUTION

Virus Evolution, 2022, **8(1)**, 1–11

DOI: https://doi.org/10.1093/ve/veac017

Advance access publication date: 5 March 2022

Research Article

The rise and spread of the SARS-CoV-2 AY.122 lineage in Russia

Galya V. Klink,[1,†] Ksenia R. Safina,[1,2,‡] Elena Nabieva,[2] Nikita Shvyrev,[3] Sofya Garushyants,[1,**] Evgeniia Alekseeva,[2] Andrey B. Komissarov,[4,§] Daria M. Danilenko,[4] Andrei A. Pochtovyi,[5,6] Elizaveta V. Divisenko,[5] Lyudmila A. Vasilchenko,[5] Elena V. Shidlovskaya,[5] Nadezhda A. Kuznetsova,[5] The Coronavirus Russian Genetics Initiative (CoRGI) Consortium, Anna S. Speranskaya,[7] Andrei E. Samoilov,[7,8] Alexey D. Neverov,[7] Anfisa V. Popova,[7] Gennady G. Fedonin,[7,1,9] The CRIE Consortium, Vasiliy G. Akimkin,[7] Dmitry Lioznov,[4,10] Vladimir A. Gushchin,[5,6] Vladimir Shchur,[3] and Georgii A. Bazykin[2,1,*]

[1]Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Bol'shoi Karetnyi per., 19, Moscow 127051, Russia, [2]Skolkovo Institute of Science and Technology (Skoltech), Nobel st., Building 1, Moscow 121205, Russia, [3]International Laboratory of Statistical and Computational Genomics, HSE University, Moscow, Russia, [4]Smorodintsev Research Institute of Influenza, Prof. Popov 15/17, Saint Petersburg 197376, Russia, [5]Federal State Budget Institution 'National Research Centre for Epidemiology and Microbiology Named after Honorary Academician N F Gamaleya' of the Ministry of Health of the Russian Federation, Gamaleya st., 18, Moscow 123098, Russia, [6]Department of Virology, Biological Faculty, Lomonosov Moscow State University, Kolmogorov st., 1, building 73, Moscow 119192, Russia, [7]Central Research Institute for Epidemiology, Novogireyevskaya st., 3a, Moscow 111123, Russia, [8]Saint Petersburg Pasteur Institute, Mira st., 14, Saint Petersburg 197101, Russia, [9]Moscow Institute of Physics and Technology, Institutskiy per., 9, Dolgoprudny, Moscow region 141701, Russia and [10]First Pavlov State Medical University, L'va Tolstogo st., 6-8, Saint Petersburg 197022, Russia

[†]https://orcid.org/0000-0001-8466-6958

[‡]https://orcid.org/0000-0002-5126-9953

[§]https://orcid.org/0000-0003-1733-1255

[**]Present address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA.

*Corresponding author: E-mail: g.bazykin@skoltech.ru

Abstract

Delta has outcompeted most preexisting variants of SARS-CoV-2, becoming the globally predominant lineage by mid-2021. Its subsequent evolution has led to the emergence of multiple sublineages, most of which are well-mixed between countries. By contrast, here we show that nearly the entire Delta epidemic in Russia has probably descended from a single import event, or from multiple closely timed imports from a single poorly sampled geographic location. Indeed, over 90 per cent of Delta samples in Russia are characterized by the nsp2:K81N + ORF7a:P45L pair of mutations which is rare outside Russia, putting them in the AY.122 sublineage. The AY.122 lineage was frequent in Russia among Delta samples from the start, and has not increased in frequency in other countries where it has been observed, suggesting that its high prevalence in Russia has probably resulted from a random founder effect rather than a transmission advantage. The apartness of the genetic composition of the Delta epidemic in Russia makes Russia somewhat unusual, although not exceptional, among other countries.

Key words: SARS-CoV2 in Russia; cross-border transmission of SARS-CoV2; AY.122; Delta lineage; ORF7a:P45L; genomic epidemiology of SARS-CoV2.

1. Introduction

In a pandemic, the global spread of viral lineages is defined by a multitude of factors including the intrinsic properties of the virus, properties of host populations, social factors, and chance. Distinguishing between these factors remains challenging; in particular, it is difficult to spot the lineages with increased fitness against the background of random frequency fluctuations. Since the start of the SARS-CoV2 pandemic, several lineages of concern have appeared and replaced preexisting lineages in different countries (World Health Organization 2021). While some of these variants are certainly characterized by changed fitness due to changes in transmissibility and/or immune avoidance (Davies et al. 2021), much of the geographical difference and dynamics of SARS-CoV-2 lineages is due to epidemiological factors that are not caused by differences in variant fitness (Endo et al. 2020; Lewis 2021; Sun et al. 2021).

The Delta (B.1.617.2 + AY.*) variant of SARS-CoV-2 that was first detected in India in late 2020 (Mlcochova et al. 2021) has remained the prevalent lineage in most countries including Russia till late 2021 (Hodcroft 2021). It was not only shown to be more infectious but also to cause higher mortality than earlier variants of concern (Fisman and Tuite 2021; Li, Lou, and Fan 2021). The fast spread of Delta may be associated with its reduced sensitivity to neutralization by both monoclonal antibodies and antibodies from sera of convalescent patients and immunized people (Planas et al. 2021) as well as increased efficiency of fusion with human cells (Arora et al. 2021b). Delta has spread rapidly in Russia, increasing

in frequency from 1 per cent in April to over 90 per cent in June (Borisova et al. 2021; Knorre et al. 2021).

The phylogeny of Delta is more structured than that of other variants of concern, and its characteristic mutations have accumulated gradually (Stern et al. 2021). While Delta clearly has increased fitness compared to ancestral strains, whether its sublineages change its properties further is less clear (Chadeau-Hyam et al. 2021; UK Health Security Agency 2021; Arora et al. 2021a). Still, the adaptive evolution of SARS-CoV-2 continues (Kistler, Huddleston, and Bedford 2021), highlighting the need for surveillance of novel variants.

Thanks to the extensive efforts of many countries in sampling and sequencing SARS-CoV-2 genomes from patients, it is possible to track the spread of different viral variants across the world. Here, we analyze the emergence and spread of the Delta variant in Russia between April and October 2021 and compare it to other countries. We show that the majority of Russian samples carry the same set of mutations, strongly suggesting that they have descended from a single source.

2. Materials and Methods

2.1 Sample collection and RT-PCR testing

De-identified samples used in this study were collected as part of the ongoing surveillance of SARS-CoV-2 variability routinely conducted at the laboratories of the CoRGI consortium (CoRGI), the Department of Molecular Diagnostic Methods at the Central Research Institute of Epidemiology (CRIE), and the Gamaleya Center. Written informed consent was obtained from all subjects in accordance with the order of the Ministry of Health of the Russian Federation of 21 July 2015 #474 n. This study was reviewed and deemed exempt by the Local Ethics Committee of Smorodintsev Research Institute of Influenza (protocol No. 152, 18 June 2020) and the Local Ethics Committee of the Gamaleya Center (protocol No. 14, 29 September 2021).

Nasopharyngeal and/or throat swabs were collected in virus transport media. Total RNA was extracted using Auto-Pure 96 Nucleic Acid Purification System (Allsheng, China) and NAmagp DNA/ RNA extraction kit (Biolabmix, Russia) for the CoRGI samples, AmpliSens® Cov-Bat-FL reagent kit (AmpliSens, Russia) for the CRIE samples, and QIAamp Viral RNA Mini Kit (Qiagen, Germany) for the Gamaleya samples. Extracted RNA was immediately tested for SARS-CoV-2 using Biolabmix SARS-CoV-2 RT-PCR Detection System (Biolabmix, Russia) based on the Hong Kong University protocol (Chu et al. 2020) for the CoRGI samples, AmpliSens® Cov-Bat-FL reagent kit (AmpliSens, Russia) for the CRIE samples, and a one-step 'SARS-CoV-2 FRT' commercial kit with catalog number EA-128 (obtained from N.F. Gamaleya NRCEM, Russia; one-step RT-qPCR reaction conditions: 50°C for 15 min, 95°C for 5 min, followed by forty-five cycles of 95°C for 10 s and 55°C for 1 min) for the Gamaleya samples. Specimens with Ct values below 30 (CoRGI, Gamaleya) or 25 (CRIE) were selected for whole-genome sequencing.

