Skolkovo Institute of Science and Technology

MOLECULAR EPIDEMIOLOGY OF SOCIALLY IMPORTANT INFECTIOUS DISEASES

Doctoral Thesis

by

KSENIIA SAFINA

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisor

Professor Georgii Bazykin

Moscow — 2021

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgement is made, and has not been submitted for any other degree.

Candidate (Kseniia Safina)
Supervisor (Prof. Georgii Bazykin)

ABSTRACT

Infectious diseases remain an unfortunate yet inherent part of human existence. More than a thousand pathogens are capable of causing infections in humans; pathogens differ in terms of their transmissibility, disease severity, treatability, geographic and demographic distribution, and other aspects of their biology and epidemiology, posing various extent of threat to public health.

The field of epidemiology studies infectious diseases through a combination of observational, experimental, and theoretical approaches; it investigates causes and factors orchestrating the spread of infection and informs decisions of healthcare departments on control and preventive measures to tackle current outbreaks and prevent future ones. Incorporation of molecular data into existing epidemiological frameworks, coined as molecular epidemiology, offers a novel dimension in understanding pathogen dynamics: the underlying structure of an outbreak, together with evolutionary and epidemiological processes shaping it, can now be explored. In this thesis, I demonstrate how methods of molecular epidemiology can be used to study two human infectious diseases.

In Chapter 3, molecular epidemiology is applied to a densely sampled HIV-1 dataset. We aimed to characterize the HIV-1 epidemic in Oryol Oblast, a Russian geographic region with a relatively small HIV-positive population, and collected viral genetic data covering at least a third of the known part of the epidemic in Oryol Oblast. We identify multiple introductions of HIV-1 into the region and describe them in terms of their date of introduction, the rate of growth represented by the reproduction number, and the composition and association with various categories such as HIV-1 subtype, transmission route, and sex. To our knowledge, this study is the most detailed description of the HIV-1 epidemic conducted among Russian regions.

In Chapter 4, we analyze genetic data on SARS-CoV-2 coming from the beginning of the first epidemic wave of COVID-19 in Russia in March-April 2020. We identify and describe multiple instances of viral introduction into Russia from abroad at the dawn of the global SARS-CoV-2 pandemic. Combining the phylogeographic approach with travel data, we attempt to estimate the total number of introduction events and their geographic sources. Using Bayesian phylogenetics, we unravel the dynamics of an outbreak at the Vreden hospital in Saint Petersburg that took place in the spring of 2020.

The results presented in this thesis prove the utility of molecular epidemiology methods in revealing patterns of viral transmission for the two epidemics that affect Russia.

Keywords: molecular epidemiology, HIV-1, SARS-CoV-2, infectious diseases in Russia

PUBLICATIONS

1. Komissarov, A.B., Safina, K.R., Garushyants, S.K. et al. Genomic epidemiology of the early stages of the SARS-CoV-2 outbreak in Russia. Nat Commun 12, 649 (2021). https://doi.org/10.1038/s41467-020-20880-z

2. Anfisa V. Popova, **Ksenia R. Safina**, Vasily V. Ptushenko, Anastasia V. Stolyarova, Alexander V. Favorov, Alexey D. Neverov, Georgii A. Bazykin. Allele-specific nonstationarity in evolution of influenza A virus surface proteins. Proceedings of the National Academy of Sciences Oct 2019, 116 (42) 21104-21112; https://doi.org/10.1073/pnas.1904246116

3. Olga A Kudryavtseva, **Ksenia R Safina**, Olga A Vakhrusheva, Maria D Logacheva, Aleksey A Penin, Tatiana V Neretina, Viktoria N Moskalenko, Elena S Glagoleva, Georgii A Bazykin, Alexey S Kondrashov, Genetics of Adaptation of the Ascomycetous Fungus *Podospora anserina* to Submerged Cultivation, Genome Biology and Evolution, Volume 11, Issue 10, October 2019, Pages 2807–2817, https://doi.org/10.1093/gbe/evz194

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Georgii Bazykin for his patient guidance and invaluable experience which he's been sharing with me for the last seven years.

Next, I would like to express my gratitude to the people who contributed a lot to both projects of this thesis.

The work on the early dynamics of SARS-CoV-2 in Russia would have been impossible without our collaborators from Smorodintsev Research Institute of Influenza in Saint Petersburg who provided the study with two-thirds of the analyzed dataset. I would like to thank Olga Shneider from the Vreden Institute who collected patient samples and epidemiological data from the hospital making possible the analysis of the Vreden outbreak. I would like to thank Sofya Garushyants for the thorough analysis of SARS-CoV-2 imports, geographic and travel data, and GISAID metadata. I would like to thank colleagues from the Higher School of Economics who performed the birth-death skyline analysis of the Vreden hospital outbreak. I am grateful to all the authors of the SARS-CoV-2 paper for contributing to the manuscript text.

Many people have contributed to the HIV-1 project as well. I owe Dmitry Kireev a debt of gratitude for making the HIV-1 project possible. Dmitry inspired this work; he also communicated with the AIDS center in Oryol Oblast, arranged transportation and sequencing of samples, and patiently consulted me on various topics throughout the project. I would like to thank our collaborators from the AIDS center in Oryol who provided us with viral samples and the relevant patient and epidemiological data, and all patients who contributed to this study. I would also like to thank Nicola Mueller for consulting me on the multi-tree model he recently implemented and for additionally implementing a more convenient class for the prior

on sampling proportion. I thank Georgii Bazykin and Dmitry Kireev for contributing to the preprint text.

I am eternally grateful to my dear colleagues and friends who supported me in so many ways during this work and beyond.

Finally, I would like to thank my parents for letting me leave to study in Moscow ten years ago and for accepting me doing things that seem obscure to them.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AIDS — acquired immunodeficiency syndrome

ART — antiretroviral therapy

BD model — birth-death model

CD4 — cluster of differentiation 4

CDC — Centers for Disease Control and Prevention

COVID-19 — coronavirus disease 2019

CTL — cytotoxic T lymphocyte

DNA — deoxyribonucleic acid

HET — heterosexual transmission

HIV — human immunodeficiency virus

IDU — injecting drug users

LCA — last common ancestor

MRCA — the most recent common ancestor

MSM — men who have sex with men

NPI — non-pharmaceutical intervention

R_e — effective reproductive number

RNA — ribonucleic acid

RT — reverse transcriptase

SARS-CoV-2 — severe acute respiratory syndrome coronavirus 2

SIR model — susceptible-infected-recovered model

SIV — simian immunodeficiency virus

CHAPTER 1: INTRODUCTION

Diseases affect our lives. A distinct proportion of the harm comes from infectious diseases which are typically caused by viruses, bacteria, fungi, or protozoa. Infectious diseases can be transmitted between people through direct or non-direct contact and occasionally may result in outbreaks given the transmission rate is high enough. Thus, investigating and monitoring infectious diseases is crucial for public health.

The field of epidemiology uses various methods to address these challenges; one of them utilizes molecular data, setting a separate section of epidemiology called molecular epidemiology. The simplest example of how molecular data can be used to study an infectious outbreak is identifying how pathogen samples collected during the outbreak are related; this can complement an incomplete history of contacts and shed light on the dynamics of pathogen transmission. In the early days of molecular biology, various experimentally determined biomarkers were used, for instance, endonuclease restriction patterns.

As sequencing technologies evolved, it soon became apparent that nucleotide sequences provide greater resolution compared to previously used biomarkers; additionally, the growing amount of sequencing data propelled the development of computational methods in bioinformatics and evolutionary biology. These changes have significantly advanced the field of molecular epidemiology. Over the last 20 years, multiple studies have applied molecular epidemiology to reveal and understand patterns of pathogen transmission and evolution.

Sequences of pathogen samples can be used to construct viral phylogenies. From the shape of the phylogeny and the distribution of various properties of samples across the phylogeny, various inferences can be made about the dynamics and mode of pathogen

transmission. However, the range of possible analyses is dependent on characteristics of available data like sampling coverage and representativeness of the distribution of samples across time and categories. The two pathogens studied in this thesis project, HIV-1 and SARS-CoV-2, differ in this respect. By means of this thesis project, I would like to illustrate how methods of molecular epidemiology can be applied to study these two pathogens in the context of the Russian population.

The first part of this thesis is focused on HIV-1. HIV-1 is a virus that causes long-term infections in humans gradually impairing the adaptive immunity and progressing to AIDS (acquired immunodeficiency syndrome) if untreated. First widely acknowledged in the 1980s, HIV-1 has been causing epidemics in multiple areas. The policies handling epidemics differ across countries. In Russia, the prevalence of HIV-1 is one of the highest among European countries. At the end of 2019, the official HIV-1 prevalence in Russia was 0.76% with 1,068,839 HIV-positive people; the true values should be higher due to people being unaware of their HIV-positive status. Molecular epidemiology has limited applicability in the context of the country-wide HIV-1 epidemic in Russia due to the small amount of genetic data available (population coverage <1%). Instead of describing the Russian-wide epidemiology of HIV-1 using sparse data, we decided to focus on a single Russian region and study in detail one HIV-1 sub-epidemic.

We selected the HIV-1 epidemic in Oryol Oblast, represented by 2,157 registered HIV-positive people; we obtained sequences of the *pol* region fragment of the HIV-1 genome from 768 patients covering more than a third of the known part of the epidemic. This part of the thesis illustrates how substantial sampling density allowed us to describe the structure of viral lineages circulating in Oryol Oblast with a good resolution.

The second part of the thesis is focused on another Russian epidemic caused by a virus very different from HIV-1. Here, I describe our work that explored the early dynamics of SARS-CoV-2 in Russia. We studied the onset of the Russian epidemic using 211 complete SARS-CoV-2 genomes collected in March and April of 2020.

This work is different from the HIV-1 project both in terms of infection properties and sampling strategy. First, unlike HIV-1, SARS-CoV-2 is being easily transmitted through contacts with infectious respiratory fluids, possesses much shorter generation time and recovery rate, and evolves slower. Second, we were only able to collect and analyse viral samples from a small and biased fraction of the infected population. These differences affected our choice of methodology and topics studied.

Despite a much lower population coverage (0.3%) and a limited number of samples, the available data were informative of the early stages of the SARS-CoV-2 epidemic, although some of our findings are sensitive to undersampling as acknowledged. The sampling collection procedure was biased towards infections observed in Saint Petersburg; in particular, about a quarter of the final dataset corresponded to the SARS-CoV-2 outbreak in a hospital in Saint Petersburg that was placed under quarantine in April 2020 together with ~700 people inside. This allowed us to apply phylodynamics to study this outbreak.

Two major goals were set during this research, both involving the analysis of genetic data using methods of molecular epidemiology:

1. To study the HIV-1 epidemic in Oryol Oblast using the *pol* region fragment of HIV-1 collected from at least a third of the infected population, to identify importations of the virus into the region and cases of viral transmission within the region (transmission lineages), to describe the composition of subtypes, sexes, and transmission routes, and the association

between these categories and transmission lineages, and to estimate the growth rate of the epidemic.

2. To study the beginning of the COVID-19 epidemic in Russia using 211 complete SARS-CoV-2 genomes, to identify cases of imports and domestic transmission (transmission lineages), and to estimate the total number of imports and suggest their geography based on the travel data available.

The thesis is structured as follows. In Chapter 2, I provide an overview of the relevant body of literature, first introducing the concept of molecular epidemiology and its commonly used methods, then reviewing the two objects of this Ph.D. project, HIV-1 and SARS-CoV-2, and putting them in the context of molecular epidemiology. In Chapter 3, I describe the first part of my thesis project on the analysis of a densely sampled HIV-1 epidemic in a particular region of Russia. Chapter 4 investigates the early dynamics of SARS-CoV-2 in Russia based on genomic data collected in March-April, 2020, making for the second part of this project. Chapter 5 concludes the thesis by summarizing the obtained results.

I would like to state my contributions to this work here. In Chapter 3, I did the bioinformatic part of the project, starting from the development of a custom consensus calling pipeline and the preparation of the dataset. In Chapter 4, I validated and corrected SARS-CoV-2 consensus sequences initially made by Artem Fadeev from Smorodintsev Research Institute of Influenza in Saint Petersburg; prepared the dataset, reconstructed the phylogeny, and inferred transmission lineages; inferred the dynamics of R_e in EpiEstim.

CHAPTER 2: LITERATURE REVIEW

This chapter is organized as follows. In Section 2.1, I discuss briefly the molecular epidemiology field. In Sections 2.2 and 2.3, I introduce the objects studied in this Thesis, HIV-1 and SARS-CoV-2 respectively, and discuss how molecular epidemiology can add a novel piece of knowledge about these two infectious diseases.

2.1. Molecular epidemiology offers a novel dimension to our understanding of diseases bringing together epidemiological and biological data

Diseases exist, and some of them affect populations heavily. What factors contribute to disease transmission, progression, and spread across the host population? How can we affect those factors to intervene in the current outbreak and to prevent future epidemics? Epidemiology is a field that studies these questions by collecting and analyzing incidence data combined with patient data (such as age, sex, ethnicity, immune status, chronic diseases, contact with disease carriers, etc.) using methods ranging from routine surveillance and simple association studies to complex mathematical models. For most diseases, incidence and patient data can be accompanied by biological markers which can be measured in the laboratory.

Remark: Epidemiology studies both infectious (e.g. HIV, HCV) and non-infectious (e.g. cardiovascular disease, cancer) diseases, however, considering the topic of this thesis, the text below focuses on the epidemiology of infectious diseases only.

2.1.1. The dawn of molecular epidemiology

One can define molecular epidemiology as a field that aims to describe and study epidemiological processes using various biomarkers. Since its early history, molecular

biology has been lending techniques that produced data suitable for an epidemiological framework, including antibiotic resistance profiles [1], seroreactivity [2], plaque morphology [3], electrophoretic mobility of polypeptides and RNAs [4,5], and endonuclease restriction patterns [6–8], with the latter probably being the most widespread among all.

DNA fingerprinting of *Mycobacterium tuberculosis* strains based on a repetitive DNA element called IS6110 serves as a prominent example of the restriction-based approach. IS6110-based typing has been used for decades to describe and distinguish *M. tuberculosis* strains before DNA sequencing came into play. IS6110 is a repetitive insertion sequence scattered across the bacterial genome. After bacterial DNA is treated with a particular restrictase, the number and location of IS6110 elements, together with polymorphisms at the restrictase sites, produce a certain electrophoretic pattern when probed for an IS6110-specific sequence [9]. These patterns, or fingerprints, have been widely used to study tuberculosis outbreaks [10,11], recent transmissions and clustering in the population [12,13], geographic dispersal [10], and to infer serial interval and incubation period [14]. Restriction-based techniques have been successfully used for many other human pathogens including herpes simplex virus [15], *Escherichia coli* [16], *Staphylococcus aureus* [17], and *Helicobacter pylori* [18]. The rapid development of sequencing technologies provided an unprecedented resolution of pathogen genotypes and made apparent the limitations of restriction-based methods in revealing the true amount of genetic diversity in compared samples [19,20]. Nowadays, restriction-based methods, as well as other DNA fingerprinting techniques, are mainly used for typing in preliminary diagnostics or when DNA sequencing is not available.

2.1.2. Modern molecular epidemiology

The increasing availability of genetic sequences has allowed scientists to incorporate phylogenetic methods into the statistical framework of epidemiology. Evolutionary, epidemiological, and immunological forces affect genetic diversity dynamics of pathogen populations shaping their phylogenies. In this section, I provide a brief description of some of the commonly used ideas and methods.

2.1.2.1. Phylogeny shape

Among the classical examples of phylogeny-shaping processes is the host immune response. Strong immune selection can enforce a **"ladder"-like asymmetrical tree shape**, a pattern well-described for the population phylogeny of influenza A virus [21,22] which is bound to generate novel mutations at epitopes and increase its antigenic distance in order to evade herd immunity [23,24]; this process can be traced by comparing patterns of mutations on the so-called "trunk" of the tree, that captures the history of the most successful strains, with side branches of the tree [25]. A "ladder"-like pattern has been also reported for intra-host HIV evolution mediated by constant CD8+ T cells and neutralizing antibodies pressure [26–29]. In contrast, the population tree of HIV follows a balanced "star"-like pattern [30], in agreement with no clear signal of cross-immunity and differential fitness being observed for HIV at the level of population. It should be noted, however, that similar patterns might result from different processes or their combinations (e.g. regular population bottlenecks will result in a "ladder"-like topology similarly to immune pressure) thus additional statistical tests should be performed before making inferences.

2.1.2.2. Coalescent theory

Another phylogenetic pattern that can inform us about genetic diversity dynamics is the distribution of branch lengths along the tree which can be mathematically described by

coalescent theory [31]. Coalescent theory in its basic form works with a neutrally evolving, non-recombining, randomly mating population of constant size N_e with non-overlapping generations. N_e is the effective population size that represents the number of individuals contributing to progeny rather than the total number of individuals in the population and reflects the amount of genetic diversity and genetic drift [30,32]. Coalescent theory describes how lineages (alleles sampled from a population) coalesce, or merge, to their common ancestor (Fig 2.1).

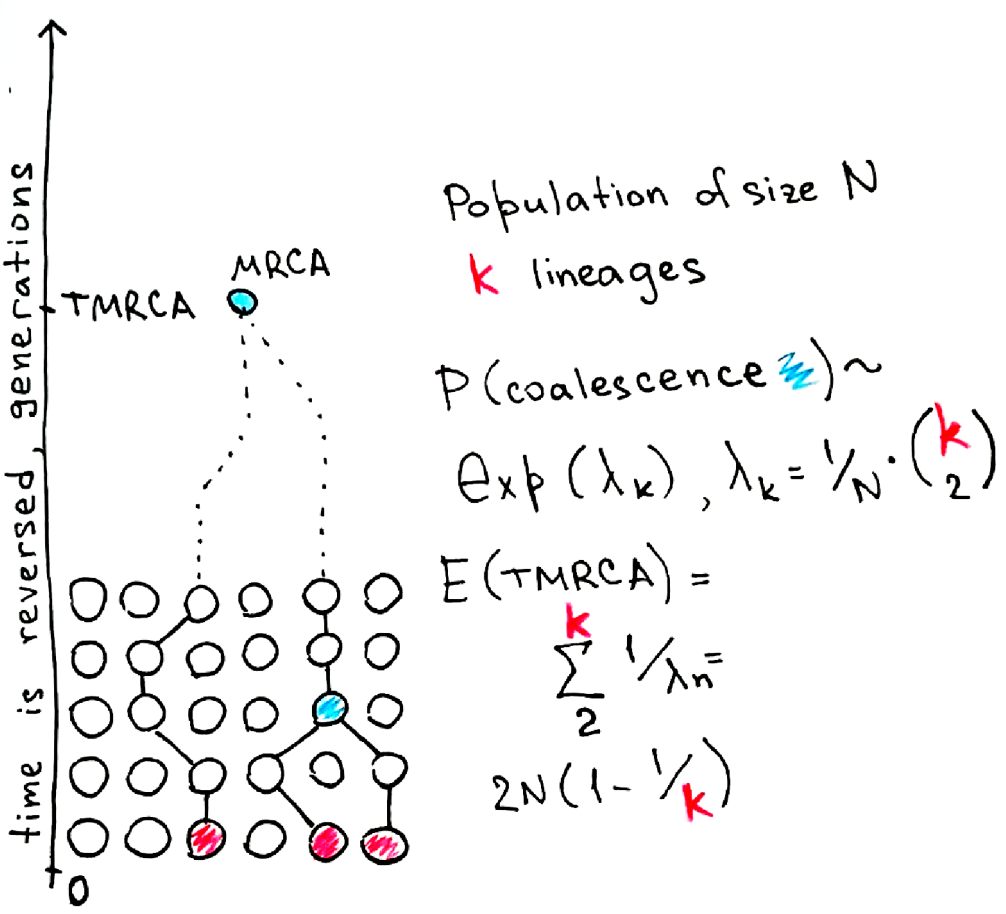


Figure 2.1. Coalescent process. Red circles correspond to the observed lineages; blue circles mark coalescent events.

Under assumptions of the model, the probability of any two lineages in a sample of k lineages to coalesce follows a geometrical (exponential, assuming N_e is large enough for

continuous approximation) distribution with mean $\frac{k(k-1)}{2N_e}$, implying that the waiting time for a coalescent event is $\frac{2N_e}{k(k-1)}$ generations. It can be easily demonstrated that the time to the most recent common ancestor (MRCA) of all *k* lineages equals $2N_e\left(1-\frac{1}{k}\right)$ and tends to $2N_e$ as long as *k* is large enough.

The equations above inform us about the expected branching pattern: waiting times increase as the number of considered lineages decreases, resulting in shorter branches near the tips of the tree and longer deeper branches; for the last two lineages, it will take on average N_e generations to coalesce. If **population size** changes with time, the rate of coalescent events changes as well, affecting the branch length pattern which can be captured by comparing with a constant size expectation; e.g. in a growing population ($N_e(t) > N_e(t+1)$, time is reversed) coalescent rate decreases with time producing longer than expected terminal branches.

When combined with genetic data for phylogenetic inference, a coalescent model is treated as an additional set of parameters (like effective population size(s) and coalescent times) which can be explicitly added to the likelihood function and estimated together with other evolutionary parameters (e.g. evolutionary rate or transition/transversion ratio) [33]. The Bayesian framework allows accompanying the likelihood function with informative priors which can enhance inference if used wisely [34–36].

One major advance of statistical frameworks that involve phylogenies is an ability to work with **time-stamped data** which is often the case for pathogenic samples [34,37]. This allows relating the reconstructed dynamics to calendar time. When working with outbreaks, time to the MRCA reconstructed by the model can inform us about the date when an outbreak might have been started as was done for the H1N1 outbreak in 2009 [38].

Estimates produced by coalescent models may be difficult to interpret in terms of epidemiology. An attempt to bring a coalescent model to a SIR model that uses a system of ordinary differential equations to describe the dynamics of three compartments (susceptible, infected, and recovered) during an outbreak showed that during an early exponential phase of the outbreak N_e estimates are indeed proportional to prevalence (the total number of infections) [39]. However, in general, there is no simple dependence of N_e on prevalence and incidence, thus N_e cannot be readily interpreted to be proportional to the number of infections. Volz [40] integrated the SIR model dynamics into the coalescent framework to directly estimate epidemiological parameters; later, a probabilistic, stochastic counterpart of this model was developed [41].

2.1.2.3. Birth-death models

Instead of a coalescent process, one can think of phylogeny as a distribution generated by **the birth-death (BD) process** [42]. In an epidemiological context, birth and death rates can be interpreted as the rate of transmission of a pathogen lineage and the rate of its death due to the host recovery, treatment, isolation, or death. The ratio between the birth rate and the death rate results in the key concept of epidemiology, **the reproduction number** R, which is defined as the average number of secondary infections produced by an infected individual. An early implementation of the BD model with constant birth and death rates [43] was soon expanded by [44,45] and allowed rates to vary over time. Direct modeling of SIR compartments in the BD model worked well compared with coalescence-based SIR models; moreover, it was shown to be better suited for describing the very noisy initial exponential phase of an outbreak [46] whose stochasticity coalescent models could not fully capture. Another important difference between BD and coalescent models is that sampling rate (or proportion) is modeled explicitly as a separate parameter in BD models. On the one hand, it

should allow accounting for sampling heterogeneity and biases; on the other hand, due to identifiability issues with BD models which make birth, death, and sampling rates interrelated, it is impossible to estimate the three rates separately — a strong prior knowledge on at least one of the parameters is required [44,45,47].

It should be noted that the coalescence and BD models, among other assumptions, operate under a neutral hypothesis. In the general case, this assumption does not hold true. Coalescent models with non-neutrality behave differently from the classical coalescent [48], for instance, more adapted lineages collapse more quickly [49]. A non-neutral extension was also developed for the BD framework allowing mutations on branches to affect birth and death rates [50].

2.1.2.4. Accompanying metadata

Finally, **pathogen samples are usually accompanied with metadata** that might include host species, sampling location, travel data, contacts with other infected individuals, risk group, therapy, and vaccination. All of these can be mapped onto phylogeny, and their impact can be tested [51,52]. For example, in [432] information about the host was incorporated into a coalescent model to study evolution of MERS-CoV (Middle East respiratory syndrome coronavirus) in humans and camels. The so-called **structured coalescent** allows different demes (in this case, hosts) to have their own population sizes and coalescent rates. The model correctly identified multiple instances of asymmetric camel-to-human migration that resulted in small tightly-clustered local outbreaks in humans, the pattern that could not be captured by a naïve parsimony-based model and, interestingly, by the structured coalescent that was forced to share the rate of coalescence across demes. In a similar way, the structured coalescent was used to study temporal patterns of transmission of yellow fever virus from wild primates to humans in Brazil in 2016-2017 [433].

Additionally, [433] used geographic coordinates of sampled locations along the dated phylogeny to study spatial patterns of viral transmission and infer the rate and direction of viral spread, an approach called **continuous phylogeography**.

2.1.2.5. The Ebola virus example

The Ebola virus disease outbreak in West Africa in 2013-2016 has probably marked the beginning of a large-scale molecular epidemiology. First identified in 1976, Ebola virus has been causing occasional zoonotic outbreaks in humans, with the West African epidemic of 2013-2016 being the most devastating one.

The epidemic in 2013-2016 has mainly affected Guinea, Sierra Leone, and Liberia resulting in 28,616 reported cases with at least 40% mortality rate [53]. Early sequencing of three viral genomes from Guinea [56] and 78 viral genomes from Sierra Leone [57] suggested that the 2013-2016 outbreak resulted from a single zoonotic transmission, being unrelated to previous outbreaks in humans [56,57]. Further sequencing efforts [54] during the epidemic provided a more detailed description of circulating lineages and allowed researchers to apply the coalescent to estimate the evolutionary rate of the virus and an upper bound of the zoonotic event (the coalescent model estimates the time to the most recent common ancestor of sampled viruses which is in general younger than the true ancestor of an outbreak).

Rapid sequencing provided an additional source of information on the ongoing transmission in real-time [58] — phylogenetic relationship of sequenced samples proved itself useful in unraveling transmission chains; for instance, it captured an episode of sexual transmission of the virus.

By the end of the epidemic, genomic data were available for more than 5% of the registered cases; this allowed for better understanding of migration patterns and timing of the

outbreak [55]. Migration events were tested for their association with various factors that could potentially explain viral dispersal; the results indicated the virus preferrably dispersed between geographically close areas, and dispersal within a country was preferred over international migration.

The same dataset was later analyzed by a different approach in [59]. The authors studied the effect of potential intervention strategies, had they been implemented at different times during the outbreak, by cropping parts of a complete geography-annotated phylogenetic tree. The total height (timespan) and the total length (sum of branch lengths) of the tree can serve as a proxy for the total duration and the total size of the epidemic, respectively. By cropping trees according to various possible control measures (for example, by prohibiting migrations over 200 km and further at different time points), the authors estimated the effect of each particular control measure on a reduction of both size and duration of the epidemic. Similarly to [55], the authors observed that long-distance viral transmission did not contribute much to the epidemic compared to short-distance dispersal. Additionally, [59] applied continuous phylogeography to infer the dispersal velocity of the virus which varied a lot during the epidemic.

Methods of molecular epidemiology become increasingly used in the context of infectious diseases; the growing body of available data on current outbreaks continuously challenges their development.

2.2. HIV-1

The human immunodeficiency viruses 1 and 2 (HIV-1 and HIV-2) are lentiviruses that cause the acquired immunodeficiency syndrome (AIDS) in humans [60]. HIV-1 is more infective and virulent compared to HIV-2 and causes epidemics worldwide while HIV-2 predominantly affects West African countries [61–63]. HIV-1 and HIV-2 differ in some aspects of their biology. I will be focusing on epidemiologically more significant HIV-1 in the text.

2.2.1. HIV-1 biology

A nearly 10kb genome of HIV-1 consists of three genes (*gag*, *pol*, and *env*) encoding structural proteins and enzymes and six regulatory genes responsible for the interaction of HIV-1 with infected cells and the host immune system [64]. Being a lentivirus, HIV-1 is able to integrate its genetic material into the host genome [65]. Two enzymes, both encoded by the *pol* gene, make that possible: reverse transcriptase (RT, RNA-dependent DNA polymerase) first converts a single-stranded RNA genome into a double-stranded DNA molecule, which is then integrated into cellular DNA by integrase. Once integrated, the HIV-1 genome is transcribed by the host RNA polymerase as a single primary transcript molecule [66] which can be either processed into various mRNAs through alternative splicing and translated into viral proteins or used as a genome source for new viral particles [67]. Drugs that are currently used in antiretroviral therapy (ART) interfere with some of these processes by targeting HIV enzymes — reverse transcriptase, protease, and integrase [68].

HIV-1 causes AIDS by depleting the population of CD4+ T cells (cluster of differentiation 4-positive T cells) which modulate the adaptive immune response [69]. CD4+ T cells are being killed by various mechanisms including apoptosis [70,71], pyroptosis [72],

and CD8+ T cells cytotoxicity [73]. When the number of CD4+ T cells drops below a certain level, the host organism becomes immunodeficient, being highly susceptible to opportunistic infections [74]. Although CD4+ T cells play a major role in AIDS development, HIV-1 can also infect other immune cells. HIV-1 enters cells using CD4 receptor molecules expressed on the surface of some immune cells including CD4+ T cells, dendritic cells, macrophages [75], and microglia [76], and co-receptor molecules, usually CCR5 or CXCR4 [77]. Interaction of viral envelope proteins with receptor and co-receptor results in a conformational change that brings HIV-1 envelope and cell membrane to close proximity and mediates membrane fusion and capsid entry into the cytoplasm [78]. A diverse range of susceptible cells and the ability of HIV-1 to integrate its DNA into the genome of host cells creates an unlimited latent reservoir of HIV-positive living cells that makes the complete eradication of HIV-1 impossible and complicates the therapy [79].

