Supplementary Information

Supplementary Note

To estimate the number of introduction events, we used the following procedure. When a Russian lineage or singleton had no predating Russian sequences in their stem (e.g., Figs. 3a, 4a), we assumed that they originated from distinct introductions. There were 5 such Russian transmission lineages (lineages 1, 2, 3, 5 and 9) which together included 18 sequences; and 33 such singletons, for a total of 38 introduction events.

Additionally, some of the Russian lineages descended from internal nodes with a mix of Russian and non-Russian sequences, and at least some of these stem Russian sequences had earlier collection dates than the earliest dates in the lineage (e.g., Fig. 3b). Similarly, in a fraction of cases, a Russian singleton descended from an internal node with multiple sequences corresponding to it, such that some of the Russian sequences on this node predated the singleton. These cases are referred to as Russian stem-derived transmission lineages and Russian stem-derived singletons, respectively. In these cases, whether the origin of the lineage or singleton corresponded to an introduction event could not be established unambiguously. Finally, each stem cluster could also originate from any number of introductions, ranging between 1 (if all transmissions within it were domestic) and the number of sequences in the cluster (if each sequence was introduced independently) (Supplementary Fig. 2).

To address this, we used the following statistical procedure. We used the fact that for a fraction of samples, direct travel data were available (Supplementary Data 4): we had information on travel abroad or absence of travel history of the sampled individuals. We assumed that these data are reflective of the fraction of sequences in the corresponding category (stem-derived transmission lineages, stem-derived singletons or stem clusters) that were introduced, and that this fraction is reflective of the entire category of samples. For transmission lineages, we assumed that if at least some individuals with travel history were present, this lineage was introduced. Therefore, for each category k, we estimated the number of introductions as $i_k = n_k t_k/(t_k + l_k)$, where n_k is the number of sequenced lineages, or sequenced samples in a non-lineage category; t_k is the number of samples among them with documented travel history; and l_k is the number of samples among them with documented absence of travel history (Supplementary Table 1).

Using this procedure, we estimate that sequences among these three categories result from additional ~1 introduction yielding a transmission lineage (one of the lineages 4, 6, 7 and 8 with predating Russian sequences at the ancestral node); ~6 introductions yielding some of the 40 singletons with Russian sequences at ancestral nodes; and ~22 introductions yielding Russian sequences in stem clusters. Therefore, we estimate the total number of introductions yielding the sampled diversity in Russia as 38+29=67. This number provides a conservative estimate for the number of introductions. It is likely an underestimate; e.g., if

many of the singletons are actually reflective of unsampled Russian transmission lineages, and the index case of these lineages was never sampled, singleton individuals without travel history may still reflect distinct introductions.

Supplementary Data

Supplementary Data 1. Acknowledgement table containing GISAID sequences used in the analysis, available as a separate file Supplementary_Data_1.tsv.

Supplementary Data 2. Sample preparation details including primer sets used, PCR temperature profiles, barcode sequences and .fastq filenames uploaded to SRA, available as a separate file Supplementary_Data_2.xlsx.

Supplementary Data 3. List of sequences excluded from the final dataset, available as a separate file Supplementary_Data_3.xlsx.