2.2 Whole-genome sequencing

For CoRGI samples, whole-genome amplification (WGA) of SARS-CoV-2 virus genome was performed using ARTIC Network protocol V3 (https://github.com/joshquick/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V3) with modifications by Itokawa et al. (2020) (Itokawa et al. 2020), using NEBNext® ARTIC SARS-CoV-2 Companion Kit (New England Biolabs, USA). For CRIE samples, WGA was performed using the SCV-2000bp primer panel (Speranskaya et al. 2020) in accordance with the protocol (Kaptelova et al. 2020). For Gamaleya samples, WGA was performed using Itokawa N2 primers (https://github.com/ItokawaK/Alt_nCov2019_primers/tree/master/Primers/ver_N2) before 23 August 2021 and ARTIC V4 primers (https://github.com/joshquick/artic-ncov2019) onward from 24 August 2021. Library preparation was performed with 1D Ligation sequencing kit (SQK-LSK109) with Native barcoding expansion (EXP-NBD196) for Oxford Nanopore sequencing library preparation and with Illumina DNA Prep kit for illumina sequencing for CoRGI samples, with the Q5 High-Fidelity DNA Polymerase (New England BioLabs) for CRIE samples, and with NEBNext Fast DNA Fragmentation & Library Prep Set for Ion Torrent (New England Biolabs, USA) according to the manufacturer's instructions for Gamaleya samples. Finally, sequencing was performed on Oxford Nanopore minION/gridION machines using R9.4.1 flowcells (CoRGI), on MiSeq using MiSeq reagent kit v2 or v3 (CoRGI and CRIE), on Illumina NextSeq 2000 with NextSeq 1000/2000 P2 Reagents (CRIE) or on Ion 540 Chip and Ion S5XL System (Thermo Fisher Scientific, USA) (Gamaley Center).

2.3 Quality control and consensus calling

Raw reads were trimmed with Trimmomatic version 0.39 (Bolger, Lohse, and Usadel 2014) for Illumina sequences and with cutadapt v3.1 (Martin 2011) and vsearch v2.17.0 (Rognes et al. 2016) for ION sequences. The reads were mapped onto the Wuhan-Hu-1 SARS-CoV-2 genome sequence (NCBI ID: MN908947.3) using minimap2 v2.17 (Li 2018) (CoRGI, Nanopore samples), BWA MEM v0.7.17 (Li 2013) (CoRGI, Illumina samples; Gamaley Center), bowtie2 (Langmead and Salzberg 2012) (CRIE). Consensus sequences were built using SAMtools v1.10 (Danecek et al. 2021), Ivar (Grubaugh et al. 2019) and Medaka (https://github.com/nanoporetech/medaka) (CoRGI, Nanopore samples), bcftools v1.9 (Danecek et al. 2021) (CoRGI, Illumina samples), BEDtools (Quinlan and Hall 2010) (CRIE) or FreeBayes v1.3.5 (Garrison and Marth 2012), bcftools v1.12 (Li et al. 2009) and bedtools v2.30.0 (Quinlan and Hall 2010) (Gamaley Center). Generated consensus sequences were deposited to the GISAID database.

2.4 Filtering of sequences

We downloaded a masked alignment of 4,452,413 SARS-CoV2 sequences from GISAID on 21 October 2021 together with accompanying metadata (see Supplementary File 3 for GISAID acknowledgments). We retained sequences characterized as follows: 'Variant' = 'VOC Delta GK/478K.V1 (B.1.617.2+AY.x) first detected in India', 'Host' = 'Human', 'Is complete?' = 'True' and 'Is high coverage?' = 'True'. 1,439 Russian and 1,428,049 non-Russian samples were retained for analysis. The estimated percent of sequenced cases during the study period and the distribution of sequenced samples across the regions of Russia are shown in Supplementary Fig. S1. GISAID IDs of retained Russian sequences are provided in Supplementary File 4.

2.5 UShER phylogenetic tree

We downloaded the public UShER mutation-annotated tree together with associated metadata on 21 September 2021 from the UCSC browser (http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/UShER_SARS-CoV-2/) and extracted the Delta subtree. To avoid duplicate entries, we removed the Russian sequences present in the UShER tree, and then added the Russian GISAID sequences to the tree using UShER (Turakhia et al. 2021), which resulted in a final tree with 28,369 leaves. Branch lengths were corrected using mutation paths obtained by matUtils (McBroome et al. 2021).

2.6 Maximum likelihood phylogenetic trees

We constructed ten datasets, each including all 1,439 Russian Delta sequences and a subset of non-Russian sequences. The number of non-Russian sequences for each country for each time period was chosen on the basis of excess mortality in the corresponding week or month according to https://github.com/dkobak/excess-mortality/blob/main/excess-mortality-timeseries.csv (Karlinsky and Kobak 2021). Specifically, for each country for which data on weekly or monthly excess mortality was available, we picked a minimum of seven sequences per week or thirty sequences per month, plus one additional sequence per fifty excess deaths. If excess mortality did not exceed zero within the time interval, the minimum number of sequences was picked. If fewer than the minimum number of sequences were available, all of them were picked. Each final subset contained 29,964 non-Russian samples. After adding the hCoV-19/Australia/VIC18574/2021 sample of the B.1.617.1 lineage as an outgroup, each dataset consisted of 31,404 samples.

For each dataset, we built a maximum likelihood phylogenetic tree using the FastTreeDbl algorithm of FastTree 2.1.11 (Price, Dehal, and Arkin 2010) with the GTR substitution model and gamma model for heterogeneity of evolutionary rates across sites. We rooted the trees and collapsed branches with less than one mutation (i.e. with length below 0.00003 mutations per site).

2.7 Phylogenetic inference of imports

Imports into Russia were inferred from the phylogenetic distribution of sequences as follows (Supplementary Fig. S2). Samples (tree tips) were marked as Russian (*R*) or non-Russian (*O*) by place of collection. All internal nodes were numbered in order along each lineage from root to tip. Moving from the nodes with the highest numbers toward the lowest (root), each node *N* was labeled according to the labels of its immediate descendants (tips or internal nodes) as follows: (i) if more than one descendant was labeled *R*, *N* was labeled *R*; (ii) if no descendants were labeled *R*, *N* was not labeled; (iii) if exactly one descendant was labeled *R*, the branch leading to this descendant was marked as an import, and *N* was not labeled. As many of the phylogenetic branches are very short and often comprise just one mutation, we found that nucleotide miscalling can result in phylogenetic misplacement of samples and therefore erroneous inference of imports. To focus on the most robust imports, for nested import events, only the deepest import was retained. This procedure resulted in a list of phylogenetically inferred imports (PIIs).

Any procedure for phylogenetic inference of transmission between regions is sensitive to differences in sequencing effort between regions and time periods. Our heuristic method provides a lower-bound number of imports, and is likely biased toward underestimates. In particular, multiple imports of similar genotypes from a poorly sampled location are likely to be inferred as a single PII.

PIIs into other countries were identified analogously. The python script for inference of PIIs is available on GitHub: https://github.com/GalkaKlink/Delta-lineage-in-Russia.

2.8 Estimation of the logistic growth rates

Logistic growth rates of the Delta lineage were estimated with the nls() function of the R language (R version 4.1.0) with initial parameters $r = 0.008$, $x_0 = 0.01$ (R Core Team 2021). For this, Delta frequencies among the Russian samples were averaged across 15 days sliding windows (spanning the 7 days before the current date, the current date, and the 7 days after the current date), and windows with fewer than 20 samples were filtered out. Confidence intervals for estimated model parameters were calculated with confint2() function from nlstools package (Baty et al. 2015).