RNA viruses are usually characterized by high mutation rates. HIV-1 possesses the highest known mutation rate of more than 1×10^{-5} mutations per site per replication [80–82], implying at least one mutation per genome is acquired after ten replication cycles. This tremendously high rate is explained by (1) the error-prone replication process performed by low-fidelity reverse transcriptase which lacks the proofreading activity [83], (2) the protective action of the cellular enzyme APOBEC3G that deaminates cytidines in single-stranded DNA molecules [84], and (3) high recombination rate [85]. Two implications follow from the high mutation rate of HIV-1. First, HIV-1 should carry a substantial mutation load and exist close to its error catastrophe [86] as was demonstrated both in simulations [87] and experiments [88]. Second, rapidly diverging HIV-1 sequences are highly informative in the phylogenetic framework.

2.2.2. Genetic bottleneck and selection during HIV-1 transmission

As mentioned in Section 2.1.2, within-host and between-host phylogenies of HIV-1 look different being shaped by different evolutionary processes. While HIV-1 evolution within a host is governed primarily by positive selection and immune escape, the sampling process and demographics play a major role at the level of population. In addition to phylogeny shape differences, within- and between-host evolution processes are characterized by different evolutionary rates, with the rate of HIV-1 evolution at the level of population being lower than that inferred from within-patient phylogenies [89]. Several factors are thought to contribute to this difference.

First, viral transmission is often coupled with an extreme genetic bottleneck that reduces the number of genotypes that establish a new infection in a recipient to a few, if not a single, variant [90,91]. In 2004, the bottleneck was demonstrated by phylogenetic methods [92]: Derdeyn et al. sequenced the *env* gene of HIV-1 populations in eight heterosexual donor-recipient pairs and compared genetic diversity between the two populations. In all eight cases, *env* sequences of a recipient were monophyletic and rooted in a much more diverse ensemble of env sequences of a donor. Inflammatory infections in the genital tract can interfere with the bottleneck probably due to disruption of the mucosal barrier [93].

Second, the bottleneck is associated with selection acting on the transmitted genotypes as was demonstrated in the study of HIV-1 sequences isolated from 137 donor-recipient pairs with recent heterosexual transmission [94]. Amino acids in proteins encoded by the *pol*, *gag*, and *nef* genes that matched the population consensus transmitted better; protein sequences carrying these amino acids were predicted to have protein stability closer to the population consensus. Amino acids presumably associated with immune escape from HLA alleles of the host were less likely to be observed in the recipient. The

transmission bias was less pronounced in men with inflammatory infections in the genital tract and in women in general, implying that a more permissive environment can mitigate the importance of the selective advantage during the transmission. Transmission in men who have sex with men (MSM) seems to be associated with weaker selection upon transmission [95] and less stringent bottleneck [96] compared to heterosexual transmission. The bottleneck is also less pronounced among injection drug users [97], although a limited amount of the transferred material can also result in a bottleneck [98].

As preferable transmission of variants with a smaller effect on the protein structure in [94] suggests, the probable reason for the observed selection is that the acquired mutations beneficial during within-host evolution frequently turn out deleterious in a new environment during or after the transmission event. The latent reservoir of HIV-1 likely contributes to the establishment of a novel infection by a basal variant by constantly supplying a sufficient number of "patient-naive" HIV-1 variants [99].

The trade-off in the fitness landscape between the current host and a recipient predicts the selective advantage of reversion mutations. Indeed, the number of transmitted CTL (cytotoxic T lymphocyte) escape mutations in the *gag* gene was shown to negatively correlate with viral load in the recipient [100] implying at least some of the CTL escape mutations lower viral fitness. One example of such mutations is the mutation T242N in the *gag* gene that was described in a mother-to-child transmission pair [101]. T242N was identified as a positively selected mutation associated with a particular maternal HLA allele and a reduced CTL recognition. In an infected child carrying different HLA alleles, this mutation gradually decreased in frequency and became fully reverted after nine months. Although beneficial in a certain context, T242N was shown to reduce viral replication capacity *in vitro* [102]. Other

reversion mutations associated with CTL escape were also described [103], further supporting the importance of the trade-off.

Impressive patterns of reversion can be readily observed from the longitudinal sequencing data on HIV-1. Coherently with the preferable transmission of consensus-like variants observed in [94], the virus steadily reverts to the population consensus during intra-patient evolution with the rate of reversion among different sites proportional to the level of their conservation [104]; importantly, unlike [94], these patterns should not depend on the route of transmission.

The *env* gene evolution provides an additional illustration of how mutation-conferred advantage varies under different circumstances. At the early stages of the infection, HIV-1 predominantly infects cells that express the CCR5 co-receptor; as the infection progresses, HIV-1 can switch its specificity to CXCR4 [105–108]. In order to do that, HIV-1 has to evolve the gp120 protein (the *env* gene product) which is responsible for co-receptor recognition. While the specificity towards CXCR4 may be beneficial by allowing the virus to infect new cell types and evade the CCR5-specific cytokine response, the CCR5-specificity is crucial for the transmission. A 32-bp deletion in CCR5 hinders the HIV-1 transmission in homozygotic carriers [109] making them highly resistant; heterozygotic carriers demonstrate lower early viral loads when infected and slower progression towards AIDS compared to the general population [110,111].

Together, the discussed factors suggest that overall, HIV-1 tends not to transfer the majority of the diversity that has been "short-sightedly" accumulated within one host to the general population and is capable of reverting some part of unnecessary burden, has it been transmitted.

2.2.3. Population genetic diversity and origin of HIV-1

The transmitted variants describe the evolution of HIV-1 at the level of population. The star-like phylogeny of HIV-1 is represented by a bunch of separate genetically distant clades raising the question of their origin. The genetic comparison of HIV-1 and HIV-2 with lentiviruses infecting other species revealed a zoonotic origin of HIV [112]. HIV was found to be genetically close to simian lentiviruses infecting non-human primates [112]. HIV-1 is most closely related to the simian immunodeficiency virus that infects chimpanzees (SIVcpz) and emerged as a result of at least four independent transmissions of SIVcpz to humans as suggested by phylogenetic analysis [112,113]. Similarly, HIV-2 has entered the human population several times being transmitted from sooty mangabeys infected with SIVsmm [114].

The identified transmission events gave rise to genetically distinct lineages called groups with four groups, M, N, O, and P, described for HIV-1, and nine groups, A-I, described for HIV-2. The cross-species transmission of SIV is an ongoing process, as illustrated by the identification of a novel HIV-1 variant P [115] and a novel HIV-2 variant I [116] among samples collected in the 2000s that were not related to the known HIV groups. Group M of HIV-1 is the most abundant group responsible for the worldwide HIV pandemic.

HIV groups are further divided into subtypes, circulating recombinant forms (CRFs), and unique recombinant forms (URFs), the latter two being products of recombination events between different subtypes. Subtypes and CRFs represent distinct clades inside the group phylogeny. Demographic processes seem to contribute a lot to the phylogenetic distinctiveness of subtypes and CRFs which are thought to appear as a result of migration and spread of a single genetic variant across a novel geographic area [117]. The founder effect agrees well with the association of subtypes with certain geographic locations [118] and with

a much higher genetic diversity of HIV-1 in West Africa where HIV-1 has initially emerged [119]; the latter also indicates bias and incompleteness of sampling.

HIV is thought to have originated in Central and West Africa through bushmeat practices early in the 20th century [119]. The role of simian bushmeat in HIV emergence is supported by a high prevalence of SIV among people involved in such activities [120]. Thus, sporadic cross-species transmission events probably have been taking place earlier as well, but socio-economic factors such as the colonization of Africa, increasing population density, high prevalence of sexually transmitted diseases, and poor medicine allowed SIV to infect a substantial number of people and to evolve the ability to spread effectively in the human population [121,122]. The mechanism that allowed simian lentiviruses to cross the barrier between non-human primates and humans is not yet completely understood, probably differs for different HIV groups, and includes adaptation to different restriction factors in a new host [123].

The ability to counteract one of the host restriction factors, tetherin, provides an illustrative example of HIV adaptation. Tetherin is a protein that inhibits the release of viral particles in HIV and other enveloped viruses [124]. It possesses cytoplasmic, transmembrane, and external domains; the external domain is modified by the addition of phosphoglyceride to the C-terminus of tetherin [124]. When an enveloped viral particle is being released, the external domain of tetherin anchors in the viral particle membrane tethering the particle to the cell surface [125]. Different SIVs manage tetherin restriction differently. The suggested ancestors of SIVcpz were able to use two proteins, Nef and Vpu, to fight tetherin, that interact with its cytoplasmic and transmembrane domain, respectively [126,127]. The Vpu protein lost the ability to inhibit tetherin in SIVcpz making the Nef protein the only mechanism of protection against tetherin in SIVcpz [127]. However, Nef from SIVcpz is

inefficient against human tetherin that carries a deletion in its cytoplasmic domain [128]. Group M HIV-1 has adapted to deal with human tetherin by regaining the anti-tetherin activity of Vpu [129]; Group N HIV-1 has also partially restored this activity, although at a cost [130]. Group O HIV-1 has instead evolved its Nef protein to recognize a different motif of tetherin [131]; interestingly, a strain of group O that regained the anti-tetherin function of Vpu has also been described [132]. Group P which is believed to have been transmitted from gorillas does not show human-to-human transmission and is sensitive to tetherin [133]. HIV-2 viruses, and their SIVsmm ancestors, do not possess the Vpu protein. Instead, SIVsmm evolved to antagonize tetherin through the Env gene product, the mechanism now being used by HIV-2 [134,135]. These observations illustrate how genetically related viruses can adapt to the same conditions through different mechanisms.

2.2.4. Phylogenetics helps to reveal the date of origin of HIV-1

In 1981, multiple cases of immunodeficiency of unknown origin were reported among homosexual men in the US. Patients demonstrated severe susceptibility to opportunistic infections and extremely low levels of lymphocytes; many developed Kaposi's sarcoma — a type of cancer caused by human herpesvirus 8 which is developed rarely in countries with a low prevalence of this virus such as the US [136,137]. In 1983, the human immunodeficiency virus was first described — "this virus as well as the previous … isolates belong to a general family of T-lymphotropic retroviruses that are horizontally transmitted in humans and may be involved in several pathological syndromes, including AIDS" [138]. Multiple cases of AIDS prior to 1981 in the US and outside of it have been retrospectively diagnosed [139,140] raising the question of when HIV has entered the human population and how it has spread across. Phylogenetic analysis quickly proved itself useful in tackling those questions.

Because the M group of HIV-1 is the only HIV group that causes epidemics globally, most published works are focused on the M group HIV-1 data, although similar works for groups N and O exist as well.

One of the earliest estimates dating the origin of the group M HIV-1 was obtained by Korber et. al. in 2000 using "unprecedented amounts of data" (159 sequences of the *env* gene of subtypes A-F with known sampling year) [141].[1] Korber et. al. first reconstructed the maximum-likelihood topology of the available samples which was then rooted using the consensus sequence of the group M. The rooted tree was then assumed to evolve under constant rate, allowing the rate of evolution and the date of the group M LCA to be estimated by fitting the total branch length of every sample versus its sampling date; the model also incorporated the uncertainty in sampling dates. The linear fit (assumed by the strict molecular clock) dated the group M origin in 1931 (1915-1941). In agreement with this estimate, the linear fit correctly predicted the date of a sequence that was obtained from a sample collected in the Democratic Republic of the Congo (DRC) in 1959 and discovered two years prior to the current work [142] (known as the ZR59 sample in literature). Interestingly, an attempt to relax the strict molecular clock assumption resulted in a poorer fit and incorrect dating of the early African sample, although relaxed clocks are considered much better suited to HIV-1 data compared to the constant rate of evolution [143].

Later works further supported and refined the produced estimate [144–147] to be slightly earlier pointing at the beginning of the 20th century. Ancient HIV-1 samples, though scarce, are extremely useful in understanding the extent of early HIV-1 genetic diversity and validating the dating results. In 2008, a sequence from another ancient HIV-positive sample from DRC was described. The DRC60 sequence was obtained from the lymph node tissue in

[1] Let us take a moment here and appreciate how impressive has been the development of computational methods and hardware over the last 20 years 🥲

a paraffin block prepared in 1960 [144]. The comparison of DRC60 and ZR59 sequences revealed the high amount of divergence in the group M HIV-1 in 1960; the estimated date of the group M origin varied depending on the inclusion/exclusion of ancient samples and on the model used. The two equally supported models that included both ancient samples produced median estimates equal to 1908 and 1921 while the exclusion of ancient samples resulted in the best-fit estimate equal to 1933.

In 2019, the most ancient nearly full-genome sequence was obtained from a paraffin block prepared in 1966 in DRC (DRC66) [147]. DRC66 further supported the substantial level of divergence of HIV-1 in DRC in the 1960s; phylogenetic analysis placed the three ancient samples close to roots of different subtypes [144,147] implying that these subtypes have already diverged by the 1960s. In [147], the origin of group M was estimated to lie between 1881 and 1918; the estimates were based on complete genomes of mostly African samples collected in 1978-2018 and did not depend much on the presence of DRC66 in the analysis. The results of these and other works, although pointing at inevitable uncertainty in the estimates that depend on models and datasets used, agree on the onset of the HIV-1 group M in the early 20th century.

2.2.5. Phylogenetics helps to reveal the origin and spread of HIV-1 subtypes

As was mentioned in 2.2.3, the global distribution of the group M subtypes is not uniform with many countries associated with a certain predominant subtype [148,149]. For instance, subtype A is the most prevalent subtype in Russia and East Africa, subtype B represents North America, Australia, and Europe, and globally the most prevalent subtype C predominates in India and Southern Africa. Phylogenetic analysis can be used to track the events that seeded HIV-1 in a particular area.

The emergence of subtype B offers an illustrative example of the founder effect. Subtype B was the first subtype described; it was identified as the cause of immunodeficiency reports among US homosexual men in the 1980s. In parallel to the growing number of AIDS reports in the US, symptoms similar to AIDS have been reported in Haitian patients [150] raising the question of the origin of subtype B. Early phylogenetic analysis of the *env* gene sequences of samples collected in the US, Haiti, and RDC suggested that the virus first migrated to Haiti from Africa and then to the US from Haiti [151]. However, poor sampling in 1988 could not eliminate the possibility of HIV-1 being seeded in Haiti by US patients.

In 2007, complete *env* and partial *gag* sequences were obtained from archival samples of five Haitian immigrants who presumably got infected in Haiti before entering the US in 1975 and developed AIDS by 1981 [144]. Phylogenetic analysis revealed that the newly recovered sequences grouped with other Haitian sequences and were located basal to all 109 non-Haitian subtype B sequences analyzed, with 96 sequences forming the "pandemic" clade of subtype B that affected multiple countries worldwide. The fact that the whole non-Haitian diversity of subtype B known by 2001 was rooted inside the Haitian subtype B samples strongly supported the Haitian origin of subtype B. In addition, the origin of subtype B was dated between 1962 and 1970, in agreement with French-speaking Haitian citizens leaving for work to DRC in the 1960s [152]; the pandemic clade was dated to 1966-1972.

Similar estimates were produced in 2016 [153] when eight complete genomes of subtype B were recovered from 1978-1979 serum samples of the US patients; additionally, the HIV-1 genome of "patient 0" who was mistakenly considered the founder of the HIV-1 outbreak in North America was sequenced. The phylogeny clearly showed that HIV-1 from patient 0 is not basal to other American samples and is nested inside the diversity of HIV-1

collected in New York. The results obtained in [144] and [153] strongly support the establishment of the major subtype B lineage through a single migration event.

Upon its migration to the US in the early 1970s, subtype B further spread to other geographic locations [154,155]. The initial association of subtype B with homosexual transmission among the US and Haitian men left an imprint on the further history of subtype B which is still strongly associated with homosexual transmission route [156], although it is now also prevalent among heterosexual transmission in some areas [157].

Similarly to subtype B, other subtypes are also thought to have emerged through migration of HIV-1 variants from Central Africa to other locations [145,158–161]; early transportation network in DRC seems to have played a major role in early HIV-1 dissemination across Africa and beyond [145]. Subtypes that co-circulate in the same area frequently result in CRFs both inside and outside of Central Africa [159–161].

For many HIV-1 subtypes and CRFs, decades of evolution made them highly genetically diverse and distant from each other [162,163]; not surprisingly, this resulted in measurable differences in their virology and epidemiology. For example, subtype A is characterized by a lower rate of disease progression compared to its recombinant with CRF02 [164] and subtype D [165], with the latter being associated with a higher frequency of treatment failure [166]; the most abundant subtype C was shown have a lower replication capacity compared to subtypes A, B, D, and some CRFs [167,168]; CRF02 was shown to have higher replication capacity compared to its parent subtypes [169]; subtypes B and C were demonstrated to interact with different HLA alleles differently affecting the level of host susceptibility and the rate of disease progression [170].

The extent to which these and other measured differences contribute to transmission rates and population spread is still unknown. Lower replication capacity was speculated to

result in slower disease progression, longer transmission period, and higher population coverage [168]. This logic could explain the high abundance of subtype C, however the link between these factors has never been clearly demonstrated, and studies produced contradicting results on the disease progression rate of subtype C [171–173]. On the other hand, subtype C did spread faster across Africa compared to subtypes A and D, as demonstrated by methods of continuous phylogeography [174]; this was speculated to be associated with its emergence in a mining community and rapid spread through a transportation network. Thus history, rather than biology, might have a more drastic effect on rapid transmission and spread of different subtypes though it remains an open question.

2.2.6. Molecular epidemiology of HIV-1

In many countries, patients entering HIV care are frequently tested for drug resistance against the first-line antiretroviral therapy used in the area by sequencing their viral samples and analyzing them for the presence of drug resistance mutations. Genetic data generated by routine molecular surveillance can also inform healthcare professionals about the ongoing HIV-1 transmission in the population.

Closely related HIV-1 sequences can be grouped into molecular clusters based on different criteria. To some extent, a molecular cluster can be assumed to reflect the true transmission cluster, i.e. the history of HIV-1 transmission involving sampled HIV-positive individuals [175], thus molecular clusters can be used to describe some aspects of this process.

Different definitions of molecular clusters have been proposed based on sequence/phylogenetic data used. Alignment-based methods define clusters based on the genetic distance threshold (e.g. by restricting maximum or mean pairwise distance) [176];

phylogeny-based methods may rely on support values (e.g. bootstrap) [177] or geographic composition [178] of internal tree nodes. The two approaches can be combined so that both genetic distance and phylogenetic support are taken into account [179]. The clustering pattern was shown to be heavily affected by sampling density, as demonstrated by the dependence of sampling density on the fraction of clustered samples [180]; the fraction is largely dominated by noise when the fraction of sampled sequences is low. Clustering pattern is also affected by the selected thresholds for obvious reasons with stricter thresholds associated with smaller clusters [181]. One approach to tackle this is to use different thresholds for different purposes, i.e. first, to define a global transmission network using a relaxed threshold, and then identify and investigate smaller clusters within this network, probably spanning a shorter period of time, using a stricter threshold [181].

When sequences are accompanied with additional information (sampling date, the date of diagnosis, self-reported transmission route, epidemiological contacts, therapy status), the inferred clusters can be used to study factors affecting the dynamics of HIV-1 transmission [182–184]. For instance, studies in multiple countries reported the association of male gender and/or MSM transmission route with cluster membership [185–188]. The comparison of MSM and heterosexual (HET) transmission routes in the UK in 2007 [189] revealed that, although HET also forms clusters (among non-B subtype samples), these clusters are associated with slower transmission compared to MSM clusters [190], in agreement with a notable but declining role of HET transmission in the UK HIV-1 epidemic [191] and a higher risk of HIV-1 transmission in MSM [192]. Another study which analysed the growth of the UK transmission clusters in 2007-2009 [193] demonstrated how the dynamics and mode of the non-B subtype epidemic in the UK had gradually changed over time accompanied by a growing role of MSM and IDU (injecting drug users) transmission,

illustrating the interaction between risk groups and the necessity of continuous monitoring of changing epidemic properties as these affect the optimal strategy of prevention efforts. [194] provides another example of interacting risk groups where a high degree of HET clustering was observed within subtype B in the UK due to joint clustering of HET and MSM samples in MSM-dominated clusters.

The [194] study compared the two most densely sampled HIV-1 epidemics, in the UK and Switzerland, where for more than 60% of HIV-positive patients the *pol* gene sequences are available through the UK HIV Drug Resistance Database and the Swiss HIV Cohort Study database, respectively. Transmission clusters were identified for both datasets and compared in terms of their sizes, densities, transmission route, and subtype composition. Overall, in terms of clustering patterns, HIV-1 epidemics in the two countries appeared similar, with all observed differences being eliminated by the correction on different sample sizes and different distributions of sampling dates, except for higher levels of HET clustering within MSM-containing clusters in the UK. Importantly, the two datasets in [194] were analyzed by the same set of methods making the results on the two epidemics comparable, in contrast to results produced by different studies involving different definitions of transmission clusters. Epidemiological parameters were inferred for separate clusters of the UK and Swiss epidemics in a range of studies using phylodynamic methods discussed in 2.1.2, e.g. [43,44,178], although to my knowledge, the two countries have never been compared directly under the same models and sets of priors.

Transmission clusters tracked in real-time allow healthcare professionals to capture a novel HIV-1 outbreak and provide a timely response. The HIV monitoring system developed in British Columbia, Canada, is a sound example of real-time molecular surveillance [195]. The system was set up to provide daily updates on transmission clusters observed in the area

based on novel HIV-1 sequences. In June 2014, the system reported a substantial growth of one of the clusters, with 8/11 new cases carrying a drug resistance mutation. As transmitted drug resistance posed a threat to public health, extensive follow-up was initiated for the members of this cluster and their reported contacts resulting in a substantial reduction of viral loads and curbed further transmission within the cluster in the next several months.

A combination of phylogenetic analysis and epidemiological data represents a tempting forensic tool that was first used in the context of HIV-1 in the 1990s. The US CDC (Centers for Disease Control and Prevention) used phylogenetic analysis to study the transmission of HIV-1 between an HIV-positive dentist and his patients. The HIV-1 sequences isolated from the dentist and five of his patients formed a separate cluster on the phylogenetic tree that also included sequences from other HIV-positive patients in the local area used as a control. The cluster of closely related sequences was genetically divergent from local control cases and was supported by 79% of bootstrap replicas. These observations, together with epidemiological data including the diagnosis of the dentist with Kaposi's sarcoma in 1986, reports of performed invasive dental procedures, and the absence of known risk factors among the patients, "strongly suggested" that the five patients were infected by the dentist [196]. In 1997, phylogenetic analysis was for the first time used as a piece of evidence in the US court — a gastroenterologist was found guilty of infecting a victim with HIV-1 by injecting her with the blood of his HIV-positive patient [197]. In a similar way, HIV-1 samples were used in a court in Sweden [198].

Phylogenetic results obtained in these and other [199] impressive works can only serve as an additional line of evidence in litigations. A major limitation of the phylogenetic approach and transmission cluster analysis in general is the fundamental inability to sample the complete transmission chain. Despite extensive efforts in contact tracing, the possibility

always exists that an unidentified member of the transmission network intervenes in the two otherwise directly related observed infection cases.

It was suggested that mutual topologies of multiple samples per patient, i.e., data on within-patient diversity, could be informative of the direction and/or directedness of the transmission [200]. For example, the direction of the transmission can be inferred if all HIV-1 sequences obtained from a presumed recipient are consistently rooted within HIV-1 sequences of a presumed (either direct or indirect) donor. The directedness of the transmission was proposed to be almost certain when mutual topologies are intermingled; simulations suggest that indirect transmission is highly unlikely for intermingled topology unless the number of transmitted lineages is rather high [200]. In many cases, these theoretical expectations are met in real data [200,201]; yet, no strict correspondence exists. For instance, intermingled topologies were observed in "sibling" cases sharing a common infection source in [201]. Intermingled, closely related topologies were also observed in two pairs of women in the study conducted in Uganda [202]; the direct transmission was assumed to be highly unlikely here because "HIV-1 is predominantly sexually transmitted in Africa, and extremely rarely transmitted sexually between women" [202], although other ways of transmission, e.g. through injecting drugs, were not ruled out. Additionally, Ratmann et al. in [202] determined the maximum genetic distance between HIV-1 repertoires in known transmission pairs and proposed to use this threshold (2.5%) to identify direct transmission. Among the general population with no transmission history, 15% of closely related pairs were coming from pairs of women which is again unexpected assuming sexual transmission is the main route of transmission. Thus, inferences made from within-patient diversity can inform population-level studies, but still cannot serve as self-sufficient evidence and should be used carefully in the legal context.

2.3. SARS-CoV-2

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel strain of coronavirus that was first identified in December 2019 and resulted in the global COVID-19 (coronavirus disease 2019) pandemic with 219 million total cases and more than 4.5 million deaths worldwide as of September 2021. The response of the scientific community to a novel virus was impressively rapid. The WHO was first notified of an outbreak of pneumonia of unknown etiology in Wuhan, China on December 31, 2019. In less than two weeks, on January 10, the first complete genome sequence of SARS-CoV-2 was published [203]. On January 22, the first version of a protocol for SARS-CoV-2 amplicon sequencing was released by ARTIC [204] allowing the researchers to monitor and study the early stages of the pandemic. Since then, more than 3 million complete genome sequences of SARS-CoV-2 have been produced and uploaded to the GISAID database [205], and more than a million are available through other databases (e.g. Genbank and COG-UK [206]), probably providing COVID-19 with the highest known quantity of molecular data among human infectious diseases.

In this Section, I am briefly reviewing the current state of knowledge on SARS-CoV-2.

2.3.1. SARS-CoV-2 biology

SARS-CoV-2 belongs to coronaviruses which infect birds and mammals. Before SARS-CoV-2, eight strains of coronaviruses had been known to cause infections in humans [207,208], including two infamous strains, SARS-CoV and MERS-CoV (Middle East respiratory syndrome coronavirus), that caused severe outbreaks in 2002-2004 [209] and 2012 [210], respectively; the remaining six strains cause mild seasonal infections. In many

cases, SARS-CoV-2 is either asymptomatic or manifests itself as a mild respiratory disease [211,212]. Sadly, in some cases, SARS-CoV-2 triggers an inadequate reaction of the immune system involving massive cytokine storm and hyperinflammation, causing severe damage to lungs and in some cases other organs [213–216].

Coronaviruses are positive-sense single-stranded enveloped RNA viruses with genome size ranging from 27 to 31 kb [217]. The genome encodes a massive replicase polyprotein, four structural proteins (E, M, N, and S), and a number of accessory proteins. The replicase polyprotein upon synthesis is cleaved into multiple non-structural proteins, including RNA-dependent RNA polymerase responsible for both replication and transcription of the virus, and exonuclease protein possessing the proofreading activity that ensures high-fidelity replication [218]. Among structural proteins, the positively-charged nucleocapsid protein (N) is responsible for tight packing of the RNA genome forming nucleocapsid, while the envelope (E), membrane (M), and spike (S) proteins are incorporated into a lipid bilayer forming the envelope around the nucleocapsid. Trimers of S protein are anchored in the membrane and exposed to the outside forming corona-like "spikes" which are responsible for the interaction of a coronavirus with receptors on the surface of host cells [219].

2.3.2. The spike protein

The spike protein consists of two domains: S1 carries a receptor-binding domain (RBD), and S2 contains a fusion peptide [219]. After the receptor is recognized by the S1 domain, host proteases act on S2 liberating the fusion peptide that mediates membrane fusion [220,221]. In some coronaviruses, including MERS-CoV, S1 and S2 are also cleaved during

S protein biosynthesis which is believed to facilitate receptor recognition and virus entry due to a conformational change in S protein [222,223].

Being the most immunologically noticeable part of a viral particle, the spike protein of SARS-CoV-2 is used as a target of vaccines developed during the ongoing pandemic. As of September 2021, seven COVID-19 vaccines are approved by WHO [224], several are undergoing approval and dozens are under development. Currently used COVID-19 vaccines use the mRNA encoding S protein delivered by lipid nanoparticles, the DNA of the spike gene delivered by adenoviral vectors, the synthesized spike protein itself, or the inactivated virus to stimulate the immune response [225]. Vaccines provide the most efficient way of protection against COVID-19; vaccine availability and efficacy is a critical issue as no effective universal treatment exists to date.

S protein determines host cell specificity as different coronaviruses interact with different cell receptors [226]. SARS-CoV-2 operates through the ACE2 receptor (angiotensin-converting enzyme) expressed on the surface of many epithelial cells, similarly to SARS-CoV [227]. SARS-CoV-2 demonstrates a higher affinity toward ACE2 [228,229] compared to SARS-CoV which is thought to be due to the insertion of four amino acids, PRRA, on the border between S1 and S2 domains that encodes a cleavage site of furin, a protease widely expressed in human cells. The insertion facilitates a conformational change and ACE2 recognition, as furin-cleaved S protein trimers possess a much higher fraction of an open, receptor-accessible conformation compared to uncleaved trimers [230]; the insertion was also shown to promote viral entry into human epithelial cell lines and affect transmission in ferrets, model animals possessing respiratory system similar to humans [231]. In SARS-CoV, S1-S2 cleavage was shown to be limited, although it could be facilitated by high concentrations of trypsin or by the insertion of a furin-like site [220]. The origin of the

insertion in SARS-CoV-2 is currently unclear as the closest known coronaviruses lack this pattern although furin sites are widespread in coronaviruses as well as other viruses [232].