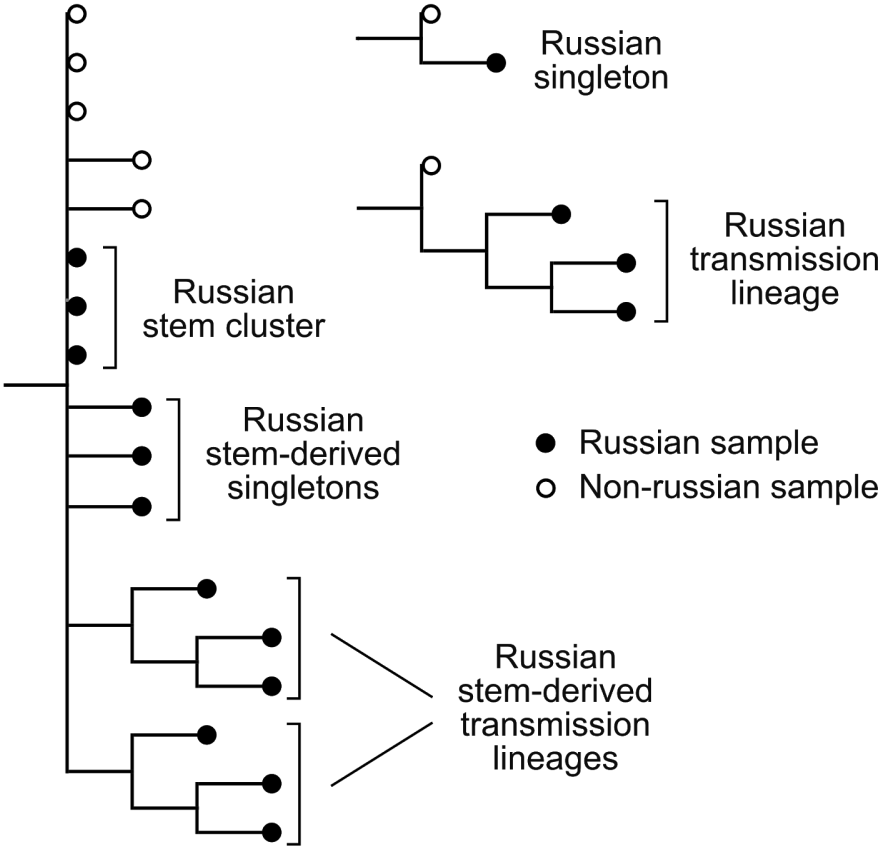
Supplementary Data 4. List of Russian SARS-CoV-2 genomes used in the analysis including samples sequenced in the study (‘Sequenced in this study’ column) and travel history (‘Travel history’ column), available as a separate file Supplementary_Data_4.xlsx.

Supplementary Data 5. Compressed archive containing Supplementary Figs. 1, 3-5, available as a separate file Supplementary_Data_5.zip.

Supplementary Figures

Supplementary Figures 1, 3-5 are available online as Supplementary_Data_5.zip.

Supplementary Figure 1. Phylogenetic tree of all analyzed sequences of SARS-CoV-2 (211 samples from Russia and 19623 from the global dataset). Leaf names are Virus IDs from GISAID. Distances are measured in the number of nucleotide substitutions. Russian samples are marked with grey background and large black dots. Phylogeny was reconstructed as described in Methods.



Supplementary Figure 2. Terminology for phylogenetic groups of samples.

Supplementary Figure 3. Russian transmission lineages and stem-derived transmission lineages. Lineage colors are as in Figs. 2-3 and 5. Notation as in Fig. 3. Stars and diamonds in (A) and (I) indicate samples associated from the Vreden hospital (see Fig. 7).

Supplementary Figure 4. Russian singletons. Notation as in Figs. 3 and 4.

Supplementary Figure 5. Russian stem clusters and stem-derived singletons. Each panel (A-L) shows an independent stem cluster together with its descendent stem-derived singletons. Notation as in Figs. 3. and 4. Triangles in (G) represent samples associated with the Vreden hospital outbreak (see Fig. 7).

Supplementary Tables

Supplementary Table 1. Estimating the number of introduction events giving rise to Russian stem-derived transmission lineages, and Russian stem-derived singletons and Russian stem clusters

	lineages	sequences	travel history		Estimated number of imports
			yes	no	
Russian stem-derived transmission lineages	4	-	1 (33%)	2	=4*0.33=1.3
Russian stem-derived singletons	-	40	1 (14%)	6	=40*0.14=5.6
Russian stem clusters	-	61	4 (36%)	7	=61*0.36=21.96
Total					28.86

Supplementary Table 2. Symptom onset dates for the 11 sequences for which these data are available. Green color is for the sequences collected on April 7; blue, on April 10; and orange, on April 14. Darker colors show sequences for which the symptom onset date differs from the collection date.

Sample id	Symptoms onset date	Collection date
4723	05.04.2020	07.04.2020
4724	05.04.2020	07.04.2020
4726	07.04.2020	07.04.2020
4728	04.04.2020	07.04.2020
4983	10.04.2020	10.04.2020
4984	10.04.2020	10.04.2020
4985	10.04.2020	10.04.2020
4988	09.04.2020	10.04.2020
5643	11.04.2020	14.04.2020
5644	14.04.2020	14.04.2020
5654	13.04.2020	14.04.2020

Supplementary Table 3. Samples per date according to collection dates. All the sequences from April 3, 7, 10 and 14 are from group 1. The sequences from April 22 belong to different groups, in particular: 3 sequences from group 1, 7 sequences from group 2 and 4 sequences from group 3.