2.9 Estimation of the effective reproduction number

We used the skyline birth-death model (BDSKY) (Stadler et al. 2013) with continuous sampling, or ψ-sampling, implemented in BEAST2 (Bouckaert et al. 2019) to infer the dynamics of the effective reproduction number R_e. We focused on the monophyletic clade corresponding to the major Delta PII. To tackle sampling heterogeneity, we filtered the major clade in two steps. First, we limited our analysis to the samples collected in Moscow. Second, we subsampled overrepresented dates (see Supplementary Fig. S3) in the Moscow dataset, because it violates the assumptions of the ψ-sampling model, namely, the assumption of continuous routine sampling. Overrepresented dates most likely correspond to additional day-specific sampling events. To remove biases from this overrepresentation, we downsampled the dataset so that it would fit with continuous ψ-sampling using the following procedure. For each date with at least ten samples, we calculated the mean number *N* of samples in a two-week interval (1 week before and 1 week after the date). Then we randomly kept kN samples for this date with $k = 1$ for the baseline analysis (see Supplementary Fig. S3), and additionally with $k = 0.5$ and 2 to check the robustness of our procedure. Our results were not sensitive to k (Fig. S7). Analyses were run for 100 million MCMC steps; convergence was assessed in Tracer (Rambaut et al. 2018). We used the skylinetools package (https://github.com/laduplessis/skylinetools) to set monthly time points for the reproduction number and sampling proportion. All priors were kept default except for those provided in Supplementary Table S1.

Independently, we estimated the R_e dynamics based on case counts using the EpiEstim package (Cori et al. 2013) v2.2-3 in R with 7-days sliding window and parametric serial interval distribution with mean 4.6 and SD 2.0. The dynamics was inferred based on the number of all new cases of SARS-CoV-2 in Moscow (gray line in Fig. 4, retrieved from https://github.com/CSSEGISandData/COVID-19), as well on the estimated number of new Delta infections inferred from the Delta logistic growth curve for Moscow (Fig. 1).

2.10 Estimation of relatedness

To measure the relatedness of samples from the same country, we calculated the mean phylogenetic distance (distance along the phylogenetic tree, $\bar{d}$) between 100 random pairs of samples from this country and compared it with the distribution of phylogenetic distances between 1000 random pairs of samples from any country. We then calculated the number of standard deviations (standard score) between $\bar{d}$ and the mean of this distribution; negative standard score corresponds to increased relatedness of samples from the same country, and positive score, to decreased relatedness. Scripts for phylogenetic clustering estimation are available on GitHub: https://github.com/GalkaKlink/Delta-lineage-in-Russia.

2.11 Visualization

The following R packages were used for visualization: tidyverse (Wickham et al. 2019), ggrepel (Slowikowski 2021), egg (Auguie 2019), stringr (Wickham 2019) and Hmisc (Harrell 2021). Phylogenetic tree was visualized using the ete3 framework (Huerta-Cepas, Serra, and Bork 2016).

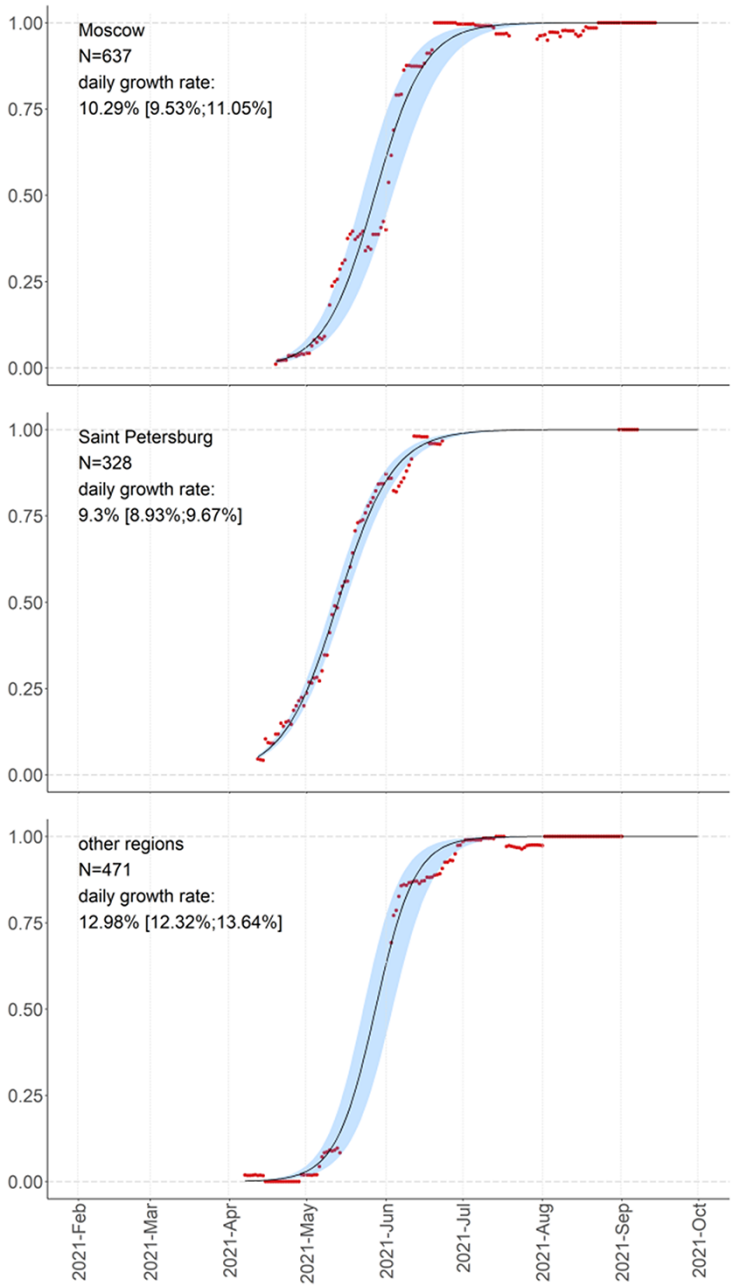

Figure 1. Frequencies of Delta variants (B.1.617.2 + AY.*) in Russia measured for 15-day sliding windows of 7 days around each day, and logistic growth estimates with 95 per cent confidence intervals.

3. Results

3.1 Delta has spread in Russia rapidly in spring 2021

Among the 4,639 high-quality Russian samples that were available in GISAID on 21 October 2021 1,439 are Delta samples, i.e. belong to pango lineage B.1.617.2 or derived lineages (AY.*). The earliest high-quality Delta sample was collected on 7 April 2021 in Moscow; two lower-quality Delta samples date to February 28 and 26 March 2021. Since April, the frequency of Delta among the Russian samples has been growing, reaching 98 per cent by early July 2021, with the estimated daily logistic growth rate of 9.74 per cent (95 per cent CI: 9.28 per cent-10.2 per cent). This growth rate is comparable with that observed in other countries (Chen et al. 2021; Public Health England 2021). The timing of this growth was similar between Russia's regions (Fig. 1).

3.2 Most Russian Delta samples are characterized by the nsp2:K81N + ORF7a:P45L combination of mutations

The vast majority of Russian Delta samples shared the same combination of mutations (Fig. 2, Supplementary Fig. S4). In addition to the mutations characteristic of Delta (Hodcroft 2021), 92.4 per cent of the Delta samples carried the nsp2:K81N (ORF1a:K261N) mutation, and 91.8 per cent carried the ORF7a:P45L mutation. The presence of the nsp2:K81N mutation puts these 92.4 per cent of

Russian Delta samples in the recently designated AY.122 pango lineage. The nsp2:K81N + ORF7a:P45L combination is rare among GISAID Delta samples worldwide (2.3 per cent); outside Russia, its frequency is the highest in Moldova (100 per cent; nine out of nine samples), followed by Ecuador (86 per cent; seventy-six out of eighty-nine samples), Kazakhstan (76 per cent; thirty-two out of forty-two samples) and Latvia (73 per cent; fifty-two out of seventy-one samples).

Outside Russia, the nsp2:K81N and ORF7a:P45L mutations are not strongly linked, and many samples carry the first but not the second (Fig. 2, Supplementary Fig. S4). The ORF7a:P45L mutation has been gained and lost repeatedly according to the global UShER tree. Notably, it is located within one of the ARTIC primers (nCoV-2019_90_RIGHT) binding site, suggesting that the nucleotide at this position may be frequently miscalled. However, in the Russian dataset, we find that the linkage between nsp2:K81N and ORF7a:P45L is nearly perfect, and these mutations co-occur in nearly all samples (Fig. 2, Supplementary Fig. S4).

The earliest nsp2:K81N + ORF7a:P45L sample in Russia dates to 19 April, and it was one of the first Delta samples obtained in Russia. The frequency of the nsp2:K81N + ORF7a:P45L combination has been steadily high between April and October, and it remained the dominant clade throughout this period (Fig. 3A).