2.3.3. The origin of SARS-CoV-2

The origin of SARS-CoV-2 itself is also unclear. Coronaviruses possess a wide range of host species; cross-species transmission is common and occasionally results in a coronavirus entering the human population [233]. Phylogenetic analysis places SARS-CoV-2 within a clade of non-human coronaviruses mostly isolated from bats [234]. The closest known relative of SARS-CoV-2 is the RaTG13 strain isolated from a bat species in 2013. Still, the genetic distance between SARS-CoV-2 and RaTG13 is 4%, implying decades of independent evolution, while the closest human coronavirus, SARS-CoV, is separated from SARS-CoV-2 by 20%. Given how frequent zoonotic transmissions in coronaviruses are, it is reasonable to suggest a zoonotic origin of SARS-CoV-2. A zoonotic origin was proven for both previous severe respiratory human coronaviruses, with civets and camels identified as intermediate hosts between bats and humans for SARS-CoV and MERS-CoV, respectively. For SARS-CoV-2, no animal host has been yet identified which is rather unsurprising given the untold diversity of coronaviruses in bats; interestingly, it took researchers more than ten years to discover a bat coronavirus showing 96% identity to SARS-CoV [235]. Another non-human coronavirus relatively close to SARS-CoV-2 was identified in pangolins [236,237]. While overall identity between pangolin-CoV and SARS-CoV-2 is lower (90%), pangolin-CoV exhibits a striking similarity to SARS-CoV-2 in the RBD region, showing only a single amino acid difference within receptor binding motif comprised of 70 amino acids. Furthermore, the spike protein of pangolin-CoV shows a comparable ACE2 affinity [238] and mediates entry to human ACE2-expressing cells more efficiently compared to

SARS-CoV-2 [239], implying some animal coronaviruses may require no adaptation to infect (or at least enter) human cells. Both convergent evolution and recombination events might have shaped the similarity of S protein in the two coronaviruses.

2.3.4. SARS-CoV-2 evolution

The proofreading activity of the viral replication machinery results in a relatively slow evolutionary rate which should be especially important for coronaviruses possessing one of the largest viral genomes. First estimates of SARS-CoV-2 evolutionary rate were around 0.001 substitutions per site per year [240,241] and later were refined on larger datasets to be slightly lower [242]. The inferred rate of evolution predicts the ancestor of the circulating strains to exist in late November — early December 2019 [242]. Simulations predict that this ancestor is likely to be a descendant of a strain that entered the human population around a month before the epidemic became established; simulations also indicate a highly stochastic nature of pre-epidemic viral dynamics as the successful establishment of a novel lineage in the human population is predicted to be twice less likely than dying out [242]. Stochastic behavior also accompanies the established pandemic as illustrated by the prominent role of superspreading events in SARS-CoV-2 transmission which makes contact tracing a critical tool in epidemic monitoring and timely response [243,244].

Despite a relatively low evolutionary rate, SARS-CoV-2 has accumulated a sizable amount of genetic diversity while circulating in the human population over (almost) the last two years. Soon after the beginning of the pandemic, the need for systematic classification of genetic diversity in an increasingly growing corpus of sequencing data became apparent. Several related classifications were proposed, all based on mutation patterns. For example, the GISAID database which possesses the largest collection of SARS-CoV-2 genomes

denotes major clades of SARS-CoV-2 after amino acid substitutions defining these clades: e.g., the clade G is named after the D614G substitution in S protein, and its descendant clade GK is denoted after an additional substitution in S protein, T478K [245]. Compared to the nine large clades defined in GISAID, the Pango classification provides a dynamic hierarchical nomenclature of SARS-CoV-2 lineages which allows for a finer resolution of growing SARS-CoV-2 diversity; Pango lineages agree with GISAID clades, e.g., clades G and GK correspond to lineages B.1 and B.1.617.2, respectively [246]. Additionally, some of the lineages of particular interest that are suspected to pose a separate threat by being more transmissive, pathogenic, or elusive of vaccines, are further labeled by the Greek Alphabet letters by WHO, e.g. the B.1.617.2 lineage recently denoted as Delta [247].

2.3.5. SARS-CoV-2 epidemiology

An unprecedented sampling density of genetic data achieved by impressive sequencing efforts of multiple countries made it possible to use and further develop methods of molecular epidemiology. Phylogenetic and phylogeographic approaches have been used to describe early SARS-CoV-2 imports into different countries including Italy, Germany, the UK, and the US [248–256]. For instance, the United Kingdom has been successfully maintaining sampling density around or higher than 10% since the beginning of the pandemic [257]; this allowed researchers to describe in detail the dynamics and properties of multiple SARS-CoV-2 introductions into the UK and assess the effect of control measures during the first epidemic wave of COVID-19 in the country; heterogeneity in sampling profiles in and outside the UK was accounted for by including incidence and travel data in the model [253]. Undersampling is critical for inferences; a severe COVID-19 outbreak in Italy in early 2020 was speculated to possibly emerge from Germany [250]. Phylogeographic simulations [250],

together with a recently developed model that directly incorporates undersampling and travel data into the phylogeographic framework [258], proved this scenario unlikely.

Of particular interest are epidemiological parameters that govern the spread of SARS-CoV-2, in particular, the reproduction number (R). The reproduction number is the key epidemiological concept defined as the average number of secondary infections produced by an infected individual. At the onset of the pandemic, the basic reproduction number characterizing the rate of transmission in a fully susceptible population was estimated from purely epidemiological data like incidence and death counts in China and other locations [259,260]; the dynamics of R through time was used to assess the effect of control measures [261].

The reproduction number can also be inferred from genetic data [249,262,263], which allows for a direct comparison of different genetic variants. The D614G substitution in S protein mentioned in 2.3.4 appeared early in the pandemic and soon became the predominant variant raising the question of its selective advantage. *In vitro* studies readily demonstrated higher infectivity of D614G [264–266], but this did not convert into an apparent signal on population data. On the UK dataset, the 614G variant tended to have a higher growth rate compared to 614D, but the difference was insignificant [267]. In Washington State, the US, the rapid growth of 614G was mainly attributed to mobility and migration patterns differences, with 614G being introduced into the state much more frequently compared to 614D; the estimated growth rates again did not differ much, and no difference in disease severity and the risk of hospitalization was observed between 614D and 614G despite a slightly higher viral load in 614G [263]. In agreement with [263], a birth-death model incorporating fitness effects of individual mutations into the rate of birth events did not find

evidence for a substantial effect of 614G on transmission fitness in the US [268]; instead, regional transmission differences seem to noticeably contribute to the early growth of 614G.

2.3.6. Concerning genetic variation of SARS-CoV-2

Unfortunately, some of the evolved variation in SARS-CoV-2 is less benign. WHO currently acknowledges four variants of concern (VOCs), named Alpha (first identified in the UK), Beta (South Africa), Gamma (Brazilia), and Delta (India) [269], each characterized by an overlapping set of mutations affecting viral transmission, disease severity, and/or vaccine performance. Many of these recurrently emerged mutations reside within the spike protein, which is expected given its central role in the SARS-CoV-2 biology (Table 2.1).

Table 2.1. Amino acid substitutions in S protein involved in the definition of VOCs. "+/-" marks mutations present as sub-lineages within a variant.

Mutation	Suggested effect	Alpha	Beta	Gamma	Delta
K417N/T	K417N reduces antibody neutralization [270]		+	+	+/-
L452R	Increased in vitro infectivity, reduced antibody neutralization [271]	+/-			+
E484K	Reduced infection- and vaccination-induced antibody neutralization [272–275]	+/-	+	+	
N501Y	Increased affinity towards ACE2 [276]	+	+	+	
P681H/R	Assumed effect on S1-S2 furin cleavage; no experimental support to date [277]	+		+/-	+

VOCs are defined by multiple mutations in and outside the spike protein; although experimental data, as well as phylodynamics [268], predict some of them to affect viral properties, effects of individual mutations are impossible to disentangle when it comes to

transmission in the human population; lineage-defining mutations are tightly linked, and some of them may need to interact in order to be beneficial for between-human transmission. Still, for VOCs, we can at least observe a cumulative effect of an ensemble of mutations on viral dynamics in population data. The increased rate of transmission was reported for all VOCs with the Delta variant being the most transmissive [278–280]. Disease severity also tends to be higher for VOCs [281–285], again being most pronounced in Delta, which increases risks of hospitalization and death more than twice compared to a non-VOC variant [281].

High transmission rate allows VOCs to outperform and outcompete other circulating variants and each other as has been recently illustrated by the Delta variant, which spent 2021 outcompeting the previously predominant variant Alpha. Rapid reshuffles like that make the issue of vaccine effectiveness critical — vaccine development and production is a rather time-consuming process that doesn't benefit our race against viral evolution and adaptation. Fortunately, despite the reduced sensitivity of VOCs to antibodies [275,286–289], current vaccines still demonstrate substantial, although in some cases reduced, effectiveness against VOCs [290–292], and importantly, seem to protect against severe cases [290,293].

Still, these observations are a red flag as several currently circulating SARS-CoV-2 lineages carry mutations present in VOCs stressing the importance of close monitoring and extensive vaccination, as a continuous persistence of SARS-CoV-2 can result in the emergence of even more transmissive, aggressive, and elusive variants.

CHAPTER 3: Molecular epidemiology of HIV in Oryol Oblast, Russia

3.1. Introduction

HIV-1 poses a substantial threat to public health in Russia. Russia is characterized by one of the highest HIV-1 prevalence rates among European countries [294]. Since its initial introduction in the late 1980s, HIV-1 has been rapidly spreading across Russia due to widespread intravenous drug usage and poor public awareness. In 2019, 97,176 new cases of HIV-1 were registered in Russia [295]. Its prevalence in the same year was estimated at 0,75% based on registered cases [295]; the actual prevalence is likely higher. Currently, the epidemic predominantly develops through heterosexual transmission. In 2019, 63.9%, 33.0%, and 2.2% of all new cases with reported transmission routes could be attributed to heterosexual, injecting drug-associated, and homosexual routes of transmission, respectively [295]. Low coverage of antiretroviral therapy (ART) contributes to poor epidemic control. In 2019, only 48.5% of the registered people living with HIV-1 in Russia were receiving therapy [296].

Analysis of molecular surveillance data can provide insights on pathways and the rate of disease spread in an ongoing epidemic. However, the Russian diversity of HIV-1 remains poorly studied. A comprehensive description of the countrywide epidemic by methods of molecular epidemiology in Russia is hindered by the fact that genetic data on HIV-1 is available for less than 1% of the infected population. Coverage provided by molecular epidemiology studies of HIV-1 in specific regions of Russia is also low [297–301].

Here, we report a detailed molecular epidemiology analysis of the HIV-1 epidemic in a single region of Russia by covering a large fraction of its HIV-positive population. We

focused on Oryol Oblast [302], a region with a population of 736,483 located in the southwestern part of the Central Federal District of Russia [303]. As of 2019, there were 2,157 registered HIV-positive people in Oryol Oblast (Supplementary Fig. A-1); we collect and analyse HIV-1 genetic data from 768 patients, thus covering over a third of the registered epidemic. Using phylogenetic analysis, we infer 82 imports of HIV-1 into Oryol Oblast that were further transmitted within this region forming transmission lineages, as well as 250 imports that did not result in observed onward transmission, indicating unhindered spread between regions of Russia. Transmission lineages were enriched in injecting drug users but not in males, reflecting the demographic properties of the epidemic. The epidemic is predominated by subtype A (87% of our dataset) followed by the recombinant CRF63 variant (7%). Using phylodynamic analysis, we show that subtype A is responsible for the moderate growth of the epidemic with R_e of 2.8 [1.7-4.4], while CRF63 demonstrates a much higher growth rate and should be closely monitored.

3.2. Methods

3.2.1. Data collection and ethics

Patients were enrolled in the study between January 1, 2018, and June 30, 2019. Over this period, 681 blood samples were collected from HIV-infected people living in the Oryol city and the remainder of Oryol Oblast (Fig. 3.1) by the local AIDS center through a routine surveillance program and regular check-ups of the registered HIV-infected people. Additionally, we included the 241 samples obtained between March 31, 2014, and November 2, 2019, in the course of a study on drug resistance, for a total of 922 samples from Oryol Oblast. HIV-1 RNA for sequencing was obtained from blood plasma left after the viral load analysis. Demographic, clinical, and epidemiological data for participants were obtained from their medical records. The assumed route of infection was recorded by interviewing the patients. Written informed consent was obtained from all subjects. The study was approved by the Local Ethics Committee of the Central Research Institute of Epidemiology (protocol 93).

3.2.2. Sequencing

Sequencing was performed between June 29, 2019, and March 30, 2021. The sequence of the *pol* region covering the protease gene and part of the reverse transcriptase gene was obtained either by Sanger or by next generation sequencing (NGS). In both cases, RNA was isolated from blood plasma using phenol-chloroform extraction. For Sanger sequencing, the AmpliSens HIV-Resist-Seq (CRIE, Russia) kit for *in vitro* diagnostics was used according to the manufacturer's instructions. For NGS, a two-step amplification procedure was used. The first step of amplification was combined with reverse transcription.

Amplification was performed according to the following protocol: 45°C — 30 min; 95°C — 15 min; 30 cycles: 95°C — 30 s, 50°C — 30 s, 72°C — 1 min 30 s; 72°C — 5 min. During this stage, a DNA fragment of approximately 1.5 kb was amplified (positions 2074-3539 in the reference HIV-1 strain HXB2, GenBank K03455). The second step of amplification was performed in four independent tubes for each sample. Amplification produced four overlapping DNA amplicons that ranged in size from 427 to 586 bp. This approach made it possible to simplify library preparation and eliminated the need for DNA fragmentation. After purification with Sera-Mag Magnetic Speedbeads (GE Healthcare Biosciences) magnetic particles, the amplified fragments were mixed in equal proportions. After barcoding, next-generation sequencing was performed on the Illumina Miseq machine (Illumina, USA) with the MiSeq Reagent Kit V3 (600 cycles). Totally, out of the 681 samples collected in this study, 562 samples were sequenced using NGS, and the remaining 119, using Sanger technology.

3.2.3. Iterative consensus calling

The diversity of Russian HIV-1 differs significantly from the widely used HXB2 reference, complicating variant calling. Furthermore, amplification of the *pol* fragment via four slightly overlapping amplicons prevented the use of *de novo* assembly tools like IVA [304]. To address these issues, we developed the following custom consensus calling pipeline, and applied it to each of the 562 samples sequenced on the Illumina platform:

1. Trim paired reads using Trimmomatic-0.33 [305] (with options ILLUMINACLIP:$adapters:2:30:10 LEADING:5 TRAILING:5 SLIDINGWINDOW:5:15 MINLEN:50);
2. Align the trimmed reads against a set of curated reference sequences from LANL HIV

(198 sequences, [306]) using blastn v.2.2.31 [307] with default options; select the closest reference sequence REF_DRAFT;

3. Transfer the coordinates of the four pairs of primers from the HXB2 reference to the selected REF_DRAFT;

   Assign REF_INIT = REF_DRAFT;

4. Map the trimmed reads against REF_INIT using smalt v.0.7.6 [308] (options -n 1 -i 1000); convert the resulting SAM file into BAM using samtools v.1.2 [309]; trim primer sequences from reads using ivar v.1.3.1 [310] based on the coordinates inferred at step 3; extract the clipped reads and the covered part of REF_INIT (<1.5kb length) from the BAM file using bedtools v.2.29.2 [311];

   Assign REF_CUR = REF_INIT; iterate steps 5-8 until no more called variants are accepted;

5. Map the clipped reads obtained at step 4 against REF_CUR using smalt;
6. Call SNPs and indel variants with lofreq v2.1.5 [312] (-C 4 --call-indels --no-default-filter --use-orphan) and filter them, again with lofreq v2.1.5 (-Q 20 -K 20 --no-defaults -v 4 -V 0 -a 0.500001 -A 0); if no more variants are called, exit;
7. Detect and mask low coverage regions (excluding called deletions) using bedtools-2.29.2 and bedops v2.4.39 [313];
8. Apply the called variants and regions to mask to REF_CUR using bcftools-1.10.2 [309] to produce REF_NEXT; assign REF_CUR = REF_NEXT.

3.2.4. Dataset preparation

To obtain data on HIV-1 diversity in Russia beyond Oryol Oblast, we downloaded the 14,365 sequences of HIV-1 collected in Russia from Genbank on 2021-08-16 ("HIV-1" AND

"Russia" query). Of those, we selected the 8,560 sequences that produced at least 800bp hit of blastn against the target *pol* fragment of at least one reference sequence from [306]. We then additionally filtered out the 1,491 sequences that were sampled outside of Russia and processed by Russian research groups thus erroneously matching the 'Russia' query; and the 26 samples from Oryol Oblast already in our dataset, leaving us with 7,043 samples. Genbank metadata was parsed using a custom python script for 6,252 samples; for the remaining 791 samples, fuller metadata was provided by our collaborators.

Together with the 922 sequences collected in Oryol Oblast, our Russian dataset comprised 7,965 sequences. Among the 922 Oryol samples, we identified 94 patients who had more than one sample sequenced. For each of these patients, we marked all samples except the earliest one for exclusion later in the pipeline.

3.2.5. Sequence alignment and processing

Sequences were putatively aligned against the HXB2 reference *pol* region using mafft [314] (option --auto) and cropped to include only the coding part of the *pol* fragment (HXB2 coordinates 2252-3539). Additionally, we filtered out the sequences that either (a) had insertions relative to the HXB2 reference longer than 50bp, or (b) were shorter than 1,100 bp, leaving us with 6,356 sequences. These sequences were further aligned more accurately by the HMM-align algorithm of HIValign [315], allowing for up to 10 codons to compensate for a frameshift. From this alignment, we excluded 154 sequences carrying premature stop codons, frameshifts, or more than ten Ns (missing data characters). Sanger sequences were somewhat shorter than sequences produced by NGS; to make sequences comparably informative, we excluded codons with more than 5% of gaps or Ns from the alignment. The resulting alignment of 1,113 sites (positions 2,253 to 3,365 in the HXB2 reference) contained

6,202 sequences, including 864 samples from Oryol Oblast.

3.2.6. Subtyping and DRM annotation

We used the SierraPy client [316] to assign HIV subtypes and predict drug resistance mutations. A minor fraction of samples were assigned to mixed variants (CRF02+A and A+B in Fig. 3.2). For further analyses, we kept only those samples corresponding to the three genetic variants most abundant in Oryol Oblast: A, B, and CRF63. Each of these three variants showed monophyly on the Russian phylogeny (Fig. 3.2), indicating that they can be identified robustly from a short genomic fragment. Still, it is possible that some of the samples were misidentified. While a similarity-based algorithm in SierraPy has likely correctly picked the closest genome in its database, some samples may in fact descend from a recent recombination event, an unfortunate possibility that we cannot eliminate with our data.

To estimate the contribution of mutations at sites associated with drug resistance (DRM), we additionally defined a set of relevant DRM sites as the 23 codons that were reported to carry DR-associated mutations in at least 5% of our samples and repeated some of our analyses on an alignment with these sites masked (Supplementary Fig. A-15,16 and Supplementary Table A-5).

3.2.7. Phylogenetic analyses

To validate subtyping and check the locations of samples from repeatedly sequenced patients, we reconstructed a putative Russian-wide phylogeny for the 6,202 samples using Fasttree2 [317] (double-precision release).

For the final dataset, we kept only the sequences of the earliest samples for all patients whose samples were sequenced more than once; we also excluded the three pairs of samples

coming from the same patient but separated by more than 10% genetic distance (Supplementary Fig. A-2), leaving us with 768 Oryol and 5,328 non-Oryol sequences. We then used Fastree2 to reconstruct phylogenies for A, B, and CRF63 variants and analyzed them separately using Treetime [318] as follows. First, the trees were rerooted to maximize the temporal signal. Next, we reconstructed the ancestral nucleotide states of internal nodes in order to convert the tree to nonbinary by removing branches that did not carry any mutations. On the resulting nonbinary tree, we then determined the geographical states of the internal nodes using a two-state (Oryol vs. non-Oryol) mugration model. Finally, based on sequences with a known year and month of sampling, we inferred the dates of internal nodes which were used as LCA estimates (see Fig. 3.5).

3.2.8. Identification of imports

We identified imports by the depth-first search of the largest clades that (1) included at least 80% of Oryol Oblast samples, (2) had bootstrap support of at least 0.8, and (3) had the last common ancestor that had at least 80% posterior probability of being an Oryol node (Fig. 3.3). For imports resulting in just a single sequence (Oryol Oblast singletons), criteria 2 and 3 were ignored. To study the dependency of the number of inferred import lineages on the number of Oryol (Fig. 3.5) and non-Oryol (Fig. 3.6) sequences available, we randomly subsampled different numbers of Oryol or non-Oryol samples in 1,000 trials, inferring in each trial the resulting number of imports using criterion (1) above.

3.2.9. Bayesian phylodynamics

In order to infer the epidemiological parameters of subtype A sub-epidemic, we used a recently developed implementation of a birth-death model in BEAST2 that jointly analyses

multiple independent clades and infers a single set of epidemiological parameters shared across all clades [263]. This implementation is capable of aggregating clades spanning different time intervals and setting any arbitrary timepoints where epidemiological parameters can change in a piecewise-constant fashion, allowing us to use priors informed by data. As birth-death models require at least one parameter to be fixed or defined in a narrow range, we put a strict prior on sampling proportion. To obtain it, we selected samples that were sequenced soon (in less than a year) after the initial diagnosis which we denote as Dataset I. In Dataset I, most samples were sequenced in 2018 and 2019, eight samples were sequenced in 2014-2017, and none before 2014. We thus put a four-dimensional prior on the sampling proportion which was allowed to change on the first day of 2014, 2018, and 2019, with the mean equal to the number of samples sequenced in a time interval divided by the number of new diagnoses in this interval based on the reported case counts (Supplementary Table A-1). We then used these priors on both Dataset I and Dataset II composed of all samples. In contrast to the sampling proportion, the reproduction number and the rate of becoming uninfectious were assumed to be time-independent and thus unidimensional. We used a relaxed lognormal uncorrelated clock with ulcd.mean and ulcd.std shared across all imports. For the ulcd.mean prior, we used the normal distribution with mean equal to the median rate inferred in [431] for the *pol* region of subtype A (0.0015 substitutions / site / year) and sigma equal to 0.001. All other priors were kept default. Priors used in the multi-tree birth-death analysis are summarized in Supplementary Table A-2. The analysis was run for 250 million steps; we discarded the first 10% steps as burn-in.

We used the same set of priors to infer the epidemiological parameters of the largest clades of variants A and CRF63. CRF63 is a relatively young subtype; we could not find independent estimates of its evolutionary rate in the literature and thus used the estimate

produced in [431] for its ancestral recombinant variant CRF02 (0.0008 substitutions / site / year); using twice as large value as a prior mean (0.0016, which is close to the estimate produced for the second parent of CRF63, subtype A) did not affect results qualitatively. The parameters were inferred using the BDSKY [44] package of BEAST2 [319]. The skylinetools package [320] was used to define appropriate time intervals for sampling proportion. Priors used in this analysis are summarized in Supplementary Table A-3. The analysis was run for 100 million steps; we discarded the first 10% steps as burn-in.

The logistic growth dynamics for the same two clades were inferred in BEAST v1.10.4 [34]. Priors are provided in Supplementary Table 4. The analysis was run for 100 million steps; we discarded the first 10% steps as burn-in.

Convergence of the produced MCMC trajectories was assessed using Tracer [321].

3.2.10. EpiEstim

We used the EpiEstim package [322] implemented in R to infer the dynamics of R_e from the reported incidence data. We constructed the distribution of serial intervals based on the estimates of HIV-1 transmission rates at different stages of infection inferred by Hollingsworth et al. in [323]. As our incidence data is provided per year, we converted the reported transmission rates to be year-wise (see Supplementary Fig. A-14). The dynamics of R_e was inferred using a five-year sliding window.

3.2.11. Analysis of transmission lineages

We tested whether the two categories of samples, male gender and IDU route of transmission, are overrepresented in transmission lineages compared to singletons. For this purpose, we reshuffled the gender or transmission route labels of Oryol Oblast samples

10,000 times and obtained the distribution of the expected number of samples of the tested category belonging to transmission lineages, and the expected number of transmission lineages carrying such samples. Additionally, we used https://github.com/appliedmicrobiologyresearch/Influenza-2016-2017 to test the co-occurence of samples from each of the two tested categories within the same lineage.

To test whether transmission lineages are preferably seeded by IDUs, we constructed two match-paired datasets. First, we sorted all transmission lineages by the earliest diagnosis in the lineage. Then, in this sorted list, we marked lineages as IDU-founded and HET-founded based on the earliest sample and selected time-matched pairs of IDU- and HET-founded lineages such that in a consecutive row of lineages with the same founder type, the earliest lineage was taken; this ensured that IDU-founded lineages could not be overrepresented due to the population structure being mostly comprised of IDUs in the 1990s. We compared the time between the date of the LCA and the date of the first diagnosis in a lineage using the paired Wilcoxon test.

Second, we sorted all lineages by the median date of diagnosis, selected the ones with both HET and IDU samples, and compared the dates of diagnosis of the earliest HET in odd lineages and the earliest IDU in even lineages, again using the paired Wilcoxon test.

Supplementary materials for this Chapter are provided in Appendix A.

3.3. Results

3.3.1. The Oryol epidemic is largely constituted by the A subtype of HIV-1

We sequenced the *pol* region fragment of 858 HIV-1 samples obtained from 768 unique patients in Oryol Oblast ("Oryol dataset"). This dataset covers more than a third of the HIV-positive population in Oryol Oblast and represents an unbiased and well-annotated dataset (Fig. 3.1, Supplementary Fig. A-1). Subtype composition in the Oryol dataset was more homogeneous compared to the non-Oryol Russian dataset represented by Genbank samples ("non-Oryol dataset") (Fig. 3.1). Subtype A was the dominant subtype in Oryol Oblast (87.2%) recapitulating the historical Russian trend [324], followed by CRF63 (7.20%) and subtype B (2.49%). The transmission route was reported for 96% of the samples in the Oryol dataset, compared to less than 50% in the non-Oryol dataset; in both datasets, heterosexual transmission (HET) was the most prevalent, followed by transmission associated with injective drug users (IDU) and men who have sex with men (MSM), although the fraction of IDU and MSM samples was much higher in the non-Oryol samples. The male-to-female ratio was the same in both datasets (1.30), although sex has been reported for only 62% of non-Oryol samples. The fractions of sexes and reported transmission routes in the Oryol dataset were representative of those in the Oryol Oblast as a whole as reported by the Oryol AIDS center (Supplementary Fig. A-1).

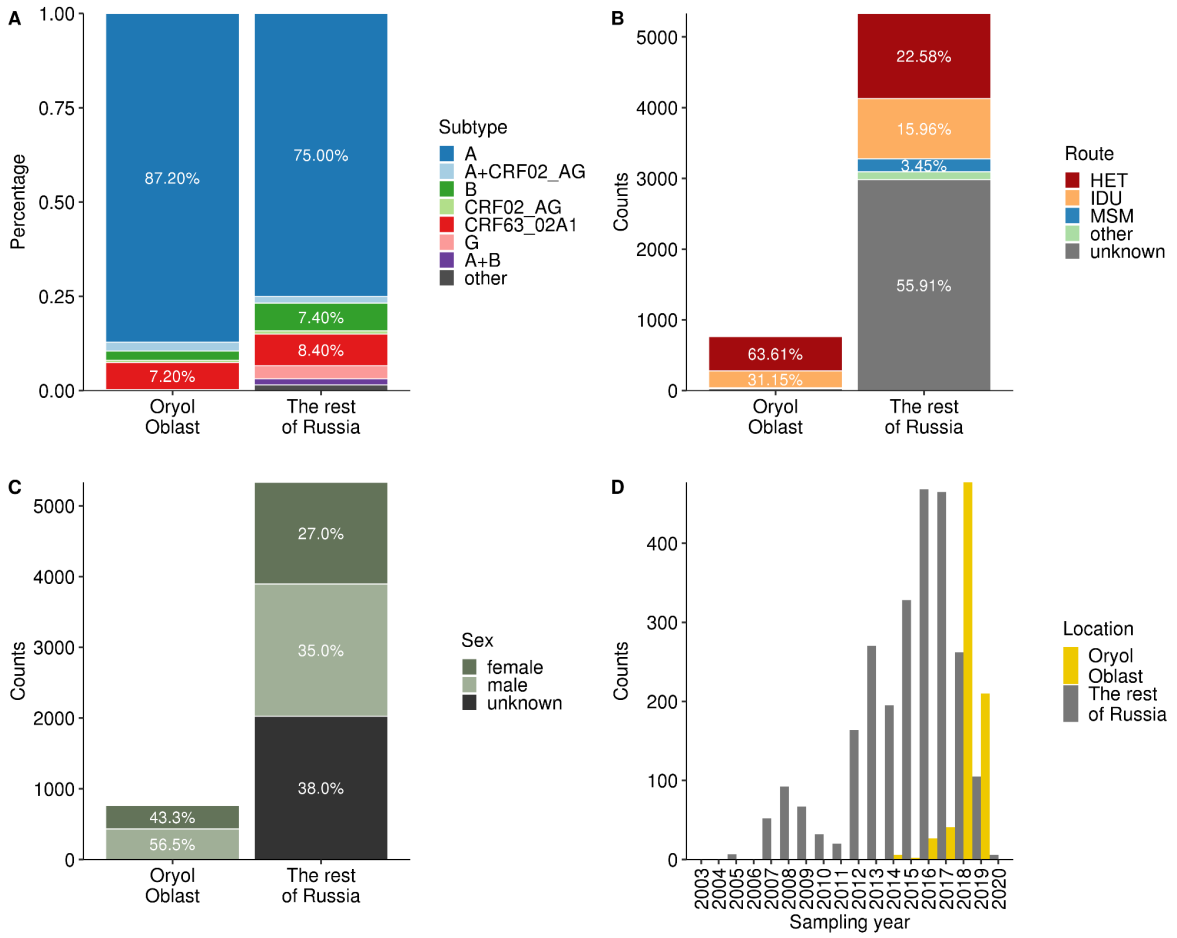


Figure 3.1. Statistics on the Oryol and non-Oryol datasets. The plots show the distribution of samples across subtypes (A), transmission routes (B), sexes (C), and sampling years (D). The seven subtypes most frequent in Russia are shown in A. Only sequences with complete sampling dates analyzed are shown in D.

The phylogenetic tree reconstructed for the combined Oryol and non-Oryol dataset had separate clades corresponding to the three most abundant variants — subtypes A and B and variant CRF63, indicating that subtyping was mostly unambiguous (Fig. 3.2). In subsequent analysis, we focused on these three variants, covering a total of 739 Oryol Oblast samples from distinct patients.