Date	April 3	April 7	April 10	April 14	April 22
Collection dates	3	17	11	7	14

Supplementary Tables 4, 5 and 6 contain the Bayesian estimates of the model parameters for the three datasets comprising groups 1, 2 and 3 (Table 4), groups 1 and 2 (Table 5) and group 1 (Table 6). The estimates of effective reproductive numbers and sampling proportions are consistent throughout all the runs. The tree height corresponds to the dating of the root. Group 1 is suspected to correspond to the first introduction event, so its root corresponds to the suspected beginning of the outbreak. The dating of the root for the two other datasets provide evidence for multiple introductions. We used Tracer [1] to summarise the results.

Supplementary Table 4. Phylodynamic parameter estimates for groups 1, 2 and 3. The parameter estimates obtained using BEAST2 with the birth-death skyline model.

Parameter	Estimate	95% confidence interval
TMRCA date	February 4	January 1 - March 7
reproductiveNumber1	0.917	[0.5978, 1.1625]
reproductiveNumber2	3.722	[2.4837, 5.046]
reproductiveNumber3	1.378	[0.4826, 2.4059]
samplingProportion1	0.0 (fixed)	--
samplingProportion2	0.788	[0.4606, 1]
samplingProportion3	0.104	[0.007, 0.2485]
samplingProportion4	0.0148	[4.77E-8, 0.0528]
clockRate	9.427E-4	[8.5E-4; 1.04E-3]

Supplementary Table 5. Phylodynamic parameter estimates for groups 1 and 2. The parameter estimates obtained using BEAST2 with the birth-death skyline model.

Parameter	Estimate	95% confidence interval
TMRCA date	March 15	February 25 - March 31
reproductiveNumber1	1.115	[0.462, 2.4224]
reproductiveNumber2	3.961	[2.5173, 5.4878]
reproductiveNumber3	1.3	[0.472, 2.26]
samplingProportion1	0.0 (fixed)	--
samplingProportion2	0.808	[0.5009, 1]
samplingProportion3	0.149	[0.0153, 0.3461]
samplingProportion4	0.017	[1.4E-7, 0.0598]
clockRate	9.403E-4	[8.4E-4; 1.04E-3]

Supplementary Table 6. Phylodynamic parameter estimates for group 1. The parameter estimates obtained using BEAST2 with the birth-death skyline model.

Parameter	Estimate	95% confidence interval
TMRCA date	March 23	March 11 - March 30
reproductiveNumber1	1.284	[0.4, 3.1]
reproductiveNumber2	4.02	[2.5, 5.7]
reproductiveNumber3	1.3	[0.5, 2.3]
samplingProportion1	0.0 (fixed)	--
samplingProportion2	0.8	[0.5, 1]
samplingProportion3	0.149	[0.02, 0.3]
samplingProportion4	0.02	[3.3E-4, 0.05]
clockRate	9.376E-4	[8.4E-4; 1.03E-3]

Supplementary Table 7. Priors used in the analysis under the birth-death skyline model. The clockRate prior was used according to the posterior estimates from the UK analysis [2]. Other priors are the same or similar to those used in the birth-death skyline analysis in [3].

Model	Parameter	Prior distribution
HKY	Gamma shape	Exponential(0.5)
	Kappa	Log Normal(1.0, 1.25)
Strict clock	Clock rate (per bp per year)	Normal(9.41e-4, 4.99e-5)*
Birth Death Skyline Serial	Effective reproductive number	Log Normal(0.8, 0.5)
	Date of infection origin	Uniform(0, 1000)
	Sampling proportion	Uniform(0, 1)
	Become uninfectious rate	36.5 (fixed)

References

1. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology* vol. 67 901–904 (2018).
2. OliverPybus, Kristian_Andersen, ljones & gbellobr. Preliminary analysis of SARS-CoV-2 importation & establishment of UK transmission lineages. *Virological* https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507/2 (2020).
3. Stadler, T. Phylodynamic Analyses of outbreaks in China, Italy, Washington State (USA), and the Diamond Princess. *Virological* https://virological.org/t/phylodynamic-analyses-of-outbreaks-in-china-italy-washington-state-usa-and-the-diamond-princess/439 (2020).