Soon after its first detection, the nsp2:K81N + ORF7a:P45L combination has become prevalent throughout Russia (Supplementary Figs S5 and S6). It was detected in all 41 Russian regions where Delta samples were collected. In the 26 regions with more than five samples of Delta, between 62 per cent and 100 per cent of samples carried the nsp2:K81N + ORF7a:P45L combination (Supplementary Table S2).

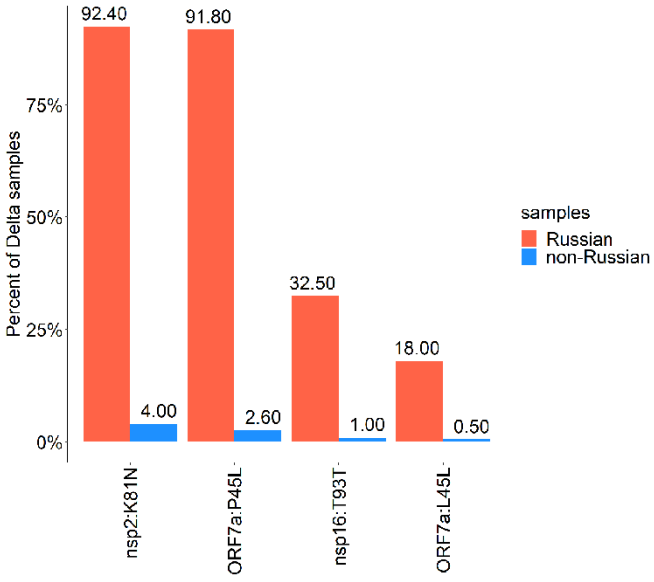

Figure 2. Mutations in the Delta lineage observed in >5 per cent of Russian Delta samples. The following mutations that characterize the major sublineage of B.1.617.2 ('21J' in Nextstrain nomenclature) and occur in >85 per cent of Delta samples both in Russia and globally are not shown: RdRp:G671S, exonuclease:A394V, nsp6:T77A, nsp3:A488S, nsp3:P1228L, nsp6:V120V, ORF7b:T40I, nsp3:P1469S, N:G215C, nsp4:D144D, nsp4:V167L, and nsp4:T492I.

3.3 Just one Delta lineage was successful in Russia although many were imported

To understand how Delta variants were imported into Russia, we used a phylogeographic analysis. Using UShER (Turakhia et al. 2021), we constructed a global phylogeny of SARS-CoV-2 Delta samples including all 1,439 Delta specimens from Russia obtained between 7 April and 29 September 2021. In a maximum parsimony-based approach, we then identified phylogenetically inferred import events (PIIs) as branches in the phylogenetic tree leading to the clades consisting of Russian samples such that their sister clades are non-Russian. For phylogenetically nested PIIs, only the deepest events were considered (Supplementary Fig. S2; see Methods). Our procedure for the detection of PIIs is conservative in that it does not allow repeated imports along the same phylogenetic lineage. It generally yields fewer imports than an alternative approach using the maximum likelihood-based algorithm of TreeTime (Sagulenko, Puller, and Neher 2018). The PIIs matched well the clusters of Russian sequences observed in phylogenies.

Using this procedure, we detected 50 PIIs of the Delta lineage. 24 of these PIIs are represented by a single sequenced Russian sample each, while each of the remaining 26 is represented by multiple Russian samples descending from them. For two early events, the first samples have known travel histories (Fig. 3B). One of them was obtained on 7 April from a person who traveled to the UAE and Turkey, and this was the earliest high-quality Russian sample of the Delta lineage. The other was obtained on 22 April from a person who traveled to India. Both these samples clustered with the Indian samples in the global UShER tree (when placed using the online version of UShER https://genome.ucsc.edu/cgi-bin/hgPhyloPlace).

Strikingly, 91.2 per cent of all samples descended from just a single PII (hereafter referred to as the 'main PII') characterized by the nsp2:K81N + ORF7a:P45L combination of mutations (Fig. 3B, C). While multiple PIIs were of the AY.122 lineage, 1312 of 1328 (98.8 per cent) Russian AY.122 sequences carried ORF7a:P45L and were descendants of the main PII (Fig. 3D). The first sample from the main PII was collected on 19 April 2021 in Moscow. The main PII was among the earliest PIIs of Delta in Russia (Fig. 3B).

When imports were inferred by Treetime as in (Matsvay et al. 2021), we detected 217 imports of the Delta lineage, of which only three imports had more than 50 Russian descendants. These three imports represent the three largest clades included in the main PII.

Phylogenetic inference of imports is sensitive to details of sampling and phylogenetic reconstruction. To estimate the robustness of our estimates, we validated them using an alternative phylogenetic approach. For this, we used the 1,428,049 non-Russian Delta sequences that were available in GISAID on 21 October 2021 after quality filtering (see Methods). We generated ten subsets of 50,000 random non-Russian samples with all 1,439 filtered Russian samples added and reconstructed the maximum likelihood (ML) phylogenetic trees for each such subsample. The inferred number of PIIs differed between replicates, mainly due to low robustness of the smaller PIIs. Nonetheless, in each of the phylogenetic replicates, over 90 per cent of Delta samples were inferred to be descendants of a single PII event (Table 1), similarly to the results obtained with the UShER tree.

3.4 Phylodynamics of the main PII clade

To infer the rate of spread of the largest introduced Delta sublineage, we performed its phylodynamic analysis using BEAST2

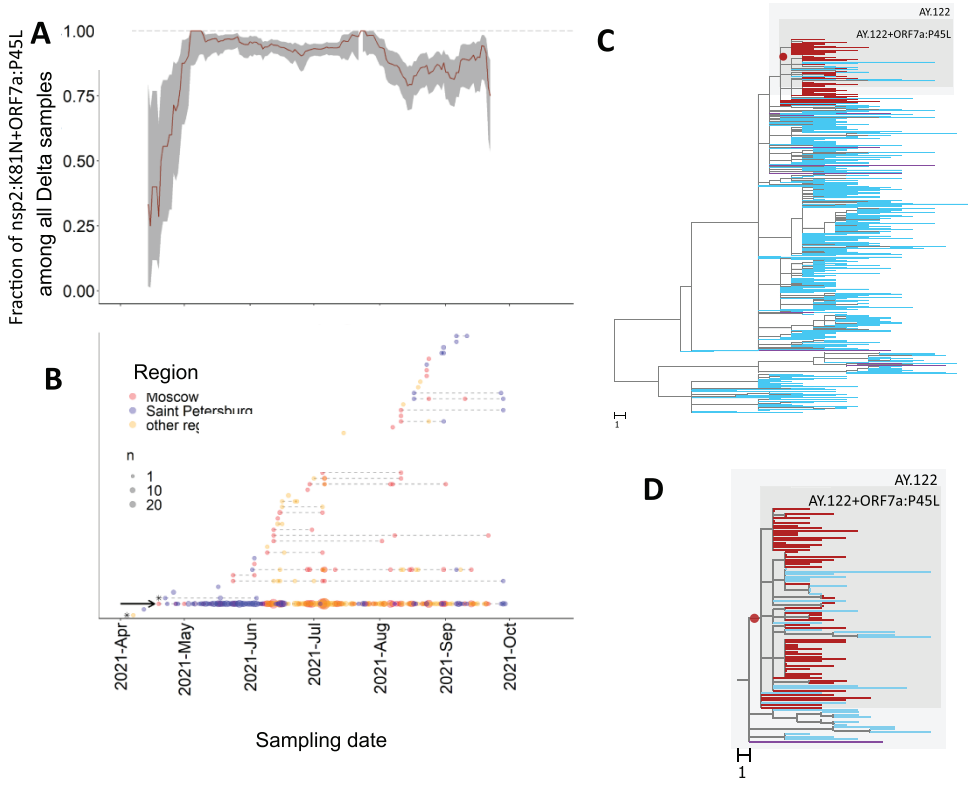

Figure 3. Dynamics of Delta sublineages in Russia. A) The fraction of the nsp2:K81N + ORF7a:P45L combination among all Delta samples from Russia in 15-day sliding window. The confidence band is the 95 per cent binomial confidence interval. B) Timeline for phylogenetically inferred imports (PIIs) of Delta subclades into Russia. Each horizontal line represents a Russian subclade descendant from a single PII, ordered by the date of the earliest sample. Circles represent samples obtained on a particular date; circle size reflects the number of samples; circle color indicates the region of sampling. The AY.122 + ORF7a:P45L sublineage is marked by an arrow. The two PIIs with known travel history for the earliest samples are marked with asterisks. C, D) UShER tree of Delta (C) and its AY.122 + ORF7a:P45L sublineage (D). For visualization purposes, 95 per cent of Russian and 99.8 per cent of non-Russian tips were pruned randomly, so some of the PIIs are not shown. The internal node corresponding to the main PII and which defines the AY.122 + ORF7a:P45L sublineage is marked by a red circle; branches leading to the Russian descendants of the main PII are colored in red; to other Russian sequences, in purple; to non-Russian sequences, in blue; internal branches, in gray. Branch lengths are measured in the number of mutations.