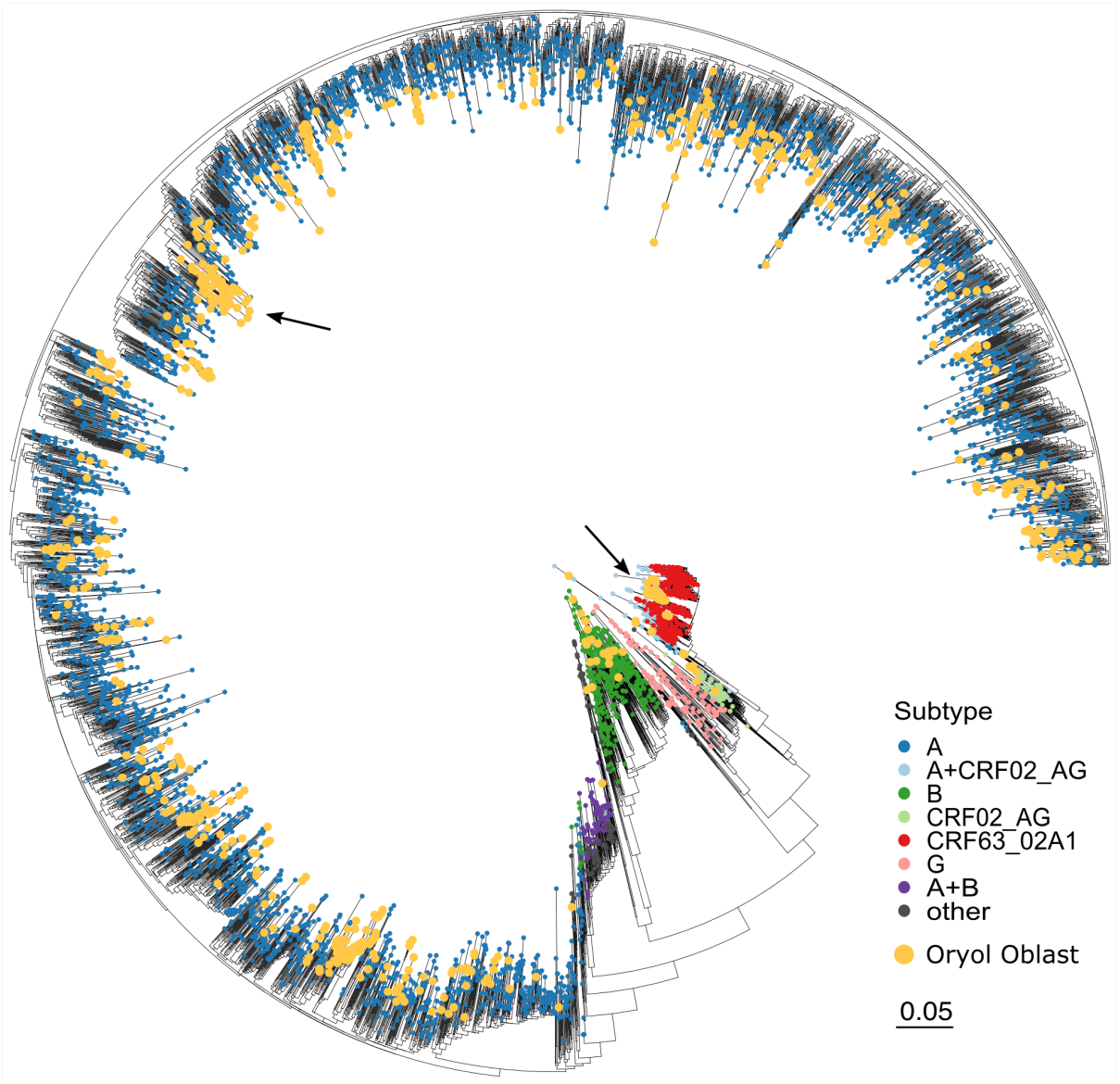


Figure 3.2. Phylogeny of the combined Russian HIV-1 dataset. When more than one sample per patient was available, only the earliest sample was used. Yellow dots mark Oryol Oblast samples. The two largest clades analysed separately, A and CRF63, are indicated with arrows. See Supplementary Fig. A-2 for the phylogeny including all samples from repeatedly sequenced patients.

3.3.2. HIV-1 has been imported into Oryol Oblast hundreds of times

To understand the interregional transmission routes of HIV-1, we first reconstructed separate phylogenetic trees for subtypes A and B and CRF63 (Supplementary Fig. A-3). Overall, the Oryol dataset samples of each of the three variants were rather scattered across

the Russian phylogeny (F_{st} = 0.01, 0.03, and 0.40 for variants A, B, and CRF63, respectively), supporting intensive transmission between regions. Still, many Oryol samples clustered on the phylogeny with other Oryol samples, indicating that many of the infections occurred within the region (Fig. 3.2, Supplementary Fig. A-2,3).

We identified introductions of HIV-1 into the region as described in 3.2.8. We attributed 489 of the 739 Oryol samples to a total of 82 imports each resulting in one or more inferred transmissions within Oryol Oblast ("Oryol transmission lineages"). The remaining 250 sequences each resulted from its own import ("singletons"), for a total of 332 imports (Fig. 3.3). CRF63 was the most homogeneous variant with 94.5% (52/55) of the samples resulting from a single import, in agreement with its substantially higher F_{st} compared to subtypes A and B. We found seven non-Oryol samples descended from Oryol transmission lineages, indicating exports.

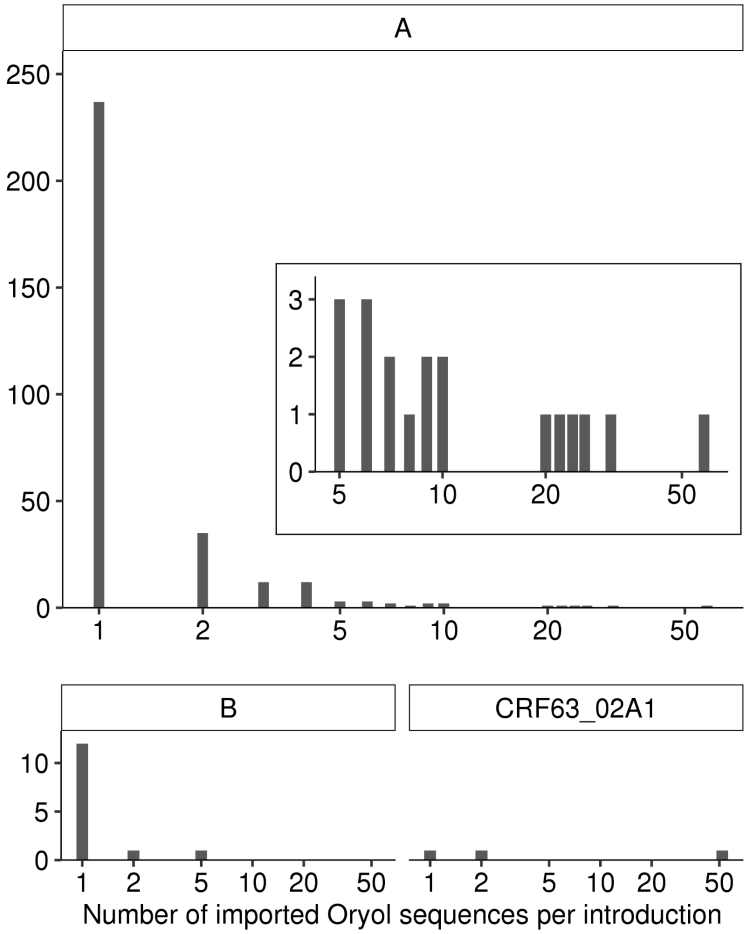


Figure 3.3. The number of Oryol sequences per import. The inset plot in A magnifies imports of subtype A of size 5 and more.

The number of imports is robust to the number of non-Oryol sequences available, already reaching a plateau when a comparable number of Oryol and non-Oryol sequences is used (Fig. 3.4A). This means that we estimate the minimal number of imports resulting in the sampled Oryol diversity robustly (at ~332). Conversely, the number of imports depends strongly on the number of Oryol samples available (Fig. 3.4B).

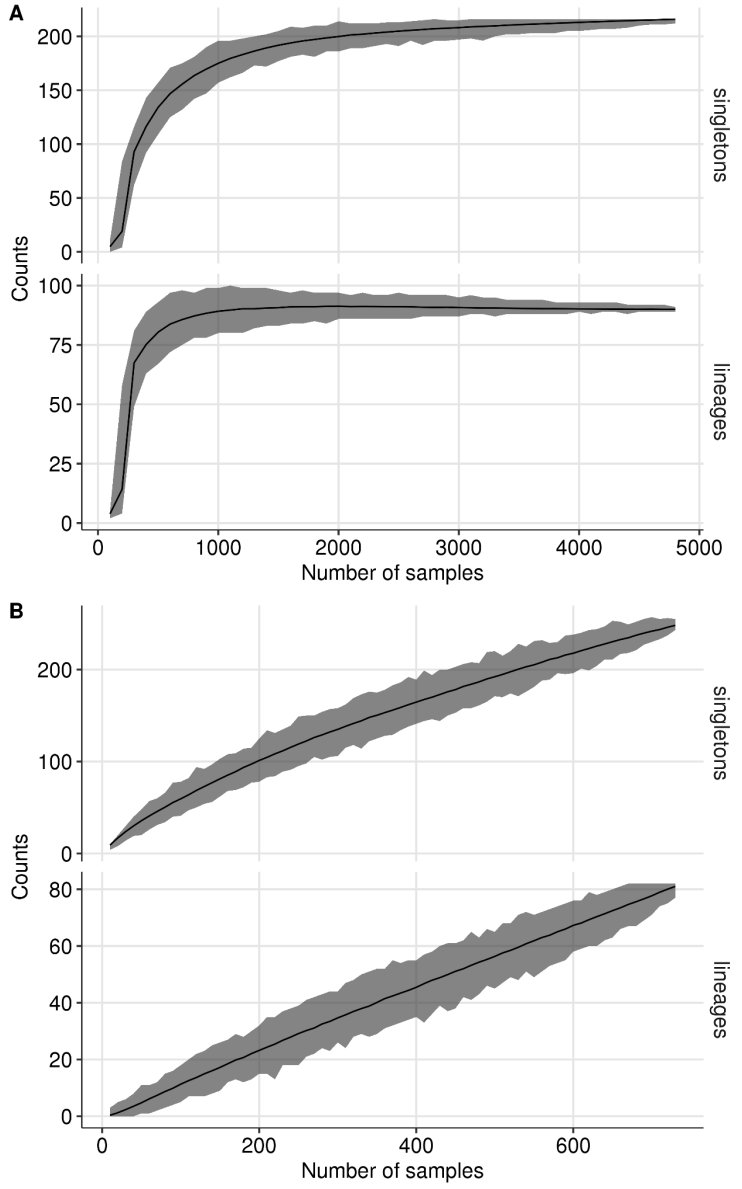


Figure 3.4. The dependence of the inferred number of singletons and transmission lineages on the number of non-Oryol (A) and Oryol (B) sequences used. The gray area reflects the range of values obtained in 1,000 subsampled sequence sets; the black line shows the mean value.

3.3.3. Early imports disproportionally contributed to the epidemic in the region

To better understand the dynamics of transmission lineages, we dated the last common ancestor (LCA) of each lineage in subtype A and CRF63 as described in 3.2.7 (Fig. 3.5). The reconstructed LCAs for individual lineages dated between 1996 and 2018,

indicating that the genetic diversity within the currently sampled transmission lineages has accumulated over decades. On average, the first positive immunoblot for a lineage was obtained 0.91 years after this lineage was established based on the LCA date estimate, although the variance of this value was very high (Supplementary Fig. A-6), in part due to lineages established after the first HIV-1 diagnosis in the lineage suggesting transmission of a non-basal variant. Lineages with earlier LCAs tended to have earlier first IB (Supplementary Fig. A-7), validating this approach. Such lineages were also larger (linear regression p-value for the LCA date is 10^{-4}; Supplementary Fig. A-8A), resulting in a disproportionate number of infections. Indeed, 50% of the lineages were established before May 2010, but they were responsible for 70% of the observed cases (Supplementary Fig. A-8B).

3.3.4. Epidemiological parameters of the subtype A sub-epidemic

Since subtype A represents 87% of all HIV-1 cases in Oryol Oblast in our dataset, we focused on this subtype to estimate the dynamics of the HIV-1 epidemic in this region. We used BEAST2 to infer epidemiological parameters of the subtype A sub-epidemic by simultaneously analyzing all transmission lineages and singletons of subtype A. The recently developed multi-tree implementation of BEAST2 [263] allows treating separate lineages as realizations of the same epidemiological process whose parameters can be jointly inferred. We modeled the subtype A sub-epidemic as a birth-death process with a constant time-independent reproductive number, constant rate of becoming noninfectious, and time-dependent multidimensional sampling proportion upon which we put a strong prior (see 3.2.9). The prior on sampling proportion was estimated based on a set of samples that were sequenced less than a year after the initial diagnosis (denoted as ‘Dataset I’ in Fig. 3.6). We used this prior both for the dataset comprising only early infections (Dataset I) and for the full

set of samples (Dataset II); the results did not differ drastically (Fig. 3.6, left vs. right column). The median effective reproductive number (R_e) was 2.8 and 2.6 for datasets I and II, respectively; the corresponding rates of becoming uninfectious were 0.37 and 0.20. The 95% HPD for R_e inferred from the full Dataset II is strictly higher than 2, implying a growing epidemic. R_e estimated from the reported yearly incidence by EpiEstim is also above 1, although it is lower than that obtained using the birth-death model (Supplementary Fig. A-14).

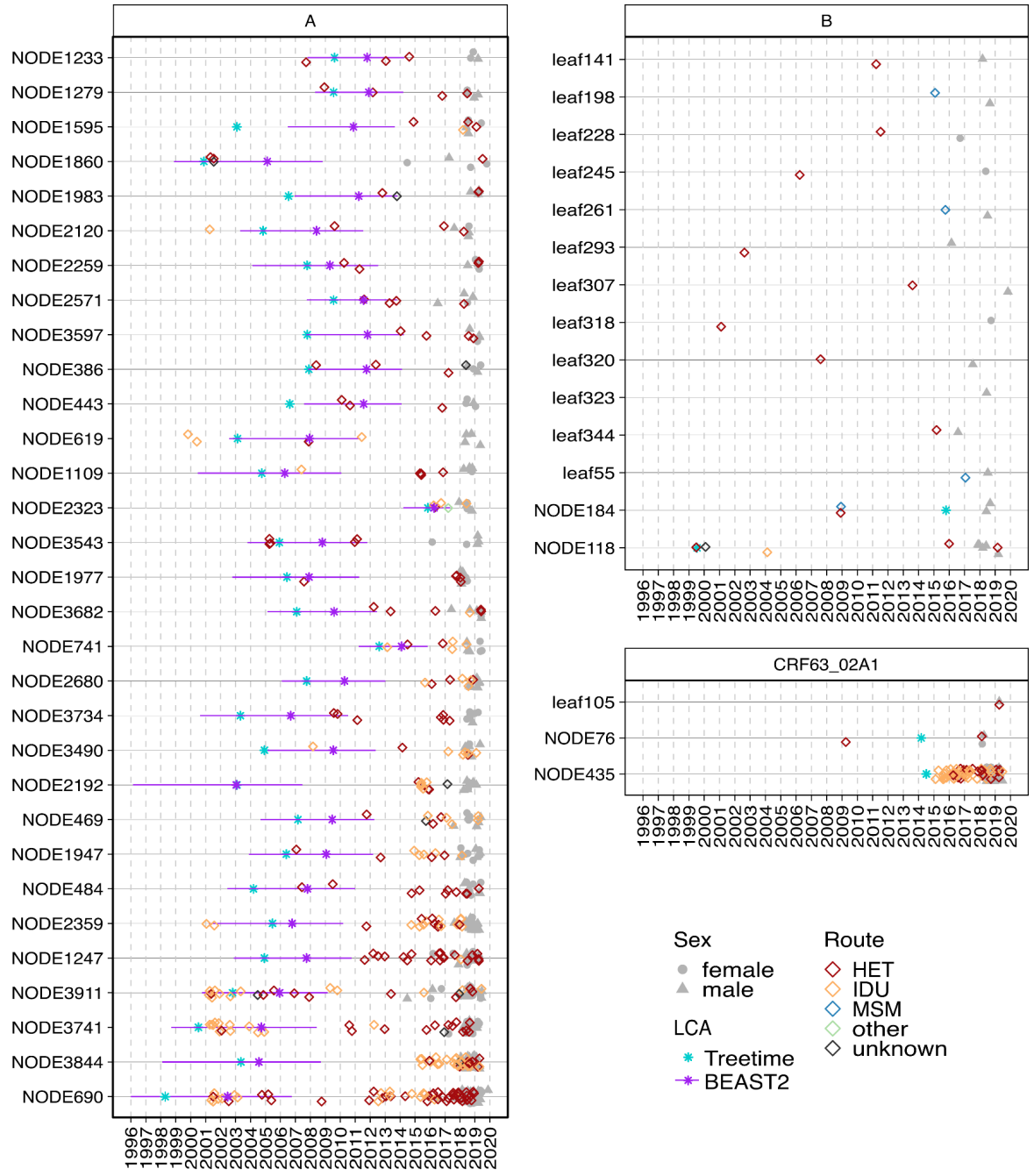


Figure 3.5. Temporal dynamics of Oryol HIV-1 lineages. Each horizontal line corresponds to a transmission lineage (with the "NODE" prefix) or a singleton (with the "leaf" prefix). Individual patients are shown twice: as an empty diamond at the date of diagnosis, with the color indicating the reported transmission route; and as a grey circle (for females) or triangle (for males) at the date of sampling. Only samples with complete sampling dates are shown. Cyan asterisks show the lineage LCA dates estimated by Treetime. For subtype A, purple asterisks and lines show the lineage LCA dates and 95% HPDs estimated by the multi-tree birth-death analysis. For subtype A, transmission lineages with at least four Oryol samples are shown. See Supplementary Fig. A-4 for all subtype A lineages and singletons, and Supplementary Fig. A-5 for all lineages and singletons sorted by the date of diagnosis.

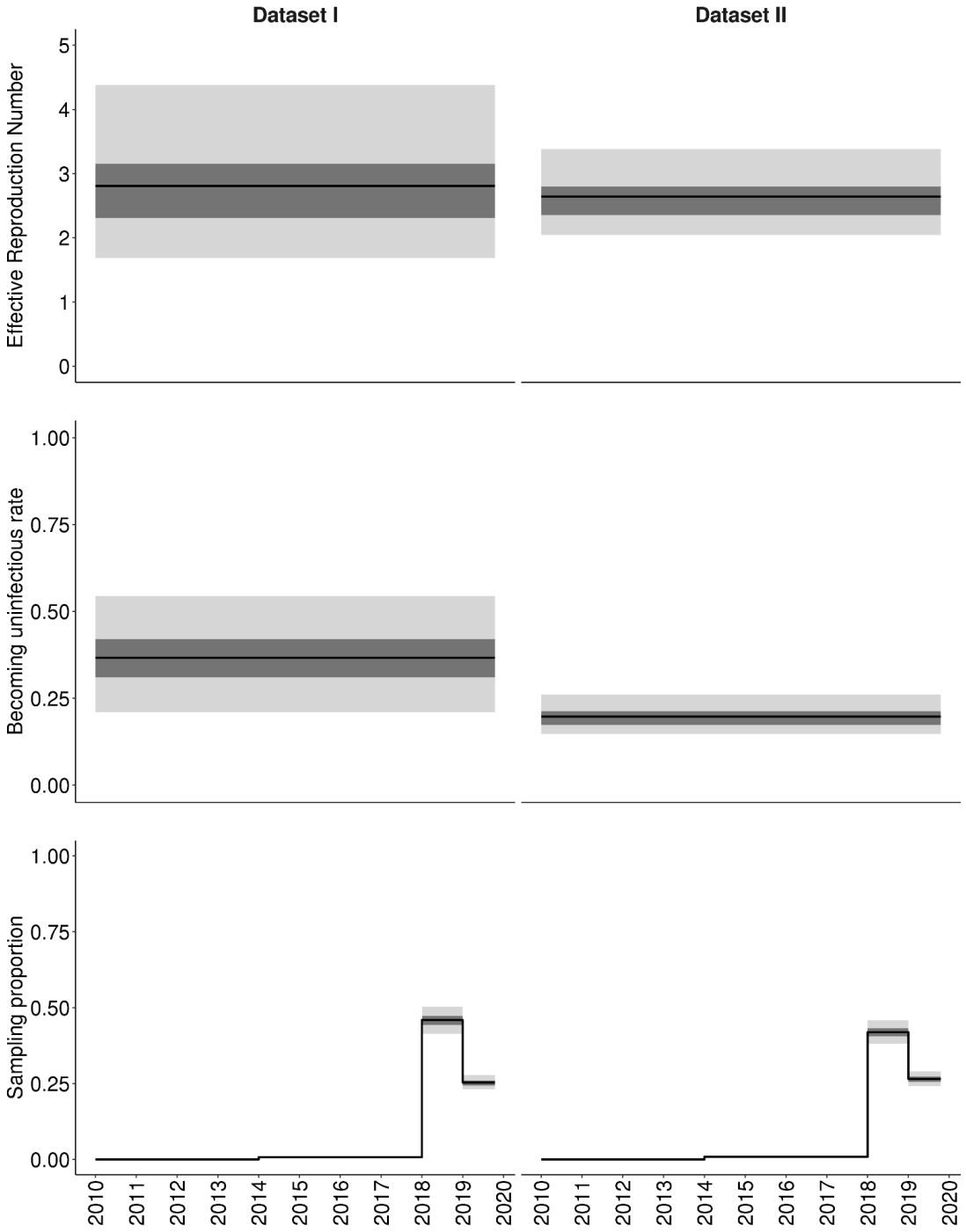


Figure 3.6. Epidemiological parameters inferred for the subtype A sub-epidemic. Dataset I, samples collected within a year of HIV-1 diagnosis; Dataset II, all samples. Solid line, median value; dark grey and light grey, 50% and 95% HPDs, respectively.

3.3.5. Bayesian phylogenetic analysis indicates the rapid growth of the CRF63 lineage in Oryol Oblast

The major CRF63 lineage (NODE435 in Fig. 3.5) is unexpectedly large for its age (Supplementary Fig. A-8B). Indeed, in less than ten years, it has reached the same size as the largest lineage of subtype A (NODE690 in Fig. 3.5) which has circulated for at least two decades. Consistently, the phylogeny of the CRF63 samples is characterized by a more recent LCA and a higher fraction of multiple merger events compared to the phylogeny of the A samples obtained at similar times (Fig. 3.7), suggesting a more rapid spread of CRF63. To test whether the two lineages indeed differ in their dynamics, we used two approaches. First, we utilized the coalescent approach to infer the growth rate of both lineages assuming logistic growth (Fig. 3.8A, Supplementary Fig. A-9). Second, we used the BDSKY model to directly infer epidemiological parameters, i.e. the reproduction number and the rate of becoming uninfectious (Fig. 3.8B). Importantly, for the largest clade of subtype A, BDSKY produced R_e estimates similar to those of a multi-tree BEAST2 implementation for multiple introductions of A (Fig. 3.6), indicating the robustness of R_e estimates. Both the coalescent and the BDSKY models suggest a higher growth rate of the CRF63 lineage compared to the largest lineage of subtype A (Fig. 3.8).

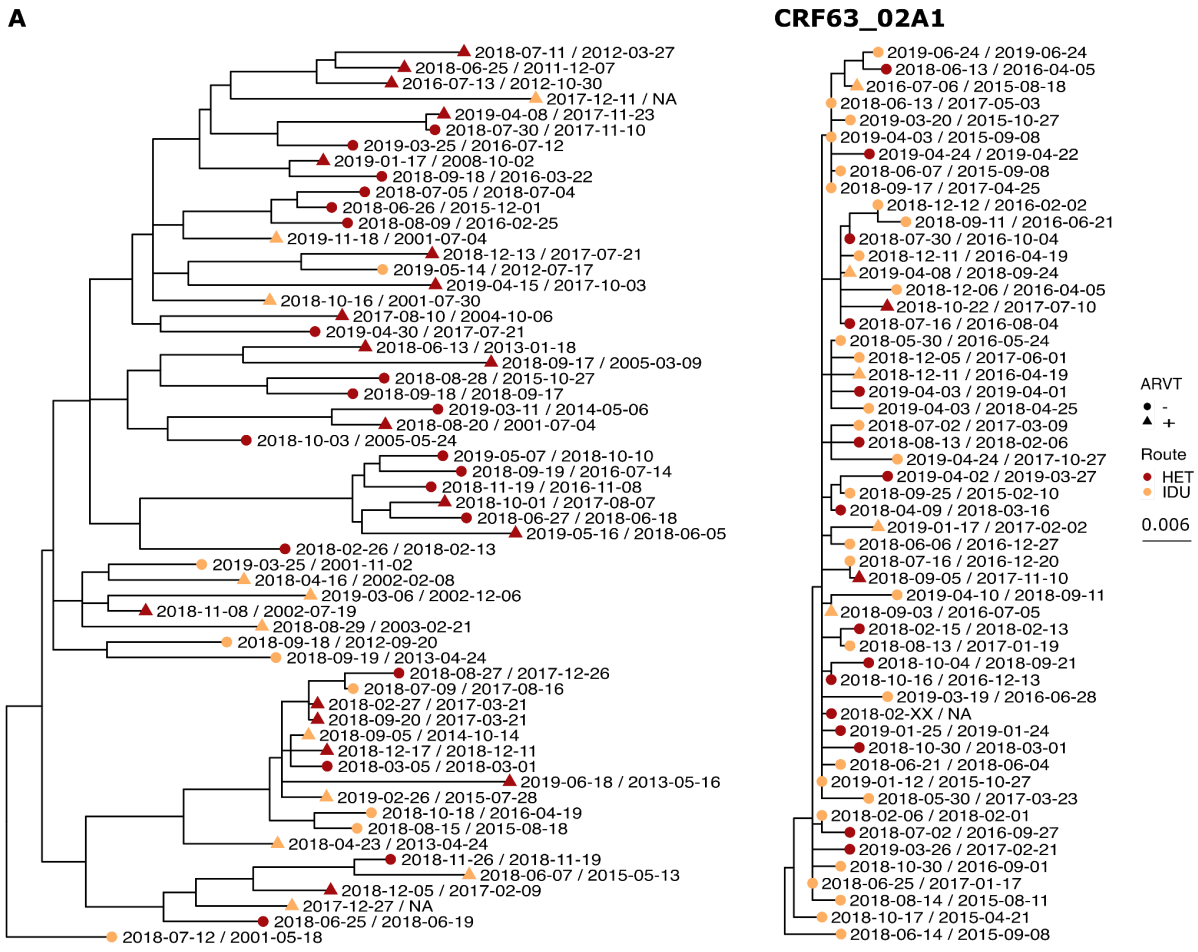


Figure 3.7. Maximum-likelihood phylogenies of the largest lineages of subtypes A and CRF63. The color indicates the reported transmission route, shape reflects the presence or absence of antiretroviral therapy at some point during the infection. The first and the second dates for each sample correspond to the date of sampling and the date of diagnosis, respectively.

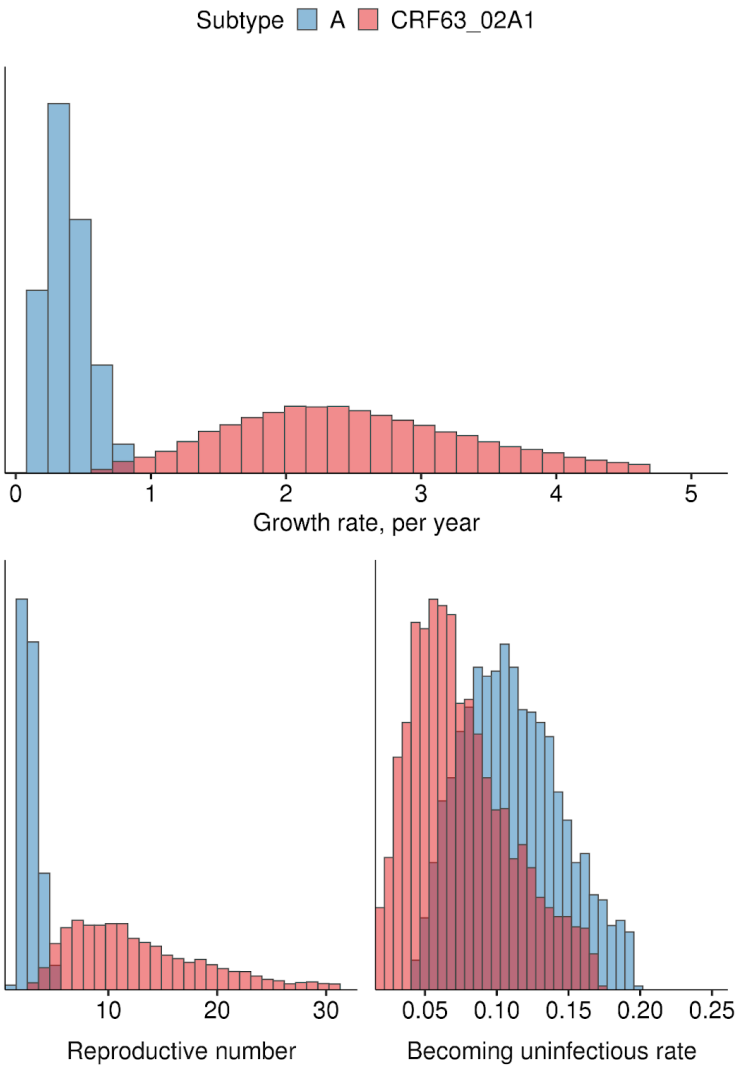


Figure 3.8. Phylodynamic inferences for the two transmission lineages shown in Fig. 3.7. A. 95% HPD of the growth rate parameter in the coalescent logistic growth model. B. 95% HPD of the reproductive number and the rate of becoming uninfectious inferred by the BDSKY model.