Table 1. PIIs into Russia estimated on ten independent ML phylogenetic trees.

ML tree	Number of PIIs	Russian samples in main PII	% Russian samples in main PII	First PII (earliest sample)	Last PII (earliest sample)	Main PII (earliest sample)
1	15	1328	92	7 April 2021	23 August 2021	19 April 2021
2	13	1397	97	7 April 2021	15 July 2021	12 April 2021
3	8	1399	97	7 April 2021	6 July 2021	12 April 2021
4	11	1397	97	7 April 2021	12 August 2021	12 April 2021
5	13	1397	97	7 April 2021	15 July 2021	12 April 2021
6	2	1436	100	7 April 2021	14 June 2021	7 April 2021
7	1	1439	100	7 April 2021	7 April 2021	7 April 2021
8	13	1397	97	7 April 2021	15 July 2021	12 April 2021
9	10	1396	97	7 April 2021	15 July 2021	12 April 2021
10	9	1397	97	7 April 2021	15 July 2021	12 April 2021

(Bouckaert et al. 2019). COVID-19 has hit Russia's regions differently and nonsynchronously; for example, the timing of epidemic waves has differed among regions (https://xn–80aesfpebagmfblc0a.xn–p1ai/information/). To minimize any effects of geographic structure, for this analysis, we focused on a single region. We considered the 333 samples collected in Moscow, which is the best-sampled of all Russia's regions.

The phylodynamic estimate of R_e of the main PII clade was 1.82 (95 per cent CI [1.49–2.16]) in May and 1.24 (95 per cent CI [1.07–1.41]) in June. In July, it dropped to 0.58 (95 per cent CI [0.40–0.77]), and rose again to 0.99 (95 per cent CI [0.79–1.20]) in August and 1.27 (95 per cent CI [0.62–1.94]) in September, the last month covered by our genetic analysis (Fig. 4).

Overall, this dynamic was consistent with epidemiological data, with increases in R_e preceding rises in case counts, in agreement with case-based R_e estimates inferred by EpiEstim (Fig. 4). Notably, the case counts before June include a large proportion of non-Delta cases; the reduction in number of non-Delta cases may partially explain why the rise in cases in May was slower than that predicted by the R_e. Nevertheless, the high R_e in May and June is consistent with the summer wave which peaked on

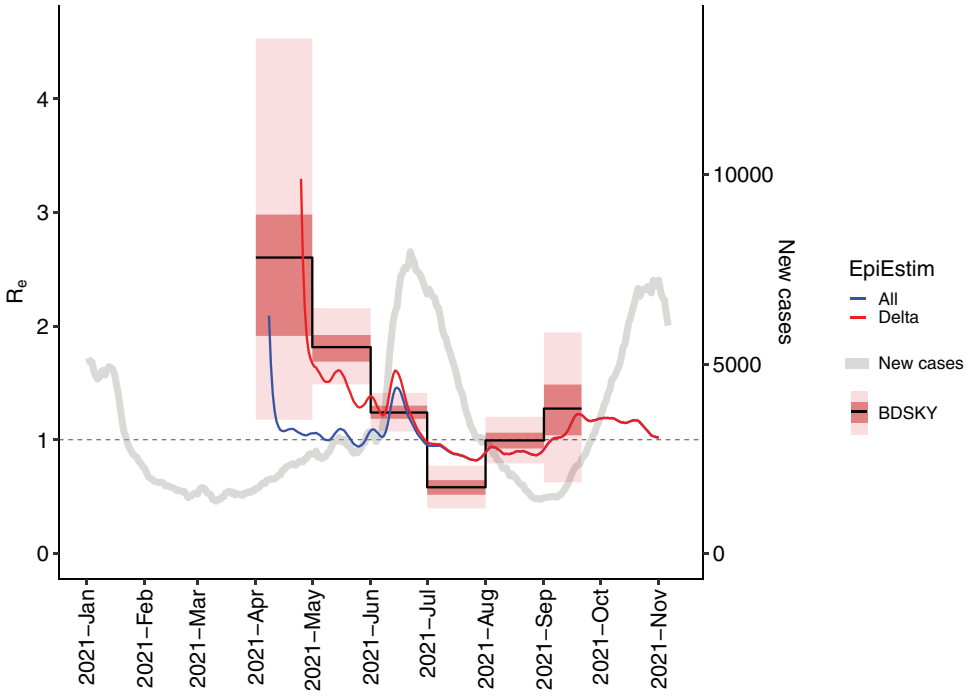

Figure 4. The dynamics of the effective reproduction number R_e for the main PII of the Delta clade in Moscow inferred by BDSKY (black line; shaded red bars show 50 per cent and 95 per cent posterior credible intervals); and for all (blue line) or for Delta (red line) SARS-CoV-2 cases in Moscow inferred by EpiEstim. The gray line shows the 7-day rolling average of the daily number of new cases in Moscow independent of the genotype.

June 25, and the low R_e in July is consistent with the decline in case counts at that time (Fig. 4). This data confirms that the main PII clade (AY.122 + ORF7a:P45L) is responsible for the summer epidemic wave, and most probably for the ongoing autumn wave. The bimodal dynamics is similar to many other countries of the Northern hemisphere, where the advent of summer helped slow down the spread, such as the UK, France, and the USA.

3.5 The success of the nsp2:K81N + ORF7a:P45L combination is probably not due to increased fitness

To explain the success of the nsp2:K81N + ORF7a:P45L combination in Russia, we hypothesized that it could arise from the fitness advantage conferred by these two mutations.

The identity of these mutations does not lend strong support to this hypothesis. nsp2 is a rapidly evolving nonstructural protein that was found to be localized to endosomes and viral replication–transcription complexes. Based on structural analysis and affinity purification mass spectrometry, it is thought to interact with multiple host proteins and mitochondrial RNA, and its suggested functions are the engagement of mitochondria to viral replication sites and modulation of cellular endosomal pathway (Gupta et al. 2021). No signs of either positive or negative selection were found at site 81 of nsp2 (https://observablehq.com/@spond/evolutionary-annotation-of-sars-cov-2-covid-19-genomes-enab?collection=@spond/sars-cov-2) using FEL and MEME algorithms of HyPhy (Kosakovsky Pond et al. 2020).

ORF7a has been shown to suppress BST2 protein that restricts the egress of viral particles from the cell (Martin-Sancho et al. 2021). It was also shown to bind to CD14+ monocytes, which reduces their antigen representation capacity and triggers the production of proinflammatory cytokines (Zhou et al. 2021). Nonsynonymous mutations in ORF7a contribute to SARS-CoV-2 clade success (Kistler, Huddleston, and Bedford 2021). C-terminal truncations of ORF7a are frequent and were shown to affect viral replication (Nemudryi et al. 2021). Nevertheless, a lineage characterized by a frameshifting deletion in ORF7a has spread rapidly in Australia (Foster and Rawlinson 2021). Site ORF7a:45 experiences episodic diversifying selection (according to MEME algorithm of HyPhy) and increase of non-reference amino acid in frequency according to (https://observablehq.com/@spond/evolutionary-annotation-of-sars-cov-2-covid-19-genomes-enab?collection=@spond/sars-cov-2). It has also been predicted to be included in the B-cell epitope (Moody et al. 2021).