3.3.6. No evidence for the preferred mechanism of transmission at the origin of lineages

In the late 1990s to early 2000s, most newly reported HIV-1 infections were among IDUs (Supplementary Fig. A-10). Consistently, we find that most of the early lineages were first sampled in IDUs (Supplementary Fig. A-11), suggesting that they were founded by

them. However, we see no evidence for IDUs being more likely to be the originators of lineages when we control for the dates of lineage origins. Indeed, in the controlled matched-pair datasets, there is no difference in time from LCA to the earliest diagnosis between IDU- and HET- founded clusters, nor is there a preference in transmission route of the earliest sample in the lineage (Supplementary Fig. A-12). This means that the high prevalence of IDUs in lineage origin reflects the temporal shift in the outbreak composition in our data rather than higher inherent spreading by IDUs.

3.3.7. Distribution of gender and transmission route categories across import lineages

We next studied whether transmission lineages in subtype A are associated with transmission route and/or sex categories by comparing lineages and singletons (see 3.2.11). We did not observe any overrepresentation of males within lineages (Fig. 3.9A) as described previously in other countries [185–187,325]. However, IDUs were overrepresented within lineages, with more samples belonging to lineages than expected by chance (Fig. 3.9B). Furthermore, there were fewer lineages carrying IDUs than expected randomly (Fig. 3.9D); in other words, an IDU was more likely to fall into a lineage with another IDU ($p = 0.0007$). No such difference was observed for males (Fig. 3.9C, $p = 0.9923$).

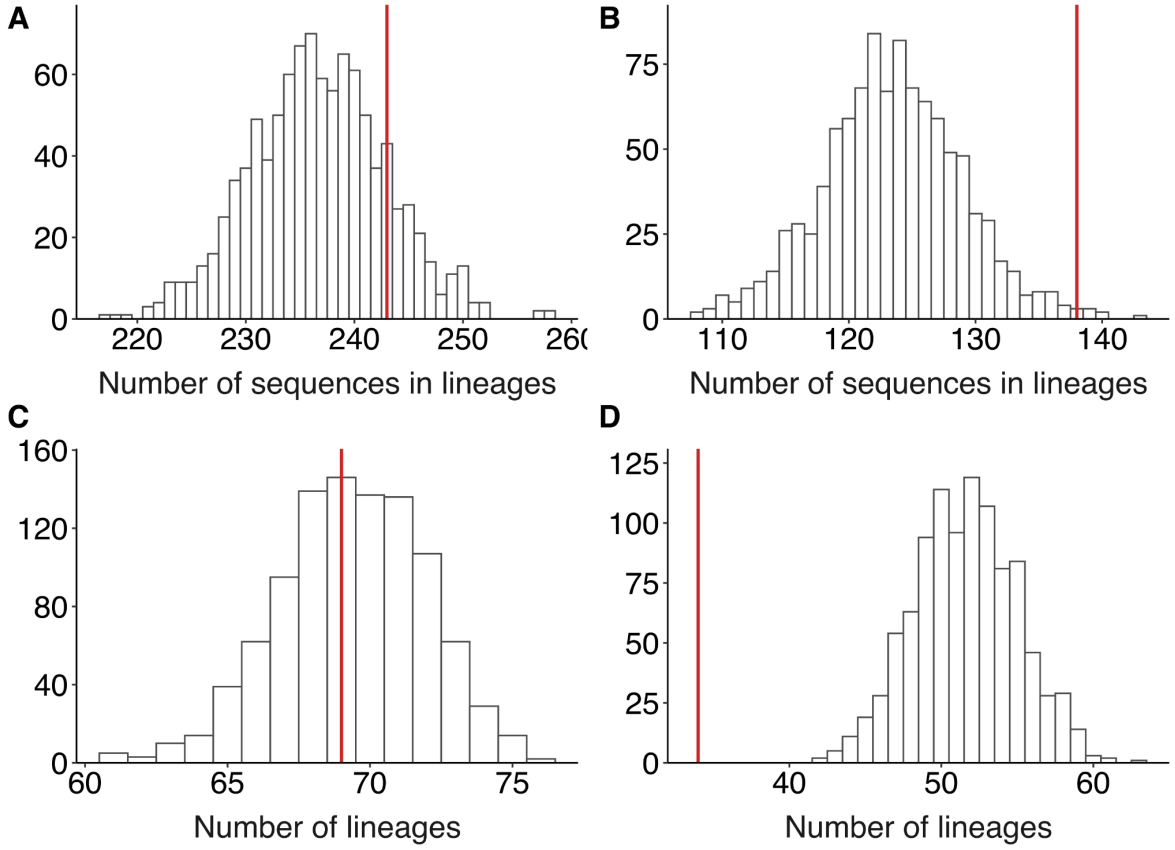


Figure 3.9. IDUs, but not males, are overrepresented within lineages. The expected distribution of the number of clustered sequences and the number of clusters carrying (A, C) males or (B, D) IDUs within subtype A. For variants B and CRF63, see Supplementary Fig. A-13.

3.3.8. MSM transmission route is underreported in Oryol Oblast

The lack of overrepresentation of males in transmission lineages is perhaps unsurprising given the heterosexual nature of the Russian HIV-1 epidemic. Still, we might expect an excess of MSM reported transmission in subtype B which is mainly associated with the MSM transmission route worldwide [156] and in Russia [326,327]. Indeed, the male-to-female ratio in subtype B is much higher than that in subtype A where we don't expect an MSM-associated bias (16/3 vs. 369/296, Fisher's exact test p=0.0169), and subtype B carries 5 out of 6 MSM-associated samples in our dataset (Fig. 10).

Is MSM transmission adequately reported in our data? Underreporting of this

transmission route can be estimated from the sex ratio among those samples for which other transmission routes are reported. Among the 14 subtype B samples with the non-MSM reported transmission route, only 3 came from females. Based on the sex ratio in subtype A, the number of males expected from this number of females is ~4. In fact, however, 11 males are observed. While the difference from the expected sex ratio is not statistically significant (Fisher's exact test, p=0.1054), if confirmed, such an excess of males would correspond to a ~2.4-fold underreporting of the MSM transmission route.

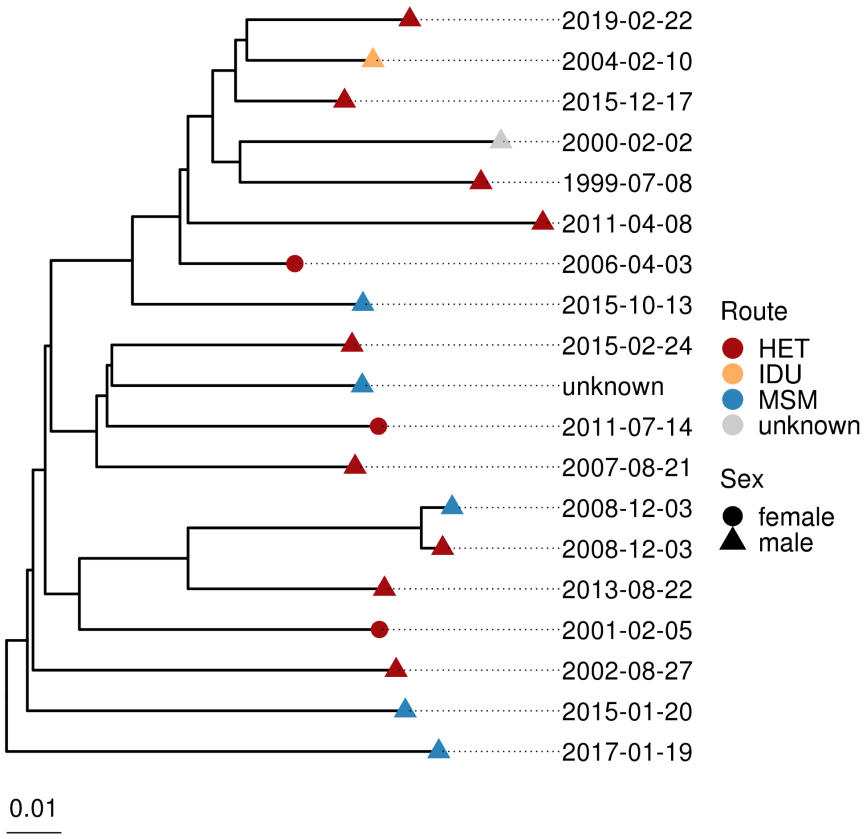


Figure 3.10. Phylogenetic tree of subtype B in Oryol Oblast. HIV-1 diagnosis dates are shown at tips.

3.4. Discussion

Since its beginning in the 1980s, the HIV-1 epidemic in Russia has been growing due to poor public awareness, stigmatization of the key risk groups, and insufficient funding [328–331]. Molecular epidemiology is informative of the characteristics of the epidemic, such as its genetic composition and reproductive number; it can also shed light on details of individual outbreaks, e.g. by identifying an increased transmission risk among a certain group or revealing a rapidly growing transmission cluster. However, molecular epidemiology methods work poorly when sampling is low, which is the case for the Russian epidemic where less than 1% of reported cases is accompanied by genetic data. Moreover, HIV-1 genetic data available in Genbank is partially obtained through target studies focusing on specific transmission routes [98,300,326,332] or outbreaks [333], making it unrepresentative of the epidemic as a whole.

Unfortunately, given the magnitude of the HIV-1 epidemic in Russia, it is infeasible to rapidly achieve sufficient genetic coverage for the entire country. Here, we instead considered a single geographic region of Russia. We focused on Oryol Oblast, a relatively small region of Russia with a total population of 736,483 [334] and a registered HIV-positive population of 2,157 (Supplementary Fig. A-1) as of 2019. The registered HIV-1 incidence is lower than that in Russia as a whole (0.29% vs. 0.73%). The relatively small size of the HIV-positive population allowed us to attain representative coverage of the local epidemic.

The dataset analysed in this work comprises sequences collected from 768 patients, or 36% of the registered HIV-positive population. Compared to non-Oryol Russian sequences available from Genbank, the Oryol dataset is unbiased, better annotated, and more up-to-date (Fig. 3.1). It also agrees well with the official sequence-independent statistics reported by the Oryol Regional Center for AIDS (Supplementary Fig. A-1). The fraction of IDUs in our

dataset is slightly below that in the official statistics (31.2% vs. 37.0%) which might be explained by lower adherence to AIDS center visits and/or lower survival among the members of this risk group [335]. In line with the official statistics, our dataset also captures the change in the predominant transmission route from IDU to heterosexual transmission in the 2000s, followed by an abrupt increase in the fraction of IDUs since 2014 when a novel designer drug became widespread [336] (Supplementary Fig. A-1,10). The male-to-female ratio is also similar to that in the official statistics (1.7 vs. 1.5) and is close to the country-wide ratio of 1.6 (as of 31.12.2019; [295]).

The subtype composition in Oryol Oblast is more homogeneous compared to non-Oryol Genbank (Fig. 3.1A), due to interregional differences and/or targeting of specific subtypes in previous studies. Still, both in Oryol Oblast and in Russia in general, the most abundant subtype is A which has historically dominated the HIV-1 epidemic in Russia [324]. Subtype A is genetically diverse and divided into several clades, or sub-subtypes. Initially, Russian sequences belonging to subtype A were annotated as A1, the most widespread subtype A clade. Recently, however, this 'Russian A1' was demonstrated to be genetically different from the African A1 [337] and received its own identifier A6. We kept the broader 'A' identifier in the text for consistency with the results produced by the SierraPy subtyping tool that we used; another widely used subtyping tool that we checked, REGA [338], utilizes an old set of reference sequences and still annotates the Russian subtype A as A1.

For the three most abundant variants (A, B, and CRF63), we used the reconstructed maximum likelihood phylogenetic trees to infer imports into Oryol Oblast. Our approach is based on the assumption that all imports are of Russian origin. While some imports could also directly come from other countries, such cases are probably rare as transborder travel is expected to be much less intense than within-country travel.

The inferred number of transmission lineages was robust to the number of non-Oryol sequences used, meaning that we have successfully resolved the sequenced genetic diversity of HIV-1 in Oryol Oblast (Fig. 3.4A). However, this number was strongly dependent on the amount of Oryol samples available: it did not saturate as we increased the number of Oryol samples in the analysis (Fig. 3.4B). This implies that the diversity of HIV-1 in Oryol oblast is higher than can be captured by sequencing of a third of the population, and/or that there is a constant import of novel HIV-1 variants into the region.

Lineages that were established early contained more samples, implying that early imports into the region significantly contributed to the current epidemic. Early imports were mostly associated with IDUs; however, we did not find evidence for preferred seeding of lineages by IDUs (Supplementary Fig. A-12), suggesting that the prevalence of different transmission routes was shaped by social factors rather than the biology of transmission.

Nearly two-thirds of all samples were attributed to transmission lineages. The remaining "singleton" sequences did not result in observable transmission within Oryol. A fraction of singletons could actually correspond to non-Oryol residents. Indeed, in Russia, the HIV data collected by an AIDS center is household-based, meaning that patients are assigned to centers based on their household registration. The attribution of sequences comprising Oryol transmission lineages as actually coming from Oryol Oblast residents is probably more reliable.

Using phylodynamic approaches, we characterized the epidemiological parameters of HIV-1 spread. These analyses have several limitations. First, we assumed that the evolutionary rate is independent of the age of the infection and the presence of therapy. Differences in infection duration and therapy between patients could have affected the estimated LCA dates. For example, the rate of evolution can be reduced by antiretroviral

therapy [339], pushing the LCA estimates to the present. Such biases could lead to some unexpected LCA datings that we observe. For example, in several instances (e.g. NODE619 and NODE2359 in subtype A or NODE184 in subtype B), two or more of the earliest diagnoses in a transmission lineage had earlier dates than the reconstructed LCA of this lineage, which is impossible as LCA by definition must correspond to the earliest transmission between samples of the lineage. We speculate that this discrepancy could have been caused by differences in the antiretroviral therapy status between sampled individuals; indeed, while ~40% of our dataset was on therapy, one of the two earliest samples in both NODE619 and NODE2359 lineages (of subtype A) and both patients in the NODE184 lineage (of subtype B) receive(d) therapy (Fig. 3.5). In theory, it may be possible to account for differences between patients in viral evolution rates, e.g. by inferring two distinct evolutionary rates, but the amount of variation in duration and adherence to therapy would be hard to account for. Overall, our results seem qualitatively robust to the choice of evolutionary rate.

Second, birth-death models are unidentifiable unless at least one of the parameters is fixed or strongly constrained. We put a strict prior on the sampling proportion, defined as the fraction of all infections being sampled. Sampling density was strongly non-uniform across years, with the vast majority of samples collected over just two years (2018-2019); however, many of these patients became infected years ago. We thus constructed a four-dimensional prior on sampling proportion based on the subset of "rapidly sequenced" samples (Dataset I, see 3.2.9), and used it for datasets that also included old infections and for which we could not make an informative assumption about sampling. While the parameters estimated from different datasets were similar (Fig. 3.6), those estimated based on Dataset I should probably be considered more reliable.

Third, our BDSKY analyses assume that sampling of infection results in its "death" due to host recovery or change of behavior. Unfortunately, this assumption usually does not hold for long-term infectious diseases such as HIV-1, especially in countries like Russia where treatment is available to less than 50% of HIV-positive people [296]. This may make our inferences about the reproduction number and the rate of becoming uninfectious under- and overestimated, respectively.

Fourth, the obtained phylodynamic estimates are relevant for the identified part of the Oryol Oblast epidemic. The unidentified part of the epidemic, i.e., associated with non-registered cases (which may comprise up to 20% in Russia [340]), can grow more rapidly due to lack of awareness of the HIV-1 diagnosis [341] and a definite absence of therapy among these people.

Fifth, our analyses in the main text use all sites, including those of drug resistance mutations (DRMs). Changes at such sites are frequently recurrent between patients, and therefore may obscure phylogenetic analyses. Although drug-resistance mutations were reported not to bias the composition of transmission lineages [342], they were shown to affect some analyses, e.g. the exact reconstructed history of transmissions [343]. To address this, we repeated the maximum-likelihood tree reconstruction, definition of transmission lineages (Supplementary Fig. A-15), the birth-death analysis on subtype A (Supplementary Table A-5), and the comparison of the two largest clades in variants A and CRF63 analyses (Supplementary Fig. A-16) on DRM-masked alignments. We found that the inclusion of DRM sites did not affect our conclusions.

With these limitations in mind, we assessed the characteristics of the epidemic. As subtype A currently prevails in Oryol Oblast, epidemiological parameters of its sub-epidemic should tone the dynamics of HIV-1 in the region. We jointly inferred R_e and the rate of

becoming uninfectious for all imports of subtype A. The estimates produced from both Dataset I and Dataset II covering subtype A sub-epidemic consistently inferred R_e above 1 (median values are 2.8 and 2.6, Fig. 3.6), implying a growing epidemic. For the above-mentioned reasons, these values should be considered as lower boundaries; the actual growth rate is probably higher. The rate of becoming uninfectious is estimated as 0.37 and 0.20 for datasets I and II, which is equivalent to the total duration of infection of 2.7 and 5 years, respectively.

Independently, we inferred the R_e dynamics from case count data using EpiEstim (Supplementary Fig. A-14). These estimates of R_e were lower, compared to the birth-death model, although still strictly higher than 1. There can be at least two reasons for this discrepancy. First, we only have incidence data starting from 2000, and the sliding window approach implemented in EpiEstim does not allow us to obtain R_e estimates for years before 2005. If R_e in the 1990s and early 2000s was in fact higher than that later in the epidemic, the EpiEstim estimate would not be reflective of the epidemic as a whole and would be an underestimate. Second, R_e can be biased by heterogeneity in sampling procedure; for instance, if the epidemic shifts to the heterosexual population but infected heterosexuals are diagnosed more slowly, the count-based R_e estimate would be lowered.

Besides subtype A, we separately studied CRF63 which recently evolved in the Siberian part of Russia as a result of recombination between CRF02 and subtype A6 and has been spreading across the regions of Russia since then. We show that in Oryol Oblast, it is mostly represented by a single introduction event that resulted in 52 identified infections. This transmission lineage was unexpectedly young for its size, motivating us to compare epidemiological dynamics of subtype A and CRF63. We compared the two largest clades of these variants using two phylodynamic approaches. First, we inferred the growth rate of both

clades assuming logistic growth; second, we inferred the reproduction number and the rate of becoming uninfectious using a birth-death model and a sampling prior used for multi-tree subtype A analysis. Both analyses indicate a higher growth rate of the CRF63 clade.

The rapid invasion of CRF63 in Oryol Oblast is consistent with the fact that this variant has been rapidly expanding in several Russian regions in recent years, and has become the dominant variant in some of them [344–346]. The differences in the rate of spread of A and CRF63 could result from biological and/or epidemiological differences between these two variants. CRF63 has originated from CRF02 and A6 [347]; different properties of CRF02 ([164,348], but see [349]) from the major Russian strain A6 may be responsible for CRF63 success, although experimental studies are required. Indeed, while the biological properties of CRF63 have not yet been studied extensively, this strain has been reported to cause a different immunological response [350]. There are also epidemiological differences between A and CRF63. CRF63 was introduced into Oryol Oblast much later than subtype A; thus, CRF63-associated infections are expected to be younger even when controlling for the year of diagnosis. Furthermore, a smaller fraction of CRF63-infected patients were reported to receive therapy (16% vs. 40% in our dataset, Fig. 3.7); this could be due to a lower rate of disease progression in CRF63 resulting in it being transmittable for longer than A6 and producing more infections, but can also be due to infections by this variant being more recent. A high fraction of IDUs in the CRF63 clade also hints at increased transmission through this group. The differences between CRF63 and A6 merit further study.

The distribution of transmission routes is informative about epidemiological patterns. We did not observe excessive clustering among males such as has been reported in many countries where the MSM-associated subtype B is prevalent [185–187,325]. This might be explained by a different structure of the HIV-1 epidemic in Russia. Compared to those areas

where most new diagnoses come from MSM contacts resulting in a majority of the HIV-positive population being male, HIV-1 in Russia heavily affects the general population through HET contacts as well as the IDUs, making the male-to-female ratio less biased (Fig. 3.1, refs on official stats on Russia).

By contrast, we do observe clustering of IDUs in transmission lineages. Part of this effect could arise from a possible fraction of non-Oryol residents outside transmission lineages (see above), e.g. if non-Oryol residents are less likely to be IDUs. However, we also observe that IDUs are more likely to co-occur within the same transmission lineages, and this bias cannot have a purely geographic nature. Therefore, the observed IDU clustering probably results from increased transmission within enclosed communities of drug users, although this partially may also come from IDU being the major transmission route in the early years of the epidemic in Russia.

The MSM transmission route was rarely (0.8%) reported in our dataset. Still, five of six MSM cases belong to subtype B in agreement with the historical association of subtype B with this route of transmission. A striking contrast between the number of males in subtype B and a low fraction of MSM probably stems from major underreporting of MSM.

In our dataset, only 39% of patients were on therapy; this fraction is lower than that reported by the Oryol AIDS center (65%). This is because a significant fraction of our dataset corresponds to recently diagnosed patients who have not started receiving therapy yet. Prior to 2016, according to government regulations, therapy was only provided to patients with CD4 counts below 350 [351]. While post-2018 recommendations propose therapy for all patients [352], and the fraction of HIV-positive people on therapy was increasing both in Oryol Oblast and in Russia in general, the funding allocated in 2019 [353] and in 2020 [354] only covered therapy for 60% and 64% of HIV-positive patients enrolled in care in Russia.

Our finding of large R_e is in line with the insufficient effectiveness of the policies currently in place to curb the HIV-1 epidemic in Russia.

Taken together, the results presented in this work offer the most in-depth molecular-based characteristic of the HIV-1 sub-epidemic within Russia. Multiple lines of evidence indicate that the HIV-1 epidemic in Oryol Oblast is clearly growing, mainly due to subtype A. We provide evidence that the frequency of the recently introduced CRF63 grows faster compared to subtype A, and suggest that this variant may eventually start to contribute more significantly to the Oryol epidemic. CRF63 should be a high-priority target for molecular surveillance and experimental studies in Oryol Oblast.

CHAPTER 4: Molecular epidemiology of SARS-CoV-2 in Russia early in the pandemic

4.1. Introduction

Russia is among the five countries with the highest number of confirmed COVID-19 cases [355]. However, the outbreak in Russia started later than in many neighboring European countries [356–358], possibly in part due to early implementation of non-pharmaceutical interventions (NPIs) limiting virus import. Early NPIs included introduction of quarantine for passengers arriving from China on January 23, 2020, closing the land border with China on January 31, cancellation of most incoming flights from China on February 1, restricting the entrance of non-Russian citizens from China on February 4, and restricting entrance from Iran and South Korea in late February [359–367]. While the earliest formally confirmed two cases in Russia dated to January and could be associated with a direct introduction from China [368], no further cases were detected until March 2, 2020, when a woman returning from Italy tested positive [369].

Nevertheless, since March 3, a steady increase in confirmed cases has started, with the initial country-wide estimated reproduction number R of ~2 [370]. Before March 21, all confirmed Russian cases were imported, while most European countries already had local transmission by this time [356,371]. Since early March, Russian regional authorities had been implementing their own NPIs. In particular, specific measures were introduced in Moscow and Saint Petersburg, the two largest transportation hubs responsible respectively for 67% [372–374] and 10% [375] of Russia's international air traffic. In Moscow, since March 5, all international travelers were temperature checked at the border; and passengers coming from countries with registered cases of SARS-CoV-2 had to report to authorities; while those

coming from countries with high case counts at the time, including China, Italy, Spain, and the UK, were quarantined [376]. Since March 14, mandatory quarantine was also applied to passengers' family members [377], and since March 16, it was introduced for all international travelers [378]. The NPIs at Saint Petersburg were timed similarly [379]. On March 13, the entrance of non-Russian citizens from Italy was restricted at the state level [380], and on March 18, entrance into Russia for all non-Russian citizens for non-emergency reasons was banned [381]. While inbound flights, mainly returning Russian citizens from abroad, were still operating as of early July 2020, passenger traffic has decreased drastically (e.g., 20-fold at the Moscow Sheremetyevo airport, the one that accepts most international flights during the pandemic [382,383]).

Here, we report the analysis of 211 SARS-CoV-2 complete genome sequences obtained in Russia between March 11, 2020 (when there were just 28 confirmed cases Russia-wide) and April 23, 2020 (when there were 62,773 confirmed cases) [384,385]. Phylogenetic analysis reveals distinct introduced lineages associated with transmission within Russia, as well as multiple individual samples phylogenetically intertwined with non-Russian sequences. The largest identified lineage corresponds to an outbreak at the Vreden hospital; phylodynamic analysis of this outbreak reveals between 2 and 3 distinct introductions and initial rapid spread curbed by subsequent establishment of quarantine.

4.2. Methods

4.2.1. Sample collection and sequencing

Nasopharyngeal and/or throat swabs were collected in virus transport media. Total RNA was extracted using the RiboPrep DNA/RNA extraction kit (AmpliSens, Russia). Extracted RNA was immediately tested for SARS-CoV-2 using LightMix ® SarbecoV E-gene plus EAV control (TIB Molbiol, Berlin, Germany) provided by the WHO Regional Office for Europe and based on Charite protocol [386]. LightMix ® SarbecoV E-gene plus EAV control was used with BioMaster qRT-PCR Kit (Biolabmix, Russia). Briefly, each 20 μL reaction mixture contained 10 μL of 2x buffer, 0,5 μL of LightMix SarbecoV E-gene reagent mix, 4.7 μL of nuclease-free water, 0.8 μL of the enzyme, and 4 μL of extracted RNA as the template. RT-PCR was performed on a LightCycler 96 RT-PCR system (Roche). The thermal cycling conditions were 55 °C for 15 min, 95 °C for 5 min, followed by 45 cycles of 95 °C for 5 s, 60 °C for 15 s and 72 °C for 15 s. Specimens with Ct values less than 30 were selected for whole-genome sequencing.

4.2.2. Ethics

Samples used in this study were collected as part of approved ongoing surveillance conducted by the Smorodintsev Research Institute of Influenza. Written informed consent was obtained from all subjects. All samples were de-identified prior to receipt by the study team. The study was presented to the Local Ethics Committee at the Smorodintsev Research Institute of Influenza. The Committee concluded (protocol #151) that the study does not make use of new identifiable biological samples and does not bring forward any new

sensitive data. Therefore, according to the rules of the Committee and national regulations this project does not require ethical approval.

4.2.3. Virus isolation

For some of the SARS-CoV-2 PCR-positive samples, viruses were isolated in Vero cell culture (ATCC #CCL-81). Cells were propagated in MEM (Gibco) supplemented with GlutaMax (Gibco), Sodium Pyruvate (Gibco), and 10% FBS (Gibco #10500). 2 days before inoculation cells were seeded in 5.5 cm^2 cell culture tubes (Nunc) at 1:4 ratio and 5% FBS. Samples were diluted 1:10 with serum-free media containing antibiotic-antimycotic (Gibco) and inoculated to cells in a volume of 0.5 ml/tube. After incubation for 2 h at 37°C, the inoculum was removed and 3 ml of serum-free media with anti-anti was added to tubes. Viruses were harvested 4-6 days post-inoculation (p.i.) when cytopathic effect (CPE) was near 80-100%, while first signs of CPE were typically observed 2-4 days p.i. For subsequent work, 0.15 ml of virus suspension was lysed in 0.5 ml RLT buffer (QIAGEN) and stored at -20°C until RNA extraction.

4.2.4. Whole-genome sequencing

RNA from primary clinical specimens and virus isolates was re-extracted using QIAamp Viral RNA Mini Kit or RNeasy Mini Kit (QIAGEN). Whole-genome amplification of the SARS-CoV-2 virus genome was performed using ARTIC Network protocol [387] with modifications. ARTIC Network primer sets were modified by adding ONT universal tags: 5'-TTTCTGTTGGTGCTGATATTGC-3' and 5'-ACTTGCCTGTCGCTCTATCTTC-3' for forward and reverse primers, respectively. 1D Ligation sequencing kit (SQK-LSK109) with

PCR barcoding expansion (EXP-PBC096) was utilized for sequencing library preparation. MinION (Oxford Nanopore) (flow cell R9.4.1) was used for whole-genome sequencing.

4.2.5. Genome assembly and consensus correction

Fast5 files produced by minION were basecalled using guppy_basecaller v3.6.0 [388]. Basecalled reads were processed by Porechop v0.2.4 [389] in two steps. First, for each sequencing run, reads were demultiplexed with default settings, with built-in barcode and adapter sequences cleaved from read ends. Second, PCR primers were trimmed from demultiplexed reads with options --end_size 70 --no_split. Processed reads corresponding to one sample were combined.

For each sample, we then mapped reads onto the Wuhan-Hu-1 SARS-CoV-2 genome sequence (NCBI ID: MN908947.3) using minimap2 v2.17 [390] with default settings and filtered out chimeric reads and reads that had secondary alignments. SAMtools-mpileup v1.10 [391] was used to produce draft consensus sequences which were then corrected as follows. Mappings were converted into .tsv files using sam2tsv [392], and for each position in the genome, we computed the frequencies of all variants present. We further considered positions with coverage 15 or higher and alternative (compared to Wuhan-Hu-1) variant frequency 50% or higher. We corrected the draft consensus sequences based on the defined set of alternative variants. Each introduced correction was assessed by visually analyzing the corresponding region of mapped reads in IGV v2.8.0 [393]. Additionally, we manually assessed all alternative variants that had coverage 100 and below. We observed several spurious mutations that were not included in final consensus sequences, including the homoplasic mutation G11083T residing at the end of the poly-T tract in the genome that was observed in five of our samples.

4.2.6. SARS-CoV-2 dataset preparation and filtering

All complete high coverage genomes of SARS-CoV-2 for all regions were downloaded from GISAID on May 26, 2020, for a total of 20,469 global sequences and 78 Russian sequences. To this dataset, we added the 136 sequences obtained in this study. Sequences shorter than 29,000 bp, sequences with more than 300 positions with missing data (Ns), sequences excluded by Nextstrain, and samples corresponding to resequencing of the same patients were removed. This led to exclusion of one Russian sample sequenced in this study (hCoV-19/Russia/Ulan-Ude-RII4560S/2020), as well as 834 non-Russian sequences.

The obtained sequences were aligned with MAFFT v7.453 [394] with the following parameters: '--addfragments --keeplength'. We utilized Wuhan-Hu-1/2019 (NCBI ID: MN908947.3) as the reference sequence. To remove low-quality bases from the alignment, 100 nucleotides from the beginning and the end were trimmed. The final alignment was used to construct the phylogenetic tree with IQ-Tree v1.6.12 [395] with the GTR substitution model and the '-fast' option. We used TreeTime v0.7.5 [318] to reconstruct the sequences of the internal tree nodes. Sequences separated from the tree root by more than ten nucleotide mutations were excluded as probable results of incorrect base calling; this included two sequences from Russia (Russia/SCPM-O-02/2020 and Russia/SCPM-O-05/2020). The final dataset contained 19,834 virus SARS-CoV-2 sequences. The resulting tree is available as [396].

4.2.7. Phylogenetic analysis

We categorized each Russian sequence into one of the five phylogenetic categories based on its phylogenetic position, defined as follows. A Russian transmission lineage is a set of two or more sequences that form a Russian-only clade. A Russian singleton is a single

Russian sequence that differs from all other Russian sequences and is not a part of a Russian transmission lineage. A Russian stem cluster is a set of Russian sequences identical to each other and some non-Russian sequences. A Russian stem-derived transmission lineage is a Russian transmission lineage whose immediate ancestor is a Russian stem cluster. A Russian stem-derived singleton is a Russian singleton whose immediate ancestor is a Russian stem cluster. These categories are schematically represented in Fig. 4.3.

As sampling dates, we used the collection dates reported in GISAID. For some samples, either the day or both the day and the month of the collection were missing. In such cases, the date was set to the latest date possible, e.g. “2020-03” to “2020-03-31” and “2020” to “2020-12-31”. The date for the lineage introduction was estimated as the earliest collection date among all samples in the particular lineage, which in fact reflects the latest possible date the lineage could have been introduced.

In phylogeographic analysis, we assumed that the possible source(-s) of introduction for Russian transmission lineages and Russian singletons were the sampling country(-ies) of the non-Russian sequences ancestral to the considered lineage; for Russian stem-derived transmission lineages and Russian stem-derived singletons, these were the non-Russian sequences identical to the ancestral stem cluster. For stem clusters, the possible source of introduction were the countries of origin for the sequences identical to those in the stem cluster. As a possible source of introduction, we only considered those countries with the earliest collection date earlier than the earliest collection date among all samples in the lineage (or than the collection date of the singleton). For patients with known travel history, we considered the country (continent) of origin as uniquely identified by phylogeography if it was either the only country (continent) on the ancestral stem or the one with the earliest collection date. For samples with no travel data, the same logic was applied, except we only

infer the country or continent of origin if it was the only one on the stem. If there were more than eight countries on the list, countries were merged into regions: Africa, Asia, Europe, North America, and South America. To study the possibility of introduction from China, we performed the same analysis but considered China, Hong Kong, and Taiwan separately from the rest of Asia.

To understand whether introductions to Russia occurred through major transportation hubs (Moscow and Saint Petersburg), we considered all Russian samples not included in Russian transmission lineages. For these samples, we calculated the branch lengths from each sample to its immediate ancestor and labeled all samples by two categories: major hubs (Moscow, Moscow region, Saint Petersburg, and Leningrad region) and other locations (samples from all other locations in Russia). On this data, we performed a permutation test, shuffling labels across the dataset 1000 times. The two-sided p-value was calculated.

4.2.8. Phylodynamics of SARS-CoV-2 in Vreden hospital

As discussed in Section 4.3.4, the Vreden samples belong to three distinct phylogenetic groups. To account for this, we constrained the phylogeny as follows: (((group1), group 2), (group 3)). We independently ran BEAST2 v2.6.2 on three datasets: (i) the whole Vreden dataset comprising groups 1, 2, and 3; (ii) groups 1 and 2; and (iii) group 1 only. The model details were as follows. The effective reproductive number was allowed to change on March 27 (which delimits the suspected out-of-hospital period) and again on April 8 (which corresponds to the introduction of quarantine). The prior on the clock rate was set to be a normal distribution with a mean of $9.41*10^{-4}$ and a standard deviation of $4.99*10^{-5}$, based on the estimates from the UK study [397]. Other priors are provided in Supplementary Table B-7. Supplementary Tables B-4, B-5, and B-6 contain the Bayesian estimates of the

model parameters for three datasets comprising groups 1, 2, and 3, groups 1 and 2, and group 1, respectively.

We ran a birth-death skyline model with multiple rho-sampling events. This sampling strategy allows us to take into account that the sampling was not continuous over time but instead was performed on specific dates (April 3, 7, 10, 14, 22).

The time-dependent R_e was independently estimated from incidence data using EpiEstim package v2.2-3 in R [322] with a 7-days sliding window and parametric serial interval distribution with a mean of 4.6 and standard deviation of 2.0.

4.2.9. Public information and data visualization

The initial Russian map was downloaded from GADM (sf, level 1) [398]. The number of confirmed cases in Russia by region was downloaded on May 26, 2020 from [399]. Patients' age data for Russian samples were extracted from GISAID metadata; for 12 samples, age data were missing. The Spearman correlation between age and collection date was calculated in R version 3.6.3 with cor.test() function. Maps were visualized with the ggplot2 v3.3.0 package in R. Phylogenetic trees were visualized with the ETE3 toolkit v2.3.2 [400] in Python v3.6 and iTOL v4 [401]. The maximum clade credibility tree was visualized with FigTree v1.4.4 [402].

Supplementary materials for this Chapter are provided in Appendix B.

4.3. Results

4.3.1. Sampling and data acquisition

Samples were obtained from hospitals and out-patient clinics as part of COVID-19 surveillance and sequenced at the Smorodintsev Research Institute of Influenza. We sequenced complete genomes of 135 samples from Russia, including 133 from Saint Petersburg, 1 from the Leningrad region, and 1 from the Republic of Buryatia. Samples were obtained between March 15 and April 23. For analysis, we combined this dataset with additional 76 genomes from Russia available at GISAID [403] as of May 26, 2020, obtained between March 11 and April 14. The resulting dataset includes 211 sequences from 25 out of the 85 regions (federal subjects) of Russia (including the Republic of Crimea), with the two regions with the largest numbers of cases, Moscow and Saint Petersburg, most densely covered. Therefore, while coverage differs between regions, this dataset is representative of the early outbreak in Russia in terms of geographic spread (Fig. 4.1a). For the phylogenetic context, we also used the 19,623 whole-length, high-quality GISAID genomes from the rest of the world available on May 26, 2020.

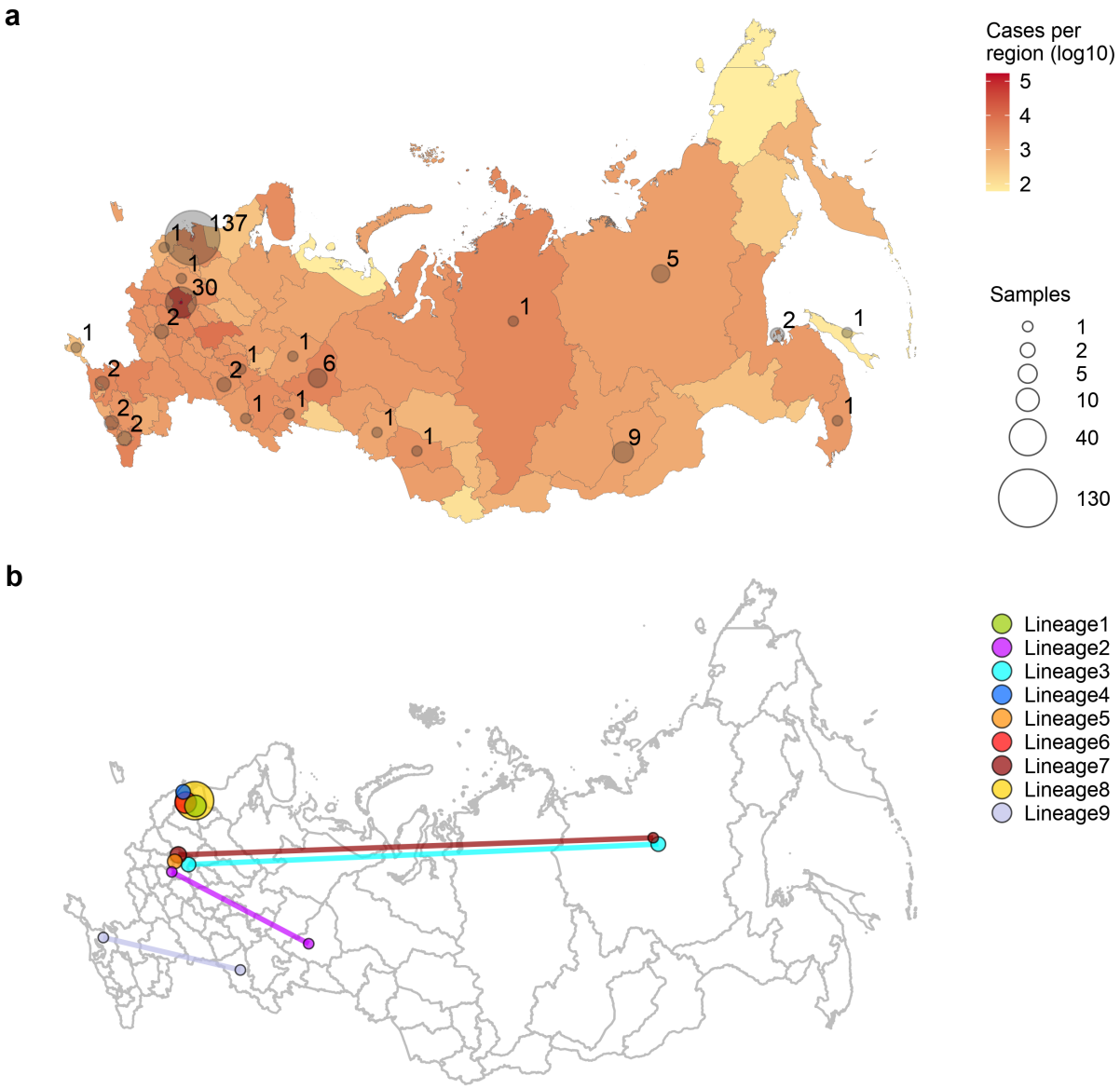


Figure 4.1. Early epidemiology of SARS-CoV-2 in the Russian Federation. (a) The number of confirmed cases (a, red) and the number of sequenced complete genomes (a, grey circles) per region of the Russian Federation (including the Republic of Crimea) as of May 26. Circle sizes are proportional to the numbers of obtained sequences, which are also shown next to the circles. For the purpose of this figure, Moscow was pooled with the surrounding Moscow Region, and Saint Petersburg was pooled with the surrounding Leningrad Region. (b) Identified Russian lineages and corresponding regions; circle size is proportional to the number of sequences belonging to the lineage at this region, and lineages spanning multiple regions are connected by lines.

4.3.2. Multiple origins of SARS-CoV-2 in Russia

Phylogenetic analysis indicates that the Russian samples are scattered across the SARS-CoV-2 evolutionary tree, representing much of its global diversity. Most samples correspond to the B.1, B.1.1, and B.1.* lineages (PANGOLIN nomenclature [404]) or clade G, GR, and GH (GISAID nomenclature [405]) which are wide-spread in Europe (Fig. 4.2). While the predominantly Asian A, B, and B.2 lineages comprised 53% of the sampled global viral diversity around the time of Russian border closure (March 27), only 4 (2%) of the Russian samples belonged to them.

We aimed to identify distinct introductions of SARS-CoV-2 into Russia. Phylogenetically, each of the 211 Russian sequences belongs to one of the three categories (Fig. 4.3). Firstly, 77 (36%) of these sequences form the 9 distinct Russian transmission lineages (Fig. 4.2,4.3), defined as monophyletic groups (clades) carrying more than one sequence all of which are Russian. These lineages indicate within-Russia transmission of introduced variants. Three of these lineages had no Russian sequences at their ancestral nodes, indicating that they originated from at least three distinct introduction events (Fig. 4.4c-d; Supplementary Note B).


19A
B|L
(1522)
(73)
(68)
Russia/StPetersburg-73603/2020|Mar-20
(508)
(35)
B.2|V
(1387)
(36)
Russia/StPetersburg-82005/2020|Mar-24
19B
(7)
(1989)
A|S
(6)
(68)
Russia/Kabardino-Balkaria-80503/2020|Mar-23
Russia/Kabardino-Balkaria-80501/2020|Mar-23
(32)
(5)
(1)
(5)
Russia/StPetersburg-RII3992/2020|Mar-18
Russia/StPetersburg-RII4936S/2020|Apr-08
Russia/Samara-73406/2020|Mar-20
(2457)
20A
Russia/StPetersburg-RII4326S/2020|Mar-25
(1)
Russia/Tver-72601/2020|Mar-19
(1)
Russia/StPetersburg-RII8935S/2020|Apr-21
(1)
Russia/Moscow-71604/2020|Mar-19
(25)
Russia/Moscow-80112/2020|Mar-23
(118)
Russia/StPetersburg-RII6790V/2020|Apr-16
Russia/StPetersburg-RII6055S/2020|Apr-15
(9)
(63)
Lineage2 (2)
(53)
(489)
Russia/Sverdlovsk-71903/2020|Mar-19
(5)
Russia/Moscow-62501/2020|Mar-11
(63)
Russia/Moscow-62505/2020|Mar-11
(1)
Russia/StPetersburg-RII8876S/2020|Apr-23
Russia/Buryatia-86204/2020|Mar-29
Russia/Crymea-80603/2020|Mar-23
(5)
Russia/StPetersburg-RII6061S/2020|Apr-15
Russia/StPetersburg-RII8298S/2020|Apr-21
Russia/StPetersburg-RII7501S/2020|Apr-20
Russia/StPetersburg-RII4546V/2020|Apr-02
Russia/StPetersburg-RII6053S/2020|Apr-15
Russia/StPetersburg-RII7464S/2020|Apr-20
Russia/StPetersburg-RII4937S/2020|Apr-08
Russia/StPetersburg-RII4597S/2020|Apr-04
Russia/Moscow-77627/2020|Mar-22
Russia/StPetersburg-RII6704V/2020|Apr-16
Russia/StPetersburg-RII4637S/2020|Apr-05
Russia/StPetersburg-RII4949S/2020|Apr-09
Russia/Pskov-83602/2020|Mar-25
Russia/StPetersburg-RII8890S/2020|Apr-23
Russia/StPetersburg-RII4944S/2020|Apr-08
B.1.5|G
Russia/Omsk-89001/2020|Apr-01
Russia/StPetersburg-RII5942S/2020|Apr-13
Russia/StPetersburg-RII7581S/2020|Apr-21
Russia/StPetersburg-RII4382V/2020|Mar-29
Russia/Sverdlovsk-82904/2020|Mar-25
Russia/StPetersburg-RII6027S/2020|Apr-14
Russia/StPetersburg-RII6031S/2020|Apr-14
Russia/StPetersburg-RII7594S/2020|Apr-21
Russia/StPetersburg-RII4384V/2020|Mar-29
Russia/StPetersburg-RII5244S/2020|Apr-10
Russia/StPetersburg-RII8319S/2020|Apr-21
Russia/StPetersburg-RII4678S/2020|Apr-06
Lineage1 (8)
(875)
B.1|G
Russia/StPetersburg-3524/2020|Mar-15
(116)
Russia/Khabarovsk-71403/2020|Mar-19
(337)
(324)
Russia/Chechenya-83801/2020|Mar-26
(203)
(5)
Russia/Chechenya-83803/2020|Mar-26
(31)
B.1.*|GH
Russia/Moscow-71602/2020|Mar-19
Russia/Vector_77628/2020|Mar-22
Russia/Moscow-67117/2020|Mar-16
Russia/Moscow-80402/2020|Mar-23
Russia/Krasnoyarsk-66501/2020|Mar-13
(4288)
Russia/Sakhalin-84401/2020|Mar-27
(2)
Russia/Sverdlovsk-71904/2020|Mar-19
(3)
20C
Russia/StPetersburg-RII3997/2020|Mar-18
(2)
Russia/StPetersburg-RII5247S/2020|Apr-10
(1)
Lineage9 (2)
(1)
Russia/Chelyabinsk-85311/2020|Mar-28
(6)
Russia/StPetersburg-RII4386V/2020|Mar-29
(59)
(38)
Russia/SCPM-O-03/2020|Mar-20
(9)
Russia/Sverdlovsk-84302/2020|Mar-27
(1)
20B
B.1.1|GR
1
Lineage
Lineage1
Lineage2
Lineage3
Lineage4
Lineage5
Lineage6
Lineage7
Lineage8
Lineage9
Vreden group
group 1
group 2
group 3

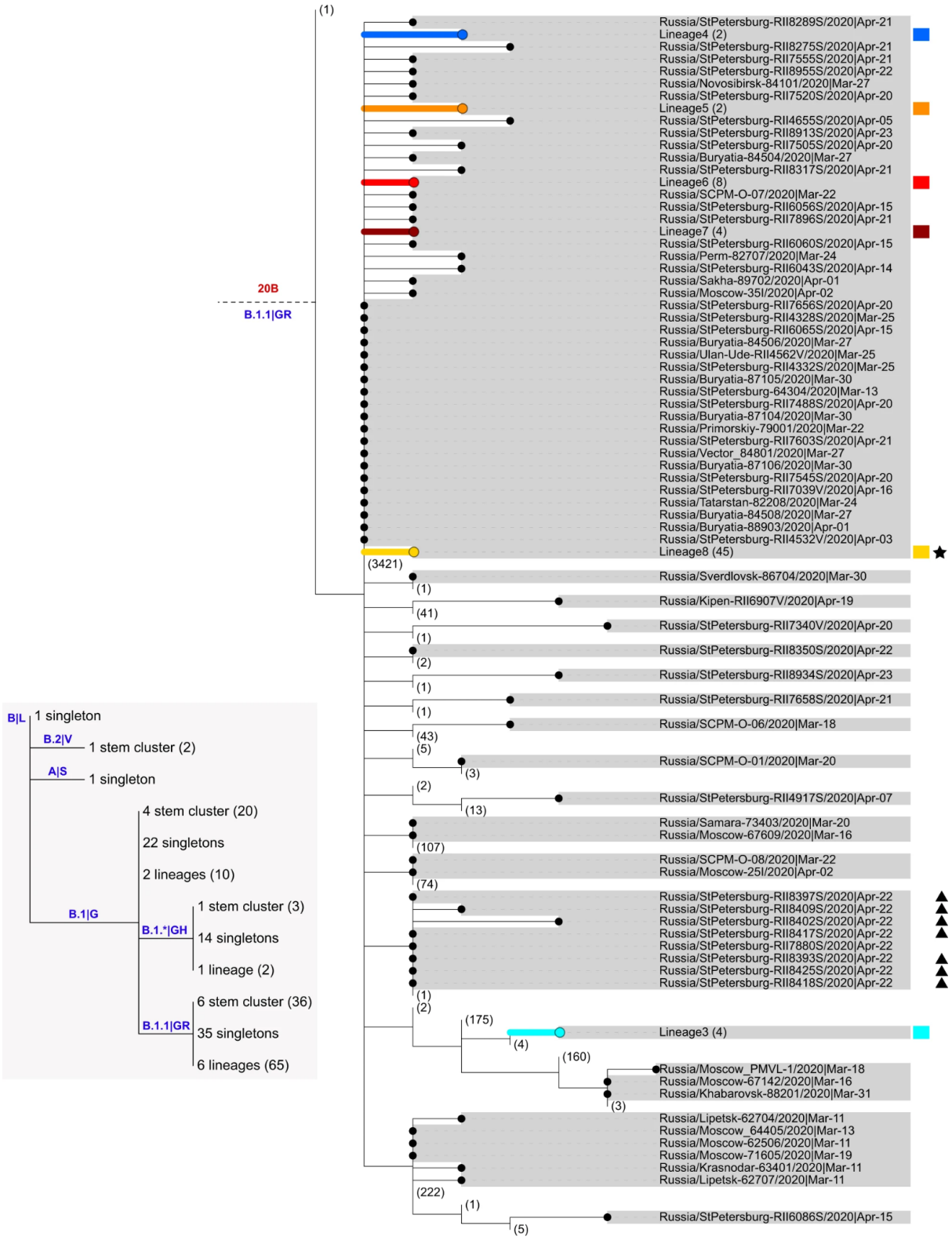


Fig. 4.2. Phylogeny of SARS-CoV-2 in Russia. Russian sequences are identified with dots and highlighted in gray. Russian transmission lineages are truncated to the founder node and highlighted with color (the color scheme is consistent between Figs. 4.1, 4.2, 4.4-4.6). Major SARS-CoV-2 lineages are labeled according to Nextstrain81 and PANGOLIN|GISAID nomenclature in red and

blue, respectively. Non-Russian sequences and lineages carrying no Russian sequences are truncated, with numbers of such sequences shown in brackets. Sequences from the Vreden hospital and lineages carrying such sequences are marked with stars, triangles, and diamonds. Branch lengths represent the number of nucleotide substitutions. "hCoV-19/" prefixes are excluded from all sample names for clarity. The inset summarizes the distribution of Russian singletons, stem clusters, and transmission lineages across major SARS-CoV-2 clades.

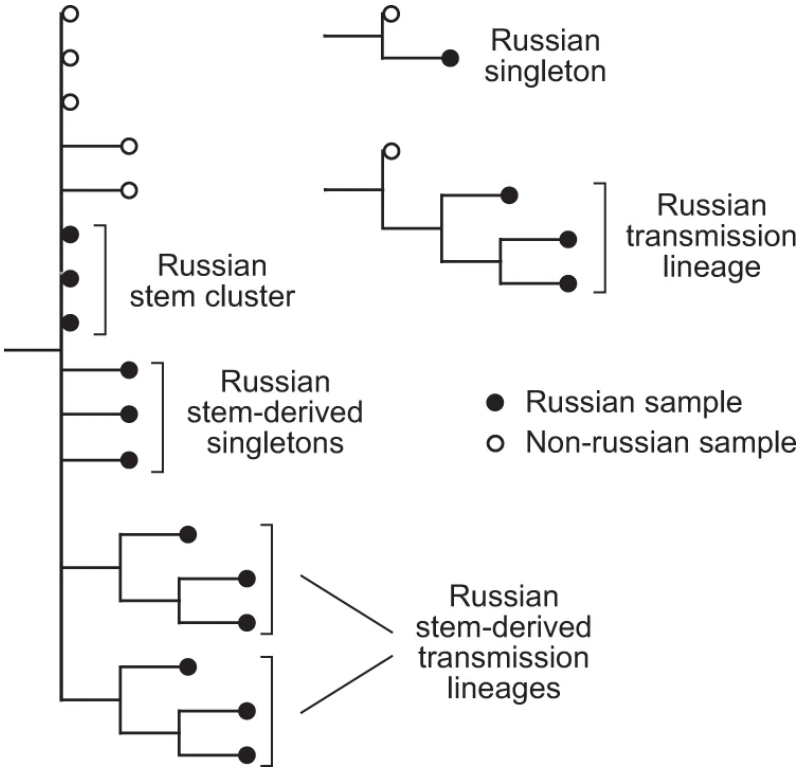


Fig. 4.3: Terminology for phylogenetic groups of samples. We categorized Russian samples into five categories: Russian transmission lineage-set of two or more sequences that form a Russian-only clade; Russian singleton-single Russian sequence that forms a clade of its own and is not a part of a Russian transmission lineage; Russian stem cluster-set of Russian sequences identical to each other and to some non-Russian sequences; Russian stem-derived transmission lineage-a Russian transmission lineage whose immediate ancestor is a Russian stem cluster; and Russian stem-derived singleton-Russian singleton whose immediate ancestor is a Russian stem cluster.

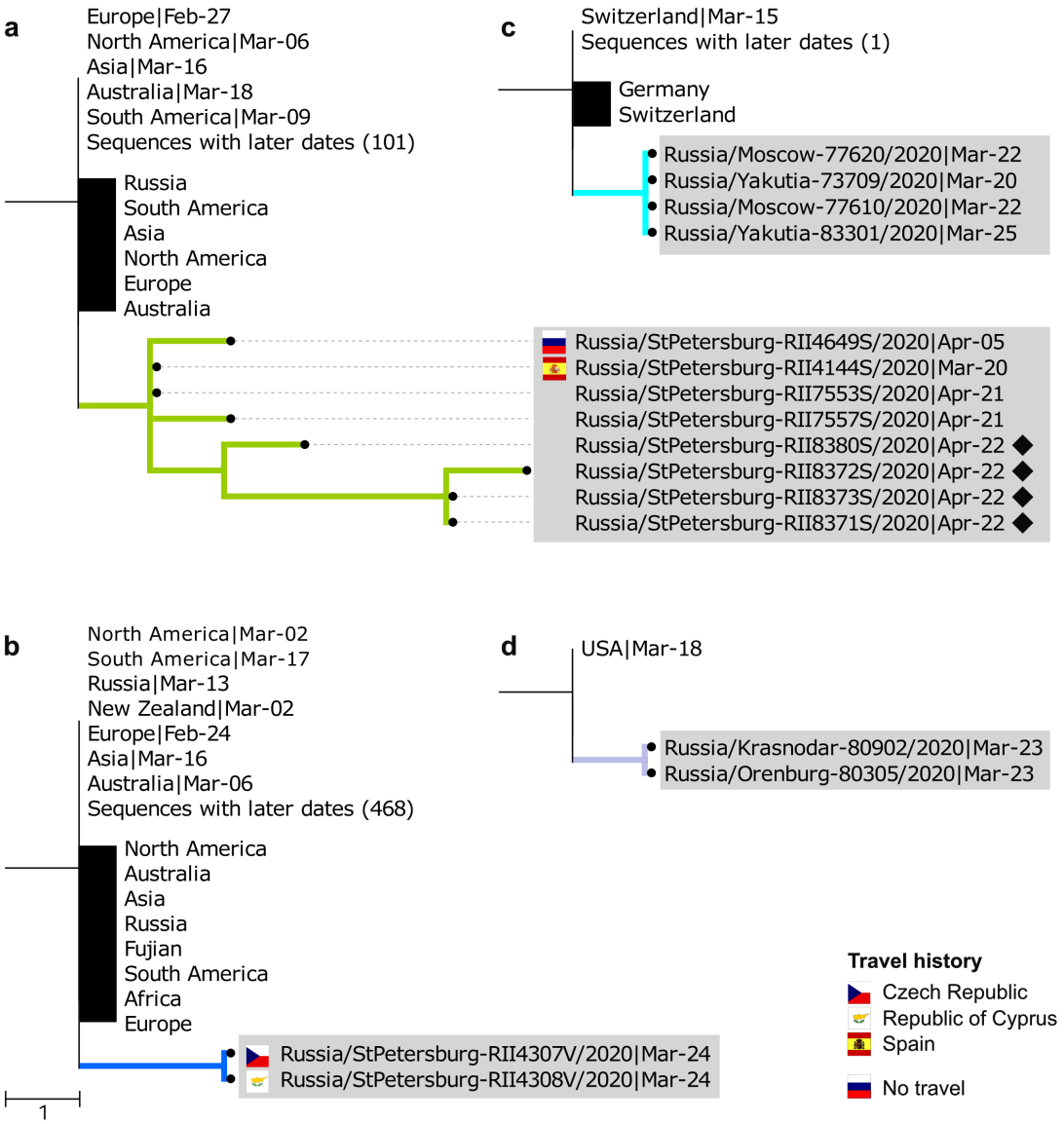


Fig. 4.4. Examples of Russian transmission lineages. Only the phylogeny of the Russian transmission lineage is shown, together with its ancestral phylogenetic node. Russian sequences are marked with dots and highlighted in gray. All other sequences corresponding to an ancestral node (black vertical line) or descendant from it (black rectangle) are truncated, with the region/country and the earliest collection date shown. (a) Lineage 1, a lineage endemic to Saint Petersburg, includes an individual with a history of travel to Spain. (b) Lineage 4 includes individuals with travel history to two different countries, suggesting recurrent introduction. (c) The ancestral node of lineage 3 uniquely maps to Switzerland. (d) The ancestral node of lineage 9 uniquely maps to the USA, and this lineage spans two different regions of Russia. Flags represent individuals with a known history of travel to the corresponding country; the Russian flag shows a known lack of travel history. Diamonds represent samples associated with group 3 of the Vreden hospital outbreak. See Supplementary Fig. B-1 for all nine transmission lineages.