Moreover, the dynamics of the nsp2:K81N + ORF7a:P45L combination outside Russia also doesn't support its increased fitness compared to other Delta variants. To show this, we estimated the logistic growth rates of this combination in those countries where it has been frequent (with >15 days with samples carrying this combination both before and after 1 July). While this lineage has been growing in most countries before 1 July (Supplementary Fig. S8), this growth was due to the weakness of competition from non-Delta variants; no systematic growth compared to other Delta lineages was observed (Supplementary Fig. S9). Across countries, the frequency of the nsp2:K81N + ORF7a:P45L combination within a month after its detection was on average higher where it emerged against the background of predominantly non-Delta variants, in line with its increased fitness compared to non-Delta. However, in many countries, nsp2:K81N + ORF7a:P45L did not take off even if it emerged when the frequency of Delta was low (Supplementary Fig. S10). The lack of a systematic fitness advantage of this lineage across the globe compared to other Delta lineages suggests that the selection that favors this variant, if it exists, is weak.

3.6 The genetic homogeneity of Delta in Russia is unusual among other countries

To compare the genetic uniformity of Delta samples observed in Russia to that in other countries, we used the same procedure to obtain a list of PII events for each country with more than 50 Delta sequences in each ML tree. For each country, we then calculated (i) the fraction of Delta samples descendant from the largest PII into this country, and (ii) the extent of relatedness of Delta samples from this country, compared to randomly chosen Delta samples (see Methods).

The contribution of the largest PII was larger in Russia than in most other countries (Fig. 5A). Moreover, while samples from most countries were scattered across the phylogenetic tree, with multiple imports contributing substantially to the local epidemics, Russian samples were unusually related (Fig. 5B). Both these observations also held for the UShER tree that was based on smaller open datasets (Supplementary Fig. S11A for results based on PIIs and Supplementary Fig. S11B for results based on imports inferred with TreeTime) as well as for the 10 ML trees that were built using subsets with all Russian and 50,000 randomly sampled non-Russian samples (Supplementary Fig. S12). The codon for ORF7a:45 is included in a binding region of the ARTIC primer, and therefore may be miscalled, which can potentially affect this result; however, the results remained the same when this codon was masked (Supplementary Fig. S13).

4. Discussion

Previously, we and others have shown that transmission of pre-Delta SARS-CoV-2 variants across Russia's border was rapid (Kozlovskaya et al. 2020; Komissarov et al. 2021). Indeed, the COVID-19 epidemic was started in Russia by a large number of near-simultaneous imports of distinct variants in early spring 2020, and many of these imports resulted in sizable Russian transmission lineages with no single lineage dominating (Komissarov et al. 2021). In subsequent months, imports have continued despite border closure, resulting in thousands of Russian transmission lineages (Matsvay et al. 2021).

By contrast, here we show that the vast majority of Delta SARS-CoV-2 variants that have spread in Russia were genetically similar, carrying the derived nsp2:K81N and ORF7a:P45L changes that are rare outside Russia.

Our ability to distinguish between viral variants resulting from specific imports is limited by the resolution provided by genomic sequences. It is impossible to distinguish between repeated imports of genetically similar or identical variants, and this could lead us to undercount imports. The maximum likelihood-based algorithm of TreeTime (Sagulenko, Puller, and Neher 2018) divides the main Russian PII into multiple import events, supporting the possibility of recurrent imports from the same source (Supplementary Fig. 11B). Moreover, if a country is relatively well sampled, but the regions that are the major sources of introductions into it are relatively poorly sampled, even genetically distinct variants may appear to descend from a single import on a phylogenetic tree (as is likely the case with the UK and the USA, Fig. 5A). Similarly, fewer PIIs into Russia could be identified in ML trees than in the UShER tree, at least partially because these trees include fewer non-Russian samples.

However, our finding of the biased composition of the Russian Delta epidemic does not depend on these concerns. At the time of import of the major Delta lineage into Russia, the global diversity of Delta variants was already high, and we would have been able to identify distinct Delta variants. Indeed, the nsp2:K81N + ORF7a:P45L combination occurs in 68 out of 80 (85 per cent) of Russian samples obtained in April, but just in 34 out of 6658 (0.5 per cent) of non-Russian samples obtained at that time.

The genetic composition of the epidemic in Russia could be less uniform than it seems if there are some unsampled well isolated Russian Regions with a different genomic composition of prevalent variants. However, AY.122 + ORF7a:P45L was the major variant in all 41 Russian regions with more than five Delta samples by the time of this study (Supplementary Table S2).

What can account for this uniformity? There are several options. Conceivably, these mutations could increase viral fitness. Both mutations characterizing the main PII, nsp2:K81N and ORF7a:P45L, are nonsynonymous, making this possibility realistic. However, neither of the two mutations is an obvious candidate for adaptiveness. There is also no evidence that the nsp2:K81N + ORF7a:P45L combination is characterized by an increased rate of spread compared to other Delta variants. While AY.122 has spread rapidly throughout Russia, and at least in Moscow this spread has been driven by a high $R_e > 1$ (Fig. 4), this rapid spread was against the background of non-Delta variants (Fig. 1). The frequency of AY.122 among Delta variants in Russia has been high almost from the start, and its early increase in frequency among Delta samples (Fig. 3A) happened while Delta cases were still low, indicating that it could be random. While it has later been established in numerous other countries in late spring and summer, its frequency has remained modest and has not increased monotonically outside Russia.

Therefore, the high prevalence of the AY.122 + ORF7a:P45L lineage in Russia is probably due to chance. As dispersal of SARS-CoV-2 within countries is more rapid than trans-border transmission, the role of chance in the spread of selectively equivalent variants is high. Indeed, some of the lineages have previously risen in frequency in Russia [e.g. B.1.1.317, (Klink et al. 2021)] and elsewhere [e.g. 'European lineage' EU1, (Hodcroft et al. 2021)] before declining, indicating that the changes defining them did not confer a substantial fitness advantage. Although PIIs of Delta variants into Russia were multiple, even closely dated PIIs differed strikingly by their success (Fig. 3B).

Several factors could contribute to the biased composition of the Delta epidemic in Russia compared to most other countries. First, it could result from a geographic bias in the origin of imports. The earliest samples carrying the nsp2:K81N + ORF7a:P45L combination are sporadic, and were often deposited months later than collected, suggesting that they could be misdated in GISAID (Supplementary Table S4). However, since mid-April, this combination started to appear nearly simultaneously among Delta samples from multiple countries. This suggests that it had originated earlier in a poorly sampled location (Supplementary Table S4). In the second quarter of 2021, the top ten countries with the strongest passenger traffic with Russia were Abkhazia, Ukraine, Turkey, Kazakhstan, UAE, Cyprus, Armenia, Finland, South Ossetia, and Egypt (https://fedstat.ru/indicator/38480). Most of these countries sequence little. Nevertheless, nsp2:K81N + ORF7a:P45L has been observed in four of them (Supplementary Table S4).

Both single and repeated importations of the AY.122 + ORF7a:P45L lineage from the same unsampled location(s) are a possibility. The latter will require the existence of a region with a high prevalence of AY.122 + ORF7a:P45L before mid-April which would be the source of multiple imports, which seems somewhat unlikely because Russia is about equally well connected with multiple countries and there is no reason for just one to play the predominant role. Alternatively, it could

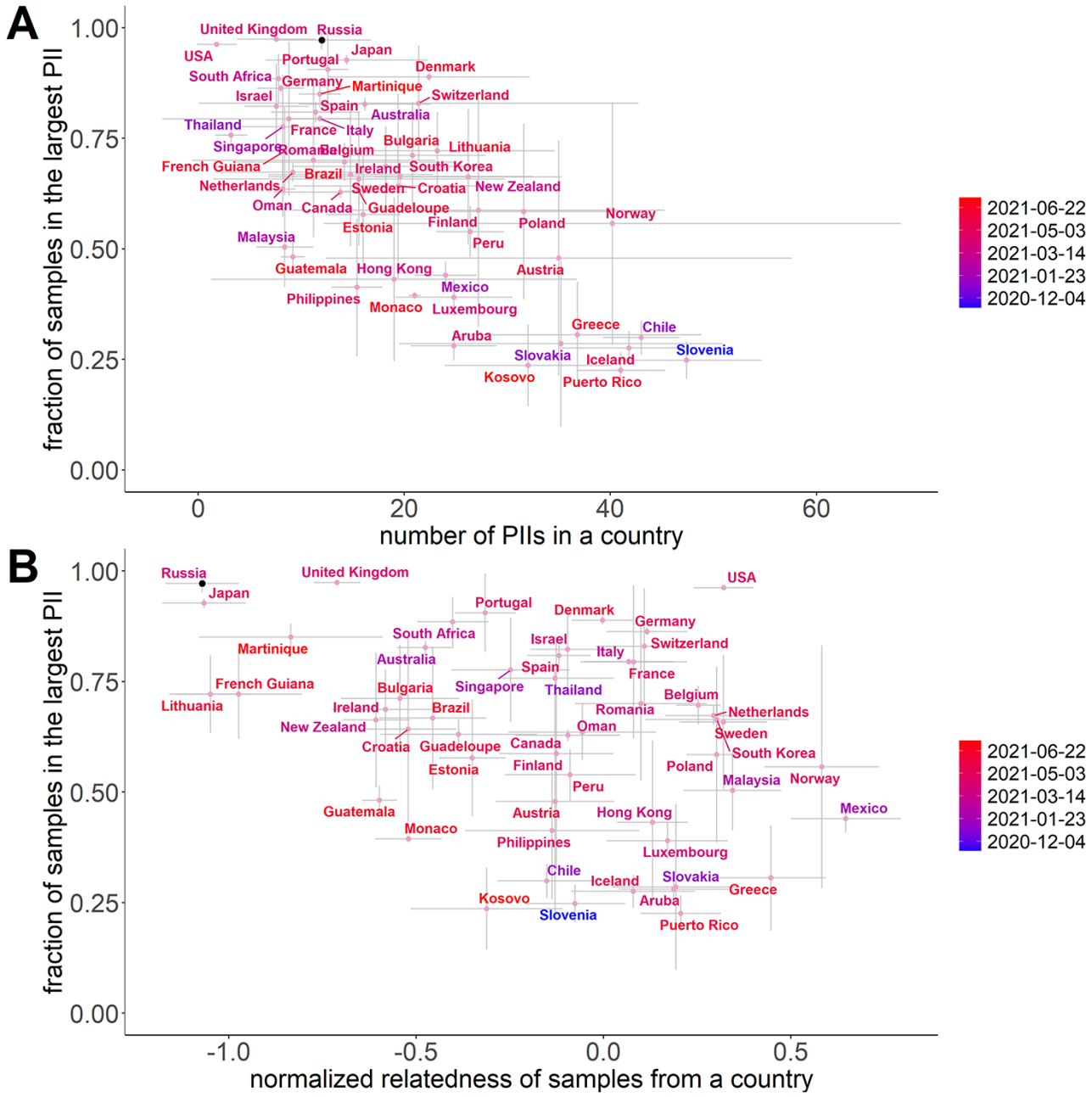

Figure 5. Fraction of Delta samples in the largest PII and (A) inferred number of PIIs or (B) relatedness of Delta samples, for countries with at least 50 Delta samples in each of the 10 ML trials (Table S3). In (B), the horizontal axis indicates the normalized relatedness of samples from the same country, compared with randomly picked samples; lower values correspond to increased relatedness (see Methods). Dots correspond to the mean (centroid) across the 10 ML trees for 29,964 non-Russian samples with added Russian sequences, with standard deviations shown as error bars. Colors indicate the date when the Delta lineage reached 1 per cent frequency in this country.

result from imports from multiple undersampled locations such that AY.122 + ORF7a:P45L has reached a high frequency in all of them—but this also seems less likely in the absence of advantage of AY.122 + ORF7a:P45L compared to other Delta variants. Overall, the hypothesis that the bottleneck has been at the border of Russia, rather than at some other location that exported into Russia, seems more parsimonious.

Second, the size of the trans-border transmission bottleneck could be affected by the measures aimed at limiting passenger traffic. Indeed, the countries with the largest contribution of a single PII to Delta cases (those in the top left in Fig. 5B and Supplementary Fig. S12) include Japan, Australia, and Singapore, some of the countries with the most stringent border policies at that time. In Russia, however, the situation was very different. International traffic through Russian airports was higher in Spring 2021 than in most months of 2020, and rose to pre-pandemic levels by late 2021 (Supplementary Table S5). Delta has been introduced into Russia repeatedly (Fig. 3B), indicating that the homogeneity of the epidemic results from a high variance of reproductive success of imports rather than from their low numbers.

Third, the success of the AY.122 + ORF7a:P45L lineage in Russia could arise from an early superspreading event. Generally, superspreading events have been crucial for SARS-CoV2 spread (Lewis 2021). However, no such event was

reported. The AY.122 + ORF7a:P45L variant started to spread near-simultaneously in Moscow, Saint Petersburg, and the remainder of Russia (Supplementary Table S2), suggesting that if true, this event took place before 19 April in a poorly sampled location within or outside Russia.

Independent of its exact mechanism, the high prevalence of just a single Delta variant in Russia highlights the high role of chance in the local spread of pathogenic lineages. This is in line with the high variance in levels of genetic differentiation (Fst) between countries early in the COVID-19 pandemic, suggesting that outbreaks in most countries could have been started by just a handful of travelers (Ruan et al. 2021). It takes few imports to start an epidemic.

Supplementary Data

Supplementary data is available at *Virus Evolution* online.

Acknowledgements

We thank Russell Corbett-Detig and Yatish Turakhia for invaluable help with UShER and Alexey Kondrashov for valuable discussions. We are grateful to all GISAID submitting and originating labs (Supplementary File 3) for rapid open release of SARS-CoV-2 sequencing data.

Funding

This study was funded by the Russian Science Foundation [grant number 21-74-20160 to GAB]. NS and VS were funded within the framework of the HSE University Basic Research Program.

Conflict of interest: None declared.

References

Arora, P. et al. (2021a) 'Delta Variant (B.1.617.2) Sublineages Do Not Show Increased Neutralization Resistance', *Cellular & Molecular Immunology*, 18: 2557–9.

——— et al. (2021b) 'B.1.617.2 Enters and Fuses Lung Cells with Increased Efficiency and Evades Antibodies Induced by Infection and Vaccination', *Cell Reports*, 37: 109825.

Auguie, B. (2019) Egg: Extensions for "Ggplot2": Custom Geom, Custom Themes, Plot Alignment, Labelled Panels, Symmetric Scales, and Fixed Panel Size. R package version 0.4.5.

Baty, F. et al. (2015) 'A Toolbox for Nonlinear Regression in R: The Package Nlstools', *Journal of Statistical Software*, 66: 51.

Bolger, A. M., Lohse, M., and Usadel, B. (2014) 'Trimmomatic: A Flexible Trimmer for Illumina Sequence Data', *Bioinformatics*, 30: 2114–20.

Borisova, N. I. et al. (2021) 'Monitoring the Spread of the SARS-CoV-2 (Coronaviridae: Coronavirinae: Betacoronavirus; Sarbecovirus) Variants in the Moscow Region Using Targeted High-throughput Sequencing', *Problems of Virology*, 66: 269–78.

Bouckaert, R. et al. (2019) 'BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis', *PLoS Computational Biology*, 15: e1006650.

Chadeau-Hyam, M. et al. (2021) *REACT-1 Round 15 Interim Report: High and Rising Prevalence of SARS-CoV-2 Infection in England from End of September 2021 Followed by a Fall in Late* October *2021*. Epidemiology.

Chen, C. et al. (2021) CoV-Spectrum: Analysis of Globally Shared SARS-CoV-2 Data to Identify and Characterize New Variants. *arXiv:210608106 [q-bio]*.

Chu, D. K. W. et al. (2020) 'Molecular Diagnosis of a Novel Coronavirus (2019-ncov) Causing an Outbreak of Pneumonia', *Clinical Chemistry*, 66: 549–55.

Cori, A. et al. (2013) 'A New Framework and Software to Estimate Time-varying Reproduction Numbers during Epidemics', *American Journal of Epidemiology*, 178: 1505–12.

Danecek, P. et al. (2021) 'Twelve Years of SAMtools and BCFtools', *GigaScience*, 10: giab008.

Davies, N. G. et al. (2021) 'Estimated Transmissibility and Impact of SARS-CoV-2 Lineage B.1.1.7 In England', *Science*, 372: eabg3055.

Endo, A. et al. (2020) 'Estimating the Overdispersion in COVID-19 Transmission Using Outbreak Sizes outside China', *Wellcome Open Research*, 5: 67.

Fisman, D. N., and Tuite, A. R. (2021) 'Evaluation of the Relative Virulence of Novel SARS-CoV-2 Variants: A Retrospective Cohort Study in Ontario, Canada', *Canadian Medical Association Journal*, 193: E1619–25.