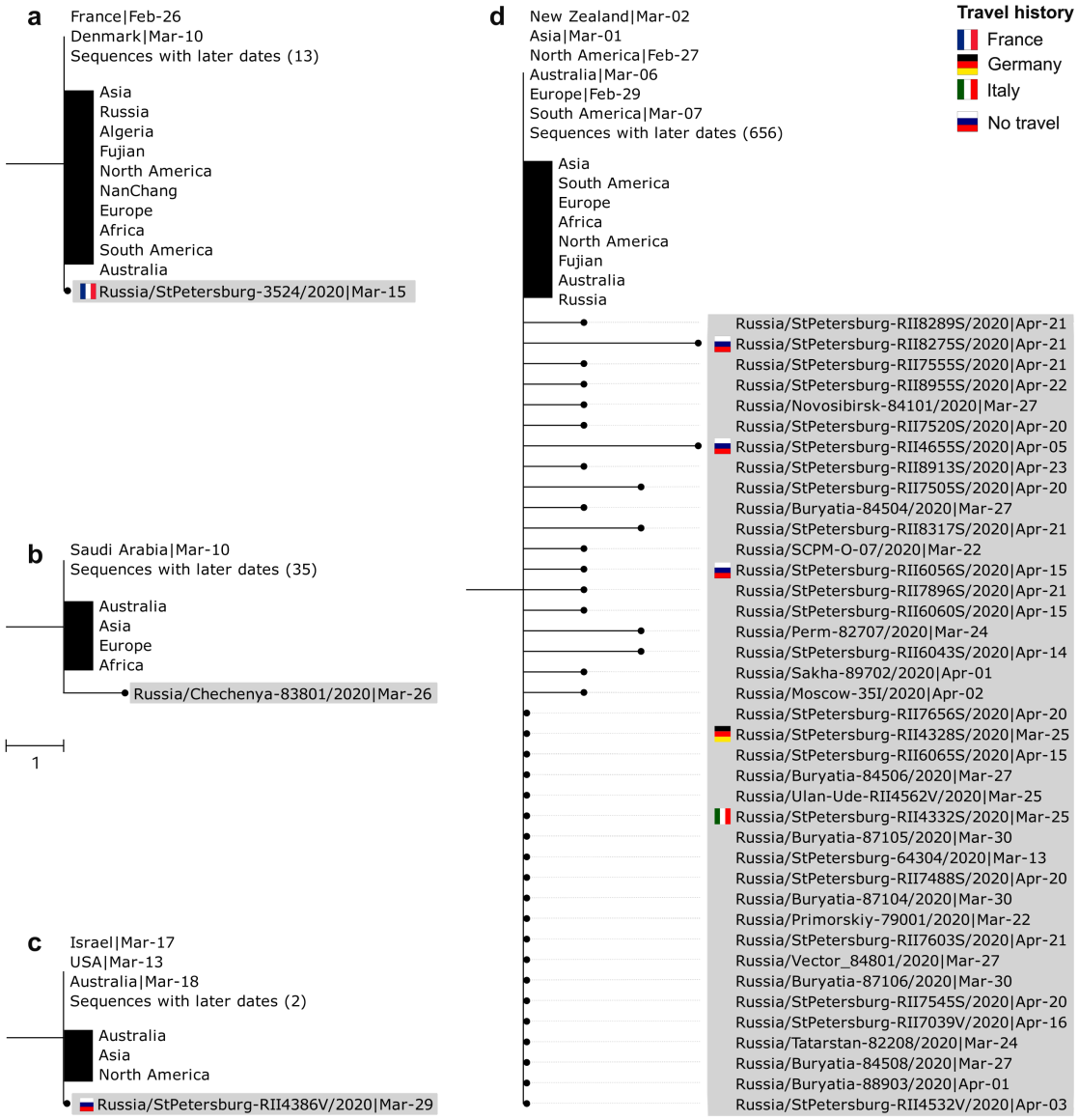


Fig. 4.5. Examples of Russian singletons and stem clusters. Notation is the same as in Fig. 4.4. (a) The singleton obtained from a patient with known travel history to France has French and Danish sequences at the ancestral node, with French sequences having an earlier date. (b) A singleton with a uniquely Saudi Arabian ancestral node. (c) A singleton with the known absence of travel history. (d) A stem cluster with associated stem-derived singletons where multiple introductions were observed. See Supplementary Figs. B-2,3,4 for all singletons and stem clusters.

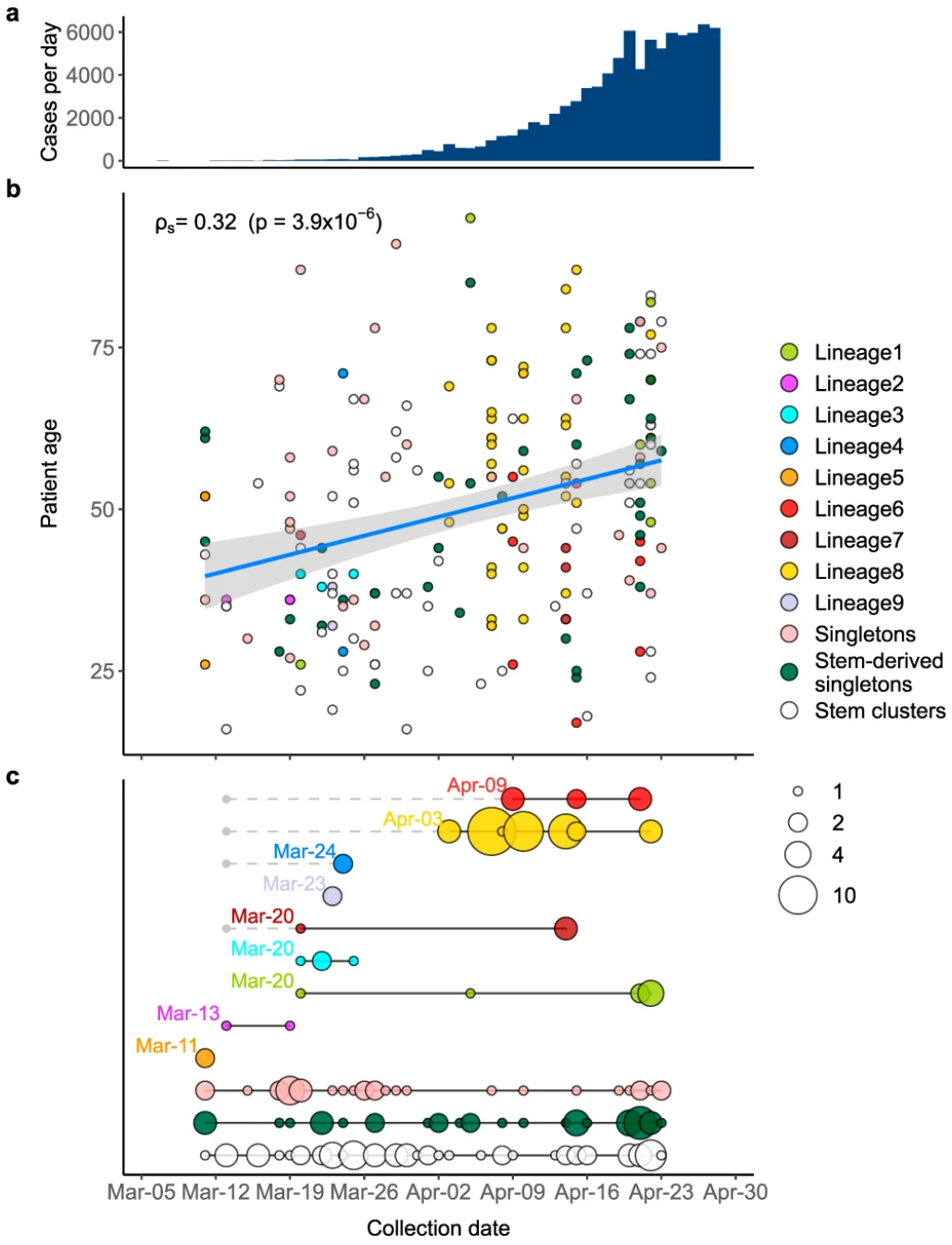


Fig. 4.6. The timeline of SARS-CoV-2 introduction into Russia. Depending on their phylogenetic position, Russian samples are classified as belonging to Russian transmission lineages, singletons, or stem clusters (Fig. 4.3). Circles correspond to Russian samples colored by category. (a) The number of new registered COVID-19 cases per day in Russia between March 5 and April 30. (b) Correlation between sample collection date and patient age. The linear fit (r = 0.32; p = 3.9 × 10-6) is shown (blue line), with the 95% confidence interval indicated as a shaded area. Spearman correlation coefficient is shown. (c) Estimated introduction dates for Russian transmission lineages, singletons, and stem clusters. The circle size is proportional to the number of samples. Black lines correspond to the full date range. For each Russian transmission lineage, the indicated date corresponds to the collection date of the earliest sample. For stem-derived Russian transmission lineages (lineages 4, 6, 7, and 8), the earliest date of the corresponding stem cluster is also shown with a gray dot.

The remaining six Russian transmission lineages carried both non-Russian and Russian sequences at their ancestral nodes (Fig. 4.4a-b). Such lineages, hereafter referred to as "stem-derived transmission lineages", could also result from distinct introduction events; alternatively, their last common ancestor could already reside in Russia. To estimate the number of introductions giving rise to the stem-derived lineages, we make use of the direct data on travel history (or lack thereof) available for a fraction of our patients. Using a statistical model, we estimate that these lineages together resulted from roughly three additional introduction events (Supplementary Note B). This number could be an underestimate due to the undersampling of diversity outside Russia. Indeed, one of the identified lineages (Fig. 4.4b) involves two samples that had travel history to two different countries, indicating likely double introduction within the same lineage.

Secondly, we observe 73 (34%) singletons that are not involved in any of the Russian transmission lineages, each possessing their own characteristic mutations not shared by any other Russian sequences (Fig. 4.5). These include 33 singletons without any Russian ancestral sequences, and 40 singletons stemming from ancestral nodes with Russian sequences (hereafter, "stem-derived singletons"). We assume that the former correspond to sole introduced cases, for a total of 33 such introductions. Most of them had probably not resulted in any within-Russia transmission. However, we find that some of the singleton sequences were sampled from patients without any travel history (Fig. 4.5c). This indicated that at least some of the singletons likely correspond to distinct introductions that yielded domestic transmission clusters, of which just one representative was sequenced. Using travel data, we estimate that stem-derived singletons resulted from ~6 additional introduction events (Supplementary Note B).

Thirdly, the remaining 61 sequences (29%) fell into 12 sets of two or more identical sequences, each of which was also identical to some of the non-Russian sequences (Fig. 4.5d). These sets are further referred to as stem clusters. Again, individual samples within a stem cluster could correspond to distinct introductions or domestic transmission. When data on travel history is available, we find that some of such clusters include multiple individuals with travel history, suggesting that identical sequences were repeatedly introduced into Russia at least in some instances (Fig. 4.5d). On the other hand, we also observe individuals without travel history, indicating domestic transmission of these variants. From travel data, the estimated number of introductions leading to stem cluster sequences is ~22.

Overall, we estimate the number of independent transmissions into Russia as ~6 resulting in transmission lineages, ~39 resulting in singletons, and ~22 resulting in stem clusters, for a total of 67 events. The uncertainty associated with this estimate is largely dependent on the approach for treating the numbers of introductions leading to stem clusters and stem-derived singletons. If each stem cluster (together with any singletons derived from) is assumed to originate from exactly one introduction, the estimated number of introductions is 48. If instead each sequence within a stem cluster and each stem-derived singleton has resulted from a distinct introduction, the estimated number of introductions rises to 143.

The earliest collection date of a sample belonging to a transmission lineage represents the latest possible date this lineage could have been introduced into Russia. For most Russian transmission lineages, the earliest sample collection dates fall into the range between March 11 and 24, indicating that the corresponding lineages were introduced not long before (Fig. 4.6b). Indeed, out of the nine Russian transmission lineages, only two (lineages 6 and 8) had later dates of the earliest sequences. However, those were stem-derived lineages, and the oldest stem sequences corresponding to them dated to March 13, suggesting that these

transmission lineages could have also been established by this date. Many (15 out of 33) of the singletons were also collected within this timeframe, although some were collected later (mean date: March 29); together with the fact that many of the singletons have not traveled (Fig. 4.5, Supplementary Figs. B-2,3), this indicates that they, in fact, correspond to as yet unsampled transmission lineages. By contrast, most stem-derived singletons were sampled at later dates (mean date: April 7, Mann-Whitney U-test, p=0.014), suggesting that they were more likely than non-stem-derived singletons to originate from the within-Russian transmission.

By the time introduction into Russia had started, the virus had already spread through other countries, with the same variant frequently present at multiple locations. Therefore, the source of most introductions could not be established unambiguously. Still, for a fraction of the samples, the phylogenetic position is consistent with the source. For example, the earliest patient with known travel history has returned to Russia from France, and her sample is nested within a clade with just French and Danish sequences at the ancestral node, with French having earlier dates and therefore arguably more plausible source (Fig. 4.5a). For two additional sequences corresponding to regional outbreaks, no direct travel data was available but the probable source could be established from media reports and was consistent with the phylogenetic position of the corresponding clades. This was the case for the import of clades from Switzerland into Yakutia (the Sakha Republic) (Fig. 4.4c) [406] and from Saudi Arabia to the Chechen Republic (Fig. 4.5b) [407].

Overall, out of the 13 patients with known travel history (11 direct + 2 from media reports), the country of origin is consistent with the sampling locations of the same or ancestral nodes in 9 cases, including the 3 cases when it is uniquely identified. In one case (Supplementary Fig. B-4b), the travel direction (Egypt) is inconsistent with the phylogenetic

position of the sample, and in the remaining three cases, there is not enough phylogeographic data to make a call. For the same 9 out of the 13 patients, we were able to correctly and uniquely identify the source continent (Europe in all cases).

Partial consistency between direct travel history and phylogenetic position motivated us to attempt to infer the sources of Russian samples phylogeographically. In the absence of travel data, we position the hypothetical source of one transmission lineage (lineage 9, Fig. 4.4d) as the USA; and of five singletons, as Chile (Supplementary Fig. B-2k), England (Supplementary Fig. B-2l), France (Supplementary Fig. B-2m), and Denmark (Supplementary Fig. B-3h,p). For 6 additional singletons, we position the hypothetical source to the continent (Europe in all cases). Finally, we estimate the hypothetical origin of one stem cluster as Sweden (Supplementary Fig. B-4c), and of two more stem clusters, as Europe. Importantly, phylogeographic inferences are strongly sensitive to sampling bias and should be treated with caution.

The individuals importing the virus and seeding the Russian transmission lineages were not a random sample of the population. Very early samples were collected from patients who were on average younger than those sampled later (Fig. 4.6b). This is consistent with the major role of younger Russians in the import of the virus into Russia [408], possibly because they comprised a larger share among the people returning from business trips or holidays.

4.3.3. Temporal dynamics of SARS-CoV-2 spread in Russia

Following introduction, the virus has spread throughout Russia. Four out of the 9 identified Russian transmission lineages, and 8 out of the 12 stem clusters, span multiple regions (Figs. 4.1b, 4.3c-d, 4.4d). As Moscow and Saint Petersburg are major transport hubs, together responsible for 77% of the international air traffic in Russia, we hypothesized that

the virus was introduced through these cities, and spread throughout Russia from them. Contrary to this hypothesis, among the Russian stem clusters and singletons, the samples from Moscow or Saint Petersburg do not sit on shorter branches than samples from other regions; in fact, branches leading to them tend to be slightly longer (mean branch length 0.88 vs. 0.37 substitutions, p=0.006, permutation test), probably because of more extensive regional sampling early in the outbreak. Thus, we see no evidence for a preferential direction of transmission within Russia, suggesting that the Russian epidemic has been seeded by near-concurrent introduction into multiple regions.

4.3.4. Vreden hospital outbreak

A major transmission cluster corresponded to the nosocomial outbreak at the Vreden Russian Research Institute of Traumatology and Orthopedics in Saint Petersburg (hereafter, the Vreden hospital) [409,410]. According to an internal investigation, the suspected patient zero at the hospital had surgery on March 27, 2020. While routine COVID-19 testing at the Vreden hospital began on March 18, the earliest samples that tested positive were collected on April 3. Quarantine was gradually introduced between April 7 and April 9, which involved a complete lockdown of the hospital, isolation of units from each other, and shutdown of the hospital-wide ventilation system. 474 patients and 270 medical workers remained inside the hospital for the following 35 days.

Our dataset contains SARS-CoV-2 genomes obtained from 52 of the Vreden hospital patients or medical workers. Phylogenetic analysis indicates that these samples form three distinct groups, each defined by its own set of mutations. The largest group, group 1, includes 41 sequences obtained between April 3 and April 22 and represents a distinct Russian transmission lineage (lineage 8, star in Fig. 4.2). This lineage derives from a very prolific

ancestral node that has seeded multiple lineages throughout the world, including five of the Russian transmission lineages, so its origin cannot be positioned phylogeographically. Group 2 contains 7 out of 9 sequences in another clade, which also carries one non-Russian (English) sequence (triangles in Fig. 4.2). Finally, group 3 includes 4 sequences and represents a clade of its own within another Russian transmission lineage (lineage 1, diamonds in Figs. 4.2 and 4.3a). While samples from group 1 came from different units located at different floors of the Vreden hospital, samples from groups 2 and 3 each came from their own unit.

Groups 1 and 2 are phylogenetically remote from group 3, with six mutations separating the most recent common ancestors (MRCAs) of groups 1 and 2 from group 3 (Fig. 4.2). Groups 1 and 2 belong to the B.1.1 lineage defined by three mutations at positions 28881, 28882 and 28883, and are further defined by mutations at positions 26750 and 1191, respectively. By contrast, group 3 belongs to the B.1.5 lineage, and is supported by the mutation at position 20268 which is widespread over the world and appeared early in phylogenetic history, as well as by two additional mutations. This provides strong evidence that group 3 originates from a separate introduction to that of groups 1 and 2.

To understand the spread of the outbreak at the Vreden hospital, we performed a Bayesian phylodynamic analysis using the birth-death skyline model [44] of BEAST2 [319]. Given the possibility of multiple introductions, we analyzed the whole Vreden dataset comprising groups 1, 2, and 3; and also its two subsets consisting of groups 1 and 2, and just of group 1. The results are summarized in Figs. 4.7-8 and Supplementary Tables B-4,5,6.

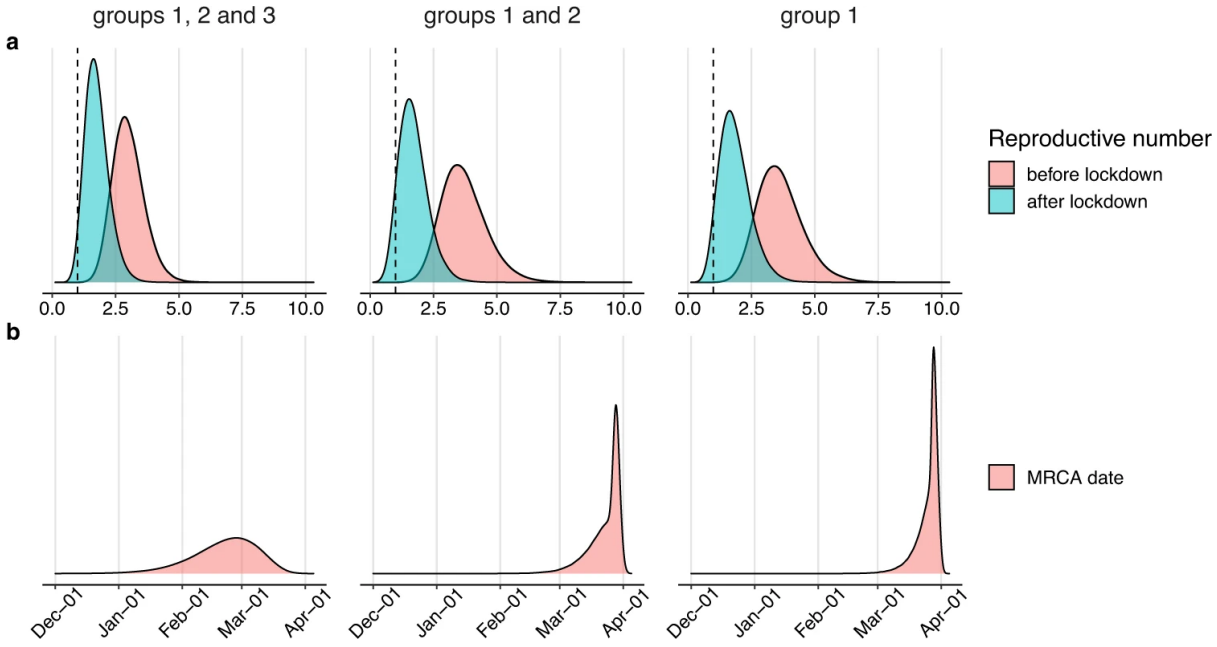


Fig. 4.7. Vreden hospital outbreak parameter estimates produced by birth-death skyline model in BEAST2. Panels show posterior distributions of effective reproductive number R_e (upper panel) with the dashed vertical line corresponding to R_e=1 and the date of the MRCA (lower panel) for analyses based on Vreden hospital samples from groups 1, 2, and 3 (left column), groups 1 and 2 (middle column) and group 1 (right column).

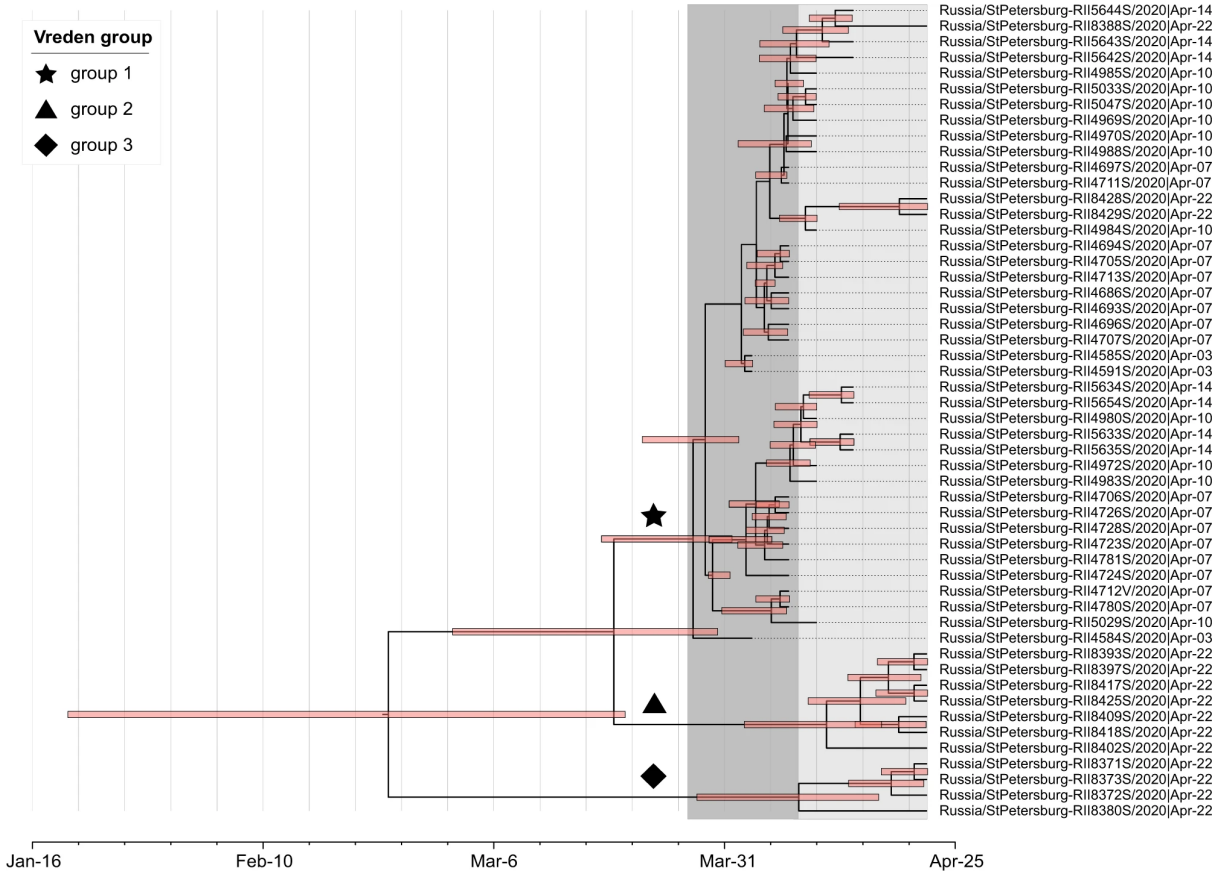


Fig. 4.8. Maximum clade credibility tree for the Vreden hospital outbreak. Groups 1, 2, and 3 are marked by a star, a triangle, and a diamond, respectively. Pink bars represent 95% credible intervals. The timeline of the outbreak is shown in gray, with the time interval from patient zero (March 27) till the introduction of quarantine (April 8) highlighted with a darker tone.

We found that the Bayesian analysis supports at least two distinct introductions of SARS-CoV-2 into the Vreden hospital. This is based on the deep split between group 3 and groups 1-2. The MRCA of all three groups dates to February 21 (95% CI January 20 — March 21). This is more than a month before the assumed date of introduction (March 27), implying that group 3 and the remaining Vreden samples were introduced independently.

A third introduction into the Vreden hospital is also highly probable. Indeed, the MRCA of groups 1 and 2 dates to March 24 (95% CI March 6 — April 1). As there was no sign of infection at the hospital before the end of March, it is quite likely that these two groups originated through separate introductions. The root of group 1 dates to March 26 (95% CI March 13 — April 2), which is consistent with the suspected illness period of the patient zero. Additional evidence that groups 1 and 2 originate from distinct introductions is provided by the fact that the clade that includes group 2 also carries a non-Russian (English) sequence (Fig. 4.2).

We estimated the phylodynamic parameters before and after the quarantine measures were introduced. In all three analyzes, the estimates were stable and consistent with each other. Based on the analysis of all three groups, we found that the effective reproductive number R_e was 3.00 (95% CI 1.85-4.25) before April 8, and dropped to 1.76 (95% CI 0.91-2.71) after April 8 (Fig. 4.7). The same estimates of the effective reproductive number R_e from group 1 only are 3.64 (95% CI 2.01-5.43) before quarantine and 1.85 (95% CI 0.77-3.06) after quarantine, respectively. These estimates are consistent with each other, and the potential effects of population structure do not create considerable biases. The substantial

decrease of the R_e upon introduction of quarantine can also be seen from the incidence data on moderately-to-severely ill patients (those deemed to require transition to specialized COVID-19 facilities; Supplementary Fig. B-5).

4.4. Discussion

The ongoing pandemic of SARS-CoV-2 has involved rapid spread of the virus across the borders of most nations within the few weeks of February and March of 2020. While Russia was behind many of the neighboring countries in the initial rise of the case counts, it has rapidly caught up in the following weeks. By analyzing the phylogenetic distribution of 211 early COVID samples from those dates, we provide details of this process, shedding light on the patterns of transborder transmission of the virus and the factors that affect it.

SARS-CoV-2 accumulates substitutions at the average rate of ~1 per 1000 nucleotides per site per year [411], which means that its genome accumulates on average just one mutation per 2-3 transmissions. Therefore, phylogenetic trees have lower resolution than transmission trees, meaning that transmission history cannot be fully resolved from phylogenetics alone. In particular, in the absence of complete data on travel history, there is no simple rule for counting the number of introductions. A common rule of thumb is counting the number of country-specific clades [397,412–415]. However, multiple introductions can result in a single clade if the viral diversity abroad is undersampled [416–418]; and a single introduction can result in multiple clades if their last common ancestor has already been introduced [397].

By using direct travel data, we show that both these problems hold. Indeed, we find transmission lineages apparently co-introduced from multiple countries (Fig. 4.4b) or singletons without any history of travel (Fig. 4.5c). The uncertainty in the number of introduction events is the highest for identical sequences with broad geographic distribution, e.g., the last common ancestor of lineage B.1.1. This node constitutes a stem cluster of 100 identical Russian sequences, as well as 4,323 sequences from outside Russia. It is the immediate ancestor to five Russian transmission lineages and 19 stem-derived Russian

singletons, so how many times this sequence has been introduced into Russia strongly affects the overall counts of the number of introductions. Travel data indicate that a stem group can carry a combination of multiple introduced and domestically transmitted sequences, complicating the inference of the number of introductions.

Under a simple statistical model combining genetic and available travel data, we estimate that the sampled diversity of SARS-CoV-2 in Russia originated from 67 introductions. Since this corresponds to roughly one introduction per every three sequences sampled, the actual number of introductions was probably much higher, and its estimate will likely increase as more sequences are sampled.

Overall, the slow mutation rate of SARS-CoV-2 together with unequal sampling among countries complicates phylogeographic inference of introduction sources [255,397]. We attempted such inference for the Russian samples nevertheless, and found that its results are largely supported by direct travel data when such data is available. The somewhat higher phylogeographic resolution for the Russian lineages compared to that in previous works [255,397] may be due to the fact that most Russian lineages originated late, when the source European lineages were already well established. Still, for most of the introductions, phylogeography is not informative of their origin; moreover, phylogeographic inferences are expected to be biased in the presence of uneven sampling between countries. This illustrates the need of combining multiple data types, including travel history, for understanding the viral origin and spread [419].

China is the 5th most popular destination for Russian citizens, accounting for 5.5% of all international travel. Overall, approximately five million people traveled between the two countries in 2019, with 65% of them traveling by land [420,421]. Although it is hard to ascribe epidemiological results to specific NPIs, our analysis suggests that the border closure

with China implemented in February has effectively curbed the virus introduction into Russia from the Asian direction. Indeed, only four of our samples belong to lineages A, B, and B.2 (GISAID clades S, L, and V, respectively), which predominantly originated in Asia; and two of those sequences are nested within other European subclades, indicating that the import was through Europe. This fraction is not representative of global case counts at that time, and is instead reflective of travel patterns and the history of border closures. It is also in contrast to the situation in other countries where the outbreaks started earlier and were probably seeded by a direct introduction from Asia [422–424].

For most of the discovered transmission lineages, the earliest sampled sequence was collected between March 11 and 24, 2020 (Fig. 7b). In the larger UK dataset, the mean time between the importation date of a lineage and its earliest sampling date within the UK was estimated to be approximately two weeks, although this depends on many factors including lineage size and sampling intensity [397]. If this can be extrapolated to Russian transmission lineages, this implies that these lineages typically originated from imports in the last week of February and the first week of March. A contributing factor could have been intensive travel around the Russian state-mandated long holidays of February 22-24 and March 7-9. Further establishment of Russian transmission lineages could be limited by NPIs, in particular, by the introduction of mandatory quarantine for incoming travelers on March 5, as well as by the overall radical reduction in international travel after these dates.