Foster, C. S. P., and Rawlinson, W. D. (2021) *Rapid Spread of a SARS-CoV-2 Delta Variant with a Frameshift Deletion in ORF7a*. *medRxiv*. 10.1101/2021.08.18.21262089.

Garrison, E., and Marth, G. (2021) Haplotype-based Variant Detection from Short-read Sequencing. *arXiv:12073907 [q-bio]*.

Grubaugh, N. D. et al. (2019) 'An Amplicon-based Sequencing Framework for Accurately Measuring Intrahost Virus Diversity Using PrimalSeq and iVar', *Genome Biology*, 20: 8.

Gupta, M. et al. (2021) 'CryoEM and AI Reveal a Structure of SARS-CoV-2 Nsp2, a Multifunctional Protein Involved in Key Host Processes', *BioRxiv*. 10.1101/2021.05.10.443524.

Harrell, F. E., Jr. (2021) Hmisc: Harrell Miscellaneous. R package version 4.6-0. <https://CRAN.R-project.org/package=Hmisc>

Hodcroft, E. B. (2021) CoVariants: SARS-CoV-2 Mutations and Variants of Interest. <https://covariants.org/>

——— et al. (2021) 'Spread of a SARS-CoV-2 Variant through Europe in the Summer of 2020', *Nature*, 595: 707–12.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016) 'ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data', *Molecular Biology and Evolution*, 33: 1635–8.

Itokawa, K. et al. (2020) 'Disentangling Primer Interactions Improves SARS-CoV-2 Genome Sequencing by Multiplex Tiling PCR. Kalendar R (Ed.)', *PLoS One*, 15: e0239403.

Kaptelova, V. et al. (2021) Protocol for SCV-2000bp: A Primer Panel for SARS-CoV-2 Full-genome Sequencing V2.

Karlinsky, A., and Kobak, D. (2021) 'Tracking Excess Mortality across Countries during the COVID-19 Pandemic with the World Mortality Dataset', *eLife*, 10: e69336.

Kistler, K. E., Huddleston, J., and Bedford, T. (2021) 'Rapid and Parallel Adaptive Mutations in Spike S1 Drive Clade Success in SARS-CoV-2', *bioRxiv*. 2021.09.11.459844.

Klink, G. V. et al. (2021) *Spread of Endemic SARS-CoV-2 Lineages in Russia*. *medRxiv*, 2021.05.25.21257695.

Knorre, D. D., Nabieva, E., and Garushyants, S. K. (2021) The CoRGI (Coronavirus Russian Genetic Initiative) Consortium, and Bazykin GA. taxameter.ru. <http://taxameter.ru/> accessed 15 Nov 2021.

Komissarov, A. B. et al. (2021) 'Genomic Epidemiology of the Early Stages of the SARS-CoV-2 Outbreak in Russia', *Nature Communications*, 12: 649.

Kosakovsky Pond, S. L. et al. (2020) 'HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. Crandall K (Ed.)', *Molecular Biology and Evolution*, 37: 295–9.

Kozlovskaya, L. et al. (2020) 'Isolation and Phylogenetic Analysis of SARS-CoV-2 Variants Collected in Russia during the COVID-19 Outbreak', *International Journal of Infectious Diseases*, 99: 40–6.

Langmead, B., and Salzberg, S. L. (2012) 'Fast Gapped-read Alignment with Bowtie 2', *Nature Methods*, 9: 357–9.

Lewis, D. (2021) 'Superspreading Drives the COVID Pandemic - and Could Help to Tame It', *Nature*, 590: 544–6.

Li, H. (2013) Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. *arXiv:13033997 [q-bio]*.

——— (2018) 'Minimap2: Pairwise Alignment for Nucleotide Sequences. Birol I (Ed.)', *Bioinformatics*, 34: 3094–100.

——— et al. (2009) 'The Sequence Alignment/Map Format and SAMtools', *Bioinformatics*, 25: 2078–9.

Li, M., Lou, F., and Fan, H. (2021) 'SARS-CoV-2 Variants of Concern Delta: A Great Challenge to Prevention and Control of COVID-19', *Signal Transduction and Targeted Therapy*, 6: 349.

Martin, M. (2011) 'Cutadapt Removes Adapter Sequences from High-throughput Sequencing Reads', *EMBnet.journal*, 17: 10.

Martin-Sancho, L. et al. (2021) 'Functional Landscape of SARS-CoV-2 Cellular Restriction', *Molecular Cell*, 81: 2656–2668.e8.

Matsvay, A. et al. (2021) *Genomic Epidemiology of SARS-CoV-2 in Russia Reveals Recurring Cross-Border Transmission Throughout 2020*. *medRxiv*, 021.03.31.21254115.

McBroome, J. et al. (2021) 'A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees'. In: Lu, J. (ed.) *Molecular Biology and Evolution*, 38: 5819–24.

Mlcochova, P. et al. (2021) 'SARS-CoV-2 B.1.617.2 Delta Variant Replication and Immune Evasion', *Nature*, 599: 114–9.

Moody, R. et al. (2021) 'Predicted B Cell Epitopes Highlight the Potential for COVID-19 to Drive Self-Reactive Immunity', *Frontiers in Bioinformatics*, 1: 709533.

Nemudryi, A. et al. (2021) 'SARS-CoV-2 Genomic Surveillance Identifies Naturally Occurring Truncation of ORF7a that Limits Immune Suppression', *Cell Reports*, 35: 109197.

Planas, D. et al. (2021) 'Reduced Sensitivity of SARS-CoV-2 Variant Delta to Antibody Neutralization', *Nature*, 596: 276–80.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010) 'FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. Poon AFY (Ed.)', *PLoS One*, 5: e9490.

Public Health England. (2021) SARS-CoV-2 Variants of Concern and Variants under Investigation in England. Technical briefing 15.

Quinlan, A. R., and Hall, I. M. (2010) 'BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features', *Bioinformatics*, 26: 841–2.

R Core Team. (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

Rambaut, A. et al. (2018) 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Susko E (Ed.)', *Systematic Biology*, 67: 901–4.

Rognes, T. et al. (2016) 'VSEARCH: A Versatile Open Source Tool for Metagenomics', *PeerJ*, 4: e2584.

Ruan, Y. et al. (2021) 'On the Founder Effect in COVID-19 Outbreaks: How Many Infected Travelers May Have Started Them All?', *National Science Review*, 8: nwaa246.

Sagulenko, P., Puller, V., and Neher, R. A. (2018) 'TreeTime: Maximum-likelihood Phylodynamic Analysis', *Virus Evolution*, 4.

Slowikowski, K. (2021) Ggrepel: Automatically Position Non-Overlapping Text Labels with "ggplot2". R package version 0.9.1. <https://CRAN.R-project.org/package=ggrepel>

Speranskaya, A. et al. (2020) *SCV-2000bp: A Primer Panel for SARS-CoV-2 Full-Genome Sequencing*. *bioRxiv*, 2020.08.04.234880.

Stadler, T. et al. (2013) 'Birth-death Skyline Plot Reveals Temporal Changes of Epidemic Spread in HIV and Hepatitis C Virus (HCV)', *Proceedings of the National Academy of Sciences*, 110: 228–33.

Stern, A. et al. (2021) *The Unique Evolutionary Dynamics of the SARS-CoV-2 Delta Variant*. *medRxiv*, 2021.08.05.21261642.

Sun, K. et al. (2021) 'Transmission Heterogeneities, Kinetics, and Controllability of SARS-CoV-2', *Science*, 371: eabe2424.

Turakhia, Y. et al. (2021) 'Ultrafast Sample Placement on Existing tRees (Usher) Enables Real-time Phylogenetics for the SARS-CoV-2 Pandemic', *Nature Genetics*, 53: 809–16.

UK Health Security Agency. (2021) Technical Briefing 28; SARS-CoV-2 Variants of Concern and Variants under Investigation in England.

Wickham, H. (2019) Stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. <https://CRAN.R-project.org/package=stringr>

——— et al. (2019) 'Welcome to the Tidyverse', *Journal of Open Source Software*, 4: 1686.

World Health Organization (2021) . Tracking SARS-CoV-2 variants. <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/> accessed 15 Nov 2021.

Zhou, Z. et al. (2021) 'Structural Insight Reveals SARS-CoV-2 ORF7a as an Immunomodulating Factor for Human CD14+ Monocytes', *iScience*, 24: 102187.