Detailed analysis of localized transmission clusters helps understand viral spread. Well-studied examples include the Diamond Princess cruise ship [425–428]; the Grand Princess cruise ship [429]; an international conference in Boston [413]; a community living facility in the Boston area [413]; and the nosocomial outbreak in the Netcare St. Augustine's Hospital in South Africa [430]. In all but one of these cases, the outbreaks were genetically

homogenous, indicating that they each arose from a single case. In the community living facility, multiple introductions have occurred, but there was a dominant clade that included nearly all the samples, while other clades were rare [413]. By contrast, at the Vreden hospital outbreak, we observe multiple (2-3) introductions, each of which gave a prolific clade. This indicates that this outbreak could have originated from multiple superspreading events. Furthermore, we estimate the initial effective reproductive number R_e during the pre-quarantine period at ~3.00, which is rather high. Multiple superspreading events and the high R_e can be due to some of the conditions specific to a hospital not specifically equipped for infection control, including dense contacts (in particular, spread by medical workers), absence of protective measures, and lack of awareness. In the second phase of the outbreak, we observe a significant decrease in R_e down to ~1.76. This change can be explained by two factors. Firstly, it can be due to increased awareness and quarantine measures which were in effect after April 7. Secondly, it can be due to a large number of people already ill, preventing further infection; indeed, around 30% of people at the hospital had been infected by April 22. We cannot quantify the contribution of these factors to the slowing rate of infection spread with available data and methods.

CHAPTER 5: CONCLUSIONS

In this Ph.D. thesis, I demonstrate how methods of molecular epidemiology can be applied to study two infectious diseases affecting the population of Russia.

The first part of the thesis is devoted to the analysis of a densely sampled HIV-1 epidemic in Oryol Oblast, Russia. Our findings can be summarized as follows:

1. The HIV-1 epidemic in Oryol Oblast resulted from at least 332 imports, with 82 imports giving rise to observable further transmission within the region.

2. Subtype A predominates the epidemic (87.2%), followed by CRF63 (7.2%) and subtype B (2.5%).

3. Subtype A is responsible for the moderate growth of the epidemic with R_e of 2.8 [1.7-4.4].

4. CRF63 which emerged in Siberia in the 2000s and was recently introduced into Oryol Oblast demonstrates more rapid growth with R_e of 11.8 [4.6-28.7] indicating the need for close monitoring and investigation of this variant.

5. Injecting drug users but not males are clustered together in transmission lineages, reflecting the structure of the HIV-1 population in the region.

6. The MSM transmission route is associated with subtype B; still, the highly unbalanced male-to-female ratio suggests underreporting of the MSM route among males infected with subtype B in our dataset.

The second part of the thesis is focused on a much less densely sampled dataset and investigates the emergence of SARS-CoV-2 in Russia early in the COVID-19 pandemic.

1. Using travel data, we estimate that the analyzed dataset of 211 Russian SARS-CoV-2 sequences resulted from at least 67 imports.

2. Among these imports, nine transmission lineages resulted in an observable domestic transmission of the virus within Russia.

3. A nosocomial outbreak in the Vreden hospital in Saint Petersburg resulted from 2-3 independent introduction events; we observe a decrease of R_e after the quarantine was introduced in the hospital.

Despite obvious social significance, neither SARS-CoV-2 nor HIV-1 possesses enough genetic data for a decent resolution of the Russian part of the epidemics; current sequencing efforts in Russia are obviously not enough to cover a sensible fraction of a large epidemic. For the HIV-1 project, we tried to tackle this by focusing on a single HIV-1 sub-epidemic for which sufficient sequencing coverage could be attained. This allowed us to provide a detailed description of the epidemic structure. Importantly, because HIV-1 requires certain types of direct contacts between people to be transmitted, it is to a notable extent isolated between various geographic areas: the number of the inferred imports was stable to the number of available sequences sampled outside Oryol Oblast even though sampling density outside Oryol Oblast was extremely low.

In contrast, SARS-CoV-2 possesses much higher mobility being characterized by both direct and indirect transmission and a much shorter serial interval. Consistently, although two-thirds of our dataset were collected in Saint Petersburg, we observed four transmission lineages and eight stem clusters carrying samples from different geographic areas. In the case of SARS-CoV-2, dense sampling of a particular geographic region would not have been informative of separate imports unless sufficient coverage outside this region had been

provided. Targeted sampling of infections observed in the Vreden hospital allowed us to apply phylodynamics to study the outbreak, although the inferred number of imports provides only a lower estimate. The inference of transmission events, including imports, is further complicated by a relatively low mutation rate of the virus.

The two viruses studied in this thesis are among the few pathogens that can be analyzed in this way in Russia. Currently, *Mycobacterium tuberculosis* (due to routine molecular surveillance of drug-resistant variants), and, to a lesser extent, influenza A and hepatitis C, are probably the only other pathogens that could possess enough genetic data obtained in Russia to be analyzed in a similar way. This thesis, describing epidemics with different sampling densities, timespan, and differences in evolutionary rate and generation time of the causative viruses, provides an illustrative example of how molecular epidemiology studies can be designed and performed to investigate Russian epidemics and local outbreaks. It also offers some practical implications.

Our results on the HIV-1 project are of promising practical use. First, we've informed the Oryol AIDS center on drug resistance predicted from assembled *pol* fragments; this will be used to adjust ART where needed. Second, transmission lineages we identified can be used to track the growth of the epidemic. The Oryol AIDS center occasionally surveys infected people and tests their reported contacts for HIV; information on actively growing lineages can make these interventions targeted and more efficient. Until now, there has been a substantial lag between sample collection and sequencing; the Oryol center agreed to collect and process new samples faster so that our team could analyse new cases on a quarterly basis and infer and report actively growing clusters.

We have also informed the Oryol center about a rapidly growing cluster of CRF63 which they promised to start working with shortly, which I think will still be an effective

intervention despite of a current two-year lag. The CRF63 cluster is probably the most important observation in our results, as it may significantly affect epidemiological situation in the region, and it definitely deserves intense surveillance and investigation of its epidemiological and biological properties.

As for the SARS-CoV-2 project, its immediate applications are less obvious as our observations are mostly retrospective (which does not affect HIV characterized by a much slower rate of transmission and a more inert epidemic). For instance, our results on the SARS-CoV-2 project agree with case data on a relatively late onset of the COVID-19 epidemic in Russia and suggest that early border closure with China has helped postpone the ignition of the epidemic in the country. This may sound as an argument for rapid border closure in this (e.g. as novel potentially more dangerous variants arise) and yet-unknown future epidemics, although careful simulations are needed to validate or predict effects like that properly. Another retrospective observation coming from our later work indicates that a significant but incomplete reduction of international travel traffic in summer 2020 was ineffective in preventing seeding new transmission lineages in Russia [434].

In the context of SARS-CoV-2, molecular epidemiology is immediately helpful in tracking existing and emerging variants. Our group has recently identified and described two potentially more transmissive endemic lineages circulating in Russia in 2020 together with the Alpha lineage (B.1.1.7) (all outcompeted by Delta in 2021) [435], and subsequently made a resource that offers the most up-to-date temporal and geographic spread of SARS-CoV-2 variants circulating in Russia (https://taxameter.ru/).

Yet, genomic coverage of the epidemic in Russia is rather low which delays observing changes. I think in the context of Russia, achieving more general, non-molecular, objectives in epidemiology is most crucial at the moment, which include intensive testing (ideally,

variant-specific PCR panels that allow for a less accurate compared to genomic data but a faster and cheaper approximation of a variant dynamics), higher adherence to self-isolation when needed, and increasing vaccination rate.

BIBLIOGRAPHY

1. Wiesner PJ, Handsfield HH, Holmes KK. Low antibiotic resistance of gonococci causing disseminated infection. N Engl J Med. 1973;288. doi:10.1056/NEJM197306072882308

2. Sarafian SK, Knapp JS. Molecular epidemiology of gonorrhea. Clinical Microbiology Reviews. 1989. doi:10.1128/cmr.2.suppl.s49

3. Ejercito PM, Kieff ED, Roizman B. Characterization of Herpes Simplex Virus Strains Differing in their Effects on Social Behaviour of Infected Cells. J Gen Virol. 1968;2: 357–364.

4. Pereira L, Cassai E, Honess RW, Roizman B, Terni M, Nahmias A. Variability in the structural polypeptides of herpes simplex virus 1 strains: potential application in molecular epidemiology. Infect Immun. 1976;13. doi:10.1128/iai.13.1.211-220.1976

5. Svensson L, Uhnoo I, Grandien M, Wadell G. Molecular epidemiology of rotavirus infections in Uppsala, Sweden, 1981: disappearance of a predominant electropherotype. J Med Virol. 1986;18: 101–111.

6. Hayward GS, Frenkel N, Roizman B. Anatomy of herpes simplex virus DNA: strain differences and heterogeneity in the locations of restriction endonuclease cleavage sites. Proc Natl Acad Sci U S A. 1975;72: 1768–1772.

7. Buchman TG, Simpson T, Nosal C, Roizman B, Nahmias AJ. The structure of herpes simplex virus DNA and its application to molecular epidemiology. Ann N Y Acad Sci. 1980;354. doi:10.1111/j.1749-6632.1980.tb27972.x

8. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol. 1993;31: 406.

9. Lari N, Rindi L, Lami C, Garzelli C. IS6110-based restriction fragment length polymorphism (RFLP) analysis of Mycobacterium tuberculosis H37Rv and H37Ra. Microb Pathog. 1999;26. doi:10.1006/mpat.1998.0270

10. Bifani PJ, Plikaytis BB, Kapur V, Stockbauer K, Pan X, Lutfey ML, et al. Origin and interstate spread of a New York City multidrug-resistant Mycobacterium tuberculosis clone family. JAMA. 1996;275. Available: https://pubmed.ncbi.nlm.nih.gov/8627966/

11. Daley CL, Small PM, Schecter GF, Schoolnik GK, McAdam RA, Jacobs WR, et al. An outbreak of tuberculosis with accelerated progression among persons infected with the human immunodeficiency virus. An analysis using restriction-fragment-length polymorphisms. N Engl J Med. 1992;326. doi:10.1056/NEJM199201233260404

12. Pena MJ, Caminero JA, Campos-Herrero MI, Rodríguez-Gallego JC, García-Laorden MI, Cabrera P, et al. Epidemiology of tuberculosis on Gran Canaria: a 4 year population study using traditional and molecular approaches. Thorax. 2003;58. doi:10.1136/thorax.58.7.618

13. Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, et al. Transmission of Tuberculosis in New York City -- An Analysis by DNA Fingerprinting and Conventional Epidemiologic Methods. New England Journal of Medicine. 1994. pp. 1710–1716. doi:10.1056/nejm199406163302403

14. ten Asbroek AH, Borgdorff MW, Nagelkerke NJ, Sebek MM, Devillé W, van Embden JD, et al. Estimation of serial interval and incubation period of tuberculosis using DNA fingerprinting. Int J Tuberc Lung Dis. 1999;3. Available: https://pubmed.ncbi.nlm.nih.gov/10331731/

15. Umene K, Sakaoka H. Populations of two eastern countries of Japan and Korea and with a related history share a predominant genotype of herpes simplex virus type 1. Arch Virol. 1997;142. doi:10.1007/s007050050213

16. Coimbra RS, Grimont F, Lenormand P, Burguière P, Beutin L, Grimont PA. Identification of Escherichia coli O-serogroups by restriction of the amplified O-antigen gene cluster (rfb-RFLP). Res Microbiol. 2000;151: 639–654.

17. Wichelhaus TA, Hunfeld KP, Böddinghaus B, Kraiczy P, Schäfer V, Brade V. Rapid molecular typing of methicillin-resistant Staphylococcus aureus by PCR-RFLP. Infect Control Hosp Epidemiol. 2001;22: 294–298.

18. Drazek ES, Dubois A, Holmes RK. Characterization and presumptive identification of Helicobacter pylori isolates from rhesus monkeys. J Clin Microbiol. 1994;32: 1799–1804.

19. Davis AH, None JW, Tsang TC, Harris DT. Direct sequencing is more accurate and feasible in detecting single nucleotide polymorphisms than RFLP: using human vascular endothelial growth factor gene as a model. Biol Res Nurs. 2007;9. doi:10.1177/1099800407308083

20. Tanahashi T, Kita M, Kodama T, Sawai N, Yamaoka Y, Mitsufuji S, et al. Comparison of PCR-Restriction Fragment Length Polymorphism Analysis and PCR-Direct Sequencing Methods for Differentiating Helicobacter pylori ureB Gene Variants. J Clin Microbiol. 2000;38: 165.

21. Fitch WM, Bush RM, Bender CA, Cox NJ. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci U S A. 1997;94: 7712–7718.

22. Bush RM, Fitch WM, Bender CA, Cox NJ. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol. 1999;16: 1457–1465.

23. Chambers BS, Parkhouse K, Ross TM, Alby K, Hensley SE. Identification of Hemagglutinin Residues Responsible for H3N2 Antigenic Drift during the 2014-2015 Influenza Season. Cell Rep. 2015;12: 1–6.

24. Hay AJ, Gregory V, Douglas AR, Lin YP. The evolution of human influenza viruses. Philos Trans R Soc Lond B Biol Sci. 2001;356: 1861–1870.

25. Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. Biol Direct. 2006;1: 1–19.

26. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol. 1999;73. doi:10.1128/JVI.73.12.10489-10502.1999

27. Sudderuddin H, Kinloch NN, Jin SW, Miller RL, Jones BR, Brumme CJ, et al. Longitudinal within-host evolution of HIV Nef-mediated CD4, HLA and SERINC5 downregulation activity: a case study. Retrovirology. 2020;17: 1–10.

28. Barton JP, Goonetilleke N, Butler TC, Walker BD, McMichael AJ, Chakraborty AK. Relative

rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nat Commun. 2016;7: 1–10.

29. Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, et al. Antibody neutralization and escape by HIV-1. Nature. 2003;422: 307–312.

30. Lemey P, Rambaut A, Pybus OG. HIV evolutionary dynamics within and among hosts. AIDS Rev. 2006;8. Available: https://pubmed.ncbi.nlm.nih.gov/17078483/

31. Kingman JFC. Origins of the Coalescent: 1974-1982. Genetics. 2000;156: 1461–1463.

32. Wang J, Santiago E, Caballero A. Prediction and estimation of effective population size. Heredity . 2016;117: 193–206.

33. Kuhner MK, Yamato J, Felsenstein J. Maximum Likelihood Estimation of Population Growth Rates Based on the Coalescent. Genetics. 1998;149: 429–434.

34. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7: 1–8.

35. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics. 2002;161: 1307.

36. Beaumont MA. Detecting population expansion and decline using microsatellites. Genetics. 1999;153: 2013.

37. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17. doi:10.1093/bioinformatics/17.8.754

38. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Déirdre Hollingsworth T, et al. Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings. Science. 2009;324: 1557.

39. Frost SDW, Volz EM. Viral phylodynamics and the search for an “effective number of infections.” Philos Trans R Soc Lond B Biol Sci. 2010;365: 1879–1890.

40. Volz EM, Koopman JS, Ward MJ, Brown AL, Frost SDW. Simple Epidemiological Dynamics Explain Phylogenetic Clustering of HIV from Patients with Recent Infection. PLoS Comput Biol. 2012;8: e1002552.

41. Popinga A, Vaughan T, Stadler T, Drummond AJ. Inferring Epidemiological Dynamics with Bayesian Coalescent Inference: The Merits of Deterministic and Stochastic Models. Genetics. 2015;199: 595–607.

42. Kendall DG. On the Generalized “Birth-and-Death” Process. Ann Math Stat. 1948;19: 1–15.

43. Stadler T, Kouyos R, von Wyl V, Yerly S, Böni J, Bürgisser P, et al. Estimating the basic reproductive number from viral sequence data. Mol Biol Evol. 2012;29. doi:10.1093/molbev/msr217

44. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci U S A. 2013;110: 228–233.

45. Louca S, Pennell MW. Extant timetrees are consistent with a myriad of diversification histories. Nature. 2020;580: 502–505.

46. Stadler T, Vaughan TG, Gavryushkin A, Guindon S, Kühnert D, Leventhal GE, et al. How well can the exponential-growth coalescent approximate constant-rate birth-death population dynamics? Proc Biol Sci. 2015;282: 20150420.

47. Louca S, McLaughlin A, MacPherson A, Joy JB, Pennell MW. Fundamental Identifiability Limits in Molecular Epidemiology. Mol Biol Evol. 2021;38: 4010–4024.

48. Desai MM, Walczak AM, Fisher DS. Genetic Diversity and the Structure of Genealogies in Rapidly Adapting Populations. Genetics. 2013;193: 565.

49. Adel Dayarian BIS. How to Infer Relative Fitness from a Sample of Genomic Sequences. Genetics. 2014;197: 913.

50. Rasmussen DA, Stadler T. Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models. 2019 [cited 26 Aug 2021]. doi:10.7554/eLife.45562

51. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009;5. doi:10.1371/journal.pcbi.1000520

52. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. Mol Biol Evol. 2010;27. doi:10.1093/molbev/msq067

53. [No title]. [cited 5 Oct 2021]. Available: http://apps.who.int/iris/bitstream/handle/10665/208883/ebolasitrep_10Jun2016_eng.pdf

54. Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A, et al. Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. Nature. 2015;524: 97–101.

55. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017;544: 309–315.

56. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N 'faly, et al. Emergence of Zaire Ebola virus disease in Guinea. N Engl J Med. 2014;371: 1418–1425.

57. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345: 1369–1372.

58. Arias A, Watson SJ, Asogun D, Tobin EA, Lu J, Phan MVT, et al. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. Virus Evol. 2016;2. doi:10.1093/ve/vew016

59. Dellicour S, Baele G, Dudas G, Faria NR, Pybus OG, Suchard MA, et al. Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. Nat Commun. 2018;9: 1–9.

60. Hazenberg MD, Otto SA, van Benthem BHB, Roos MT, Coutinho RA, Lange JMA, et al. Persistent immune activation in HIV-1 infection is associated with progression to AIDS. AIDS. 2003. pp. 1881–1888. doi:10.1097/00002030-200309050-00006

61. Marlink R, Kanki P, Thior I, Travers K, Eisen G, Siby T, et al. Reduced rate of disease development after HIV-2 infection as compared to HIV-1. Science. 1994;265: 1587–1590.

62. Gilbert PB, McKeague IW, Eisen G, Mullins C, Guéye-NDiaye A, Mboup S, et al. Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal. Stat Med. 2003;22: 573–593.

63. Reeves JD, Doms RW. Human immunodeficiency virus type 2. J Gen Virol. 2002;83: 1253–1265.

64. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW Jr, Swanstrom R, et al. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature. 2009;460: 711–716.

65. Cohn LB, Silva IT, Oliveira TY, Rosales RA, Parrish EH, Learn GH, et al. HIV-1 integration landscape during latent and active infection. Cell. 2015;160: 420–432.

66. Parada CA, Roeder RG. A novel RNA polymerase II-containing complex potentiates Tat-enhanced HIV-1 transcription. EMBO J. 1999;18: 3688–3701.

67. Schwartz S, Felber BK, Benko DM, Fenyö EM, Pavlakis GN. Cloning and functional analysis of multiply spliced mRNA species of human immunodeficiency virus type 1. J Virol. 1990;64. doi:10.1128/JVI.64.6.2519-2529.1990

68. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, et al. Antiretroviral Therapy for the Prevention of HIV-1 Transmission. N Engl J Med. 2016;375: 830–839.

69. Alimonti JB, Ball TB, Fowke KR. Mechanisms of CD4+ T lymphocyte cell death in human immunodeficiency virus infection and AIDS. J Gen Virol. 2003;84: 1649–1661.

70. Terai C, Kornbluth RS, Pauza CD, Richman DD, Carson DA. Apoptosis as a mechanism of cell death in cultured T lymphoblasts acutely infected with HIV-1. Journal of Clinical Investigation. 1991. pp. 1710–1715. doi:10.1172/jci115188

71. Li CJ, Friedman DJ, Wang C, Metelev V, Pardee AB. Induction of apoptosis in uninfected lymphocytes by HIV-1 Tat protein. Science. 1995;268: 429–431.

72. Doitsh G, Galloway NLK, Geng X, Yang Z, Monroe KM, Zepeda O, et al. Cell death by pyroptosis drives CD4 T-cell depletion in HIV-1 infection. Nature. 2013;505: 509–514.

73. Borrow P, Lewicki H, Hahn BH, Shaw GM, Oldstone MB. Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. J Virol. 1994;68: 6103–6110.

74. Mukadi Y, Perriëns JH, St Louis ME, Brown C, Prignot J, Willame JC, et al. Spectrum of immunodeficiency in HIV-1-infected patients with pulmonary tuberculosis in Zaire. Lancet. 1993;342: 143–146.

75. Bour S, Geleziunas R, Wainberg MA. The human immunodeficiency virus type 1 (HIV-1) CD4 receptor and its central role in promotion of HIV-1 infection. Microbiol Rev. 1995;59: 63–93.

76. Jordan CA, Watkins BA, Kufta C, Dubois-Dalcq M. Infection of brain microglial cells by human immunodeficiency virus type 1 is CD4 dependent. J Virol. 1991;65: 736.

77. Bleul CC, Wu L, Hoxie JA, Springer TA, Mackay CR. The HIV coreceptors CXCR4 and CCR5 are differentially expressed and regulated on human T lymphocytes. Proceedings of the National Academy of Sciences. 1997. pp. 1925–1930. doi:10.1073/pnas.94.5.1925

78. Dimitrov AS, Louis JM, Bewley CA, Marius Clore G, Blumenthal R. Conformational Changes in HIV-1 gp41 in the Course of HIV-1 Envelope Glycoprotein-Mediated Fusion and Inactivation. Biochemistry. 2005;44: 12471.

79. Tae-Wook Chun ASF. Latent reservoirs of HIV: Obstacles to the eradication of virus. Proc Natl Acad Sci U S A. 1999;96: 10958.

80. Abram ME, Ferris AL, Shao W, Alvord WG, Hughes SH. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. J Virol. 2010;84: 9864–9878.

81. Smith RA, Loeb LA, Preston BD. Lethal mutagenesis of HIV. Virus Research. 2005. pp. 215–228. doi:10.1016/j.virusres.2004.11.011

82. Zanini F, Puller V, Brodin J, Albert J, Neher RA. In vivo mutation rates and the landscape of fitness costs of HIV-1. Virus Evol. 2017;3. doi:10.1093/ve/vex003

83. Roberts JD, Bebenek K, Kunkel TA. The accuracy of reverse transcriptase from HIV-1. Science. 1988;242: 1171–1173.

84. Sadler HA, Stenglein MD, Harris RS, Mansky LM. APOBEC3G contributes to HIV-1 variation through sublethal mutagenesis. J Virol. 2010;84: 7396–7404.

85. Rhodes TD, Nikolaitchik O, Chen J, Powell D, Hu W-S. Genetic Recombination of Human Immunodeficiency Virus Type 1 in One Round of Viral Replication: Effects of Genetic Distance, Target Cells, Accessory Genes, and Lack of High Negative Interference in Crossover Events. J Virol. 2005;79: 1666.

86. Holmes EC. Adaptation and immunity. PLoS Biol. 2004;2: E307.

87. Tripathi K, Balagam R, Vishnoi NK, Dixit NM. Stochastic Simulations Suggest that HIV-1 Survives Close to Its Error Threshold. PLoS Comput Biol. 2012;8: e1002684.

88. Loeb LA, Essigmann JM, Kazazi F, Zhang J, Rose KD, Mullins JI. Lethal mutagenesis of HIV with mutagenic nucleoside analogs. Proc Natl Acad Sci U S A. 1999;96: 1492.

89. Alizon S, Fraser C. Within-host and between-host evolutionary rates across the HIV-1 genome. Retrovirology. 2013;10: 49.

90. Zhu T, Mo H, Wang N, Nam DS, Cao Y, Koup RA, et al. Genotypic and phenotypic characterization of HIV-1 patients with primary infection. Science. 1993;261: 1179–1181.

91. Wolinsky SM, Wike CM, Korber BT, Hutto C, Parks WP, Rosenblum LL, et al. Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. Science. 1992;255: 1134–1137.

92. Derdeyn CA, Decker JM, Bibollet-Ruche F, Mokili JL, Muldoon M, Denham SA, et al. Envelope-Constrained Neutralization-Sensitive HIV-1 After Heterosexual Transmission. Science. 2004;303: 2019–2022.

93. Haaland RE, Hawkins PA, Salazar-Gonzalez J, Johnson A, Tichacek A, Karita E, et al.

Inflammatory Genital Infections Mitigate a Severe Genetic Bottleneck in Heterosexual Transmission of Subtype A and C HIV-1. PLoS Pathog. 2009;5: e1000274.

94. Carlson JM, Schaefer M, Monaco DC, Batorsky R, Claiborne DT, Prince J, et al. Selection bias at the heterosexual HIV-1 transmission bottleneck. Science. 2014;345. doi:10.1126/science.1254031

95. Tully DC, Ogilvie CB, Batorsky RE, Bean DJ, Power KA, Ghebremichael M, et al. Differences in the Selection Bottleneck between Modes of Sexual Transmission Influence the Genetic Composition of the HIV-1 Founder Virus. PLoS Pathog. 2016;12: e1005619.

96. Li H, Bar KJ, Wang S, Decker JM, Chen Y, Sun C, et al. High Multiplicity Infection by HIV-1 in Men Who Have Sex with Men. PLoS Pathog. 2010;6: e1000890.

97. Bar KJ, Li H, Chamberland A, Tremblay C, Routy JP, Grayson T, et al. Wide variation in the multiplicity of HIV-1 infection among injection drug users. J Virol. 2010;84. doi:10.1128/JVI.00077-10

98. Masharsky AE, Dukhovlinova EN, Verevochkin SV, Toussova OV, Skochilov RV, Anderson JA, et al. A Substantial Transmission Bottleneck among Newly and Recently HIV-1-Infected Injection Drug Users in St Petersburg, Russia. J Infect Dis. 2010;201: 1697–1702.

99. Immonen TT, Leitner T. Reduced evolutionary rates in HIV-1 reveal extensive latency periods among replicating lineages. Retrovirology. 2014;11: 1–11.

100. Goepfert PA, Lumm W, Farmer P, Matthews P, Prendergast A, Carlson JM, et al. Transmission of HIV-1 Gag immune escape mutations is associated with reduced viral load in linked recipients. J Exp Med. 2008;205: 1009–1017.

101. Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, Feeney M, et al. HIV evolution: CTL escape mutation and reversion after transmission. Nat Med. 2004;10: 282–289.

102. Martinez-Picado J, Prado JG, Fry EE, Pfafferott K, Leslie A, Chetty S, et al. Fitness Cost of Escape Mutations in p24 Gag in Association with Control of Human Immunodeficiency Virus Type 1. J Virol. 2006;80: 3617.

103. Thobakgale CF, Prendergast A, Crawford H, Mkhwanazi N, Ramduth D, Reddy S, et al. Impact of HLA in Mother and Child on Disease Progression of Pediatric Human Immunodeficiency Virus Type 1 Infection. J Virol. 2009;83: 10234.

104. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, et al. Population genomics of intrapatient HIV-1 evolution. 2015 [cited 25 Sep 2021]. doi:10.7554/eLife.11282

105. Connor RI, Sheridan KE, Ceradini D, Choe S, Landau NR. Change in Coreceptor Use Correlates with Disease Progression in HIV-1–Infected Individuals. J Exp Med. 1997;185: 621–628.

106. Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, et al. Identification of a major co-receptor for primary isolates of HIV-1. Nature. 1996;381: 661–666.

107. Long EM, Rainwater SMJ, Lavreys L, Mandaliya K, Overbaugh J. HIV type 1 variants transmitted to women in Kenya require the CCR5 coreceptor for entry, regardless of the genetic complexity of the infecting virus. AIDS Res Hum Retroviruses. 2002;18: 567–576.

108. Clevestig P, Maljkovic I, Casper C, Carlenor E, Lindgren S, Navér L, et al. The X4 phenotype of HIV type 1 evolves from R5 in two children of mothers, carrying X4, and is not linked to transmission. AIDS Res Hum Retroviruses. 2005;21: 371–378.

109. Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, Horuk R, et al. Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. Cell. 1996;86: 367–377.

110. Ioannidis JP, Rosenberg PS, Goedert JJ, Ashton LJ, Benfield TL, Buchbinder SP, et al. Effects of CCR5-Delta32, CCR2-64I, and SDF-1 3'A alleles on HIV-1 disease progression: An international meta-analysis of individual-patient data. Ann Intern Med. 2001;135. doi:10.7326/0003-4819-135-9-200111060-00008

111. Meyer L, Magierowska M, Hubert JB, Mayaux MJ, Misrahi M, Le Chenadec J, et al. CCR5 delta32 deletion and reduced risk of toxoplasmosis in persons infected with human immunodeficiency virus type 1. The SEROCO-HEMOCO-SEROGEST Study Groups. J Infect Dis. 1999;180. doi:10.1086/314933

112. Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a zoonosis: scientific and public health implications. Science. 2000;287. doi:10.1126/science.287.5453.607

113. Paul M. Sharp BHH. Origins of HIV and the AIDS Pandemic. Cold Spring Harb Perspect Med. 2011;1. doi:10.1101/cshperspect.a006841

114. Yamaguchi J, Devare SG, Brennan CA. Identification of a new HIV-2 subtype based on phylogenetic analysis of full-length genomic sequence. AIDS Res Hum Retroviruses. 2000;16. doi:10.1089/08892220050042864

115. Plantier J-C, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemée V, et al. A new human immunodeficiency virus derived from gorillas. Nat Med. 2009;15: 871–872.

116. Ayouba A, Akoua-Koffi C, Calvignac-Spencer S, Esteban A, Locatelli S, Li H, et al. Evidence for continuing cross-species transmission of SIVsmm to humans: characterization of a new HIV-2 lineage in rural Côte d'Ivoire. AIDS. 2013;27. doi:10.1097/01.aids.0000432443.22684.50

117. Rambaut A, Robertson DL, Pybus OG, Peeters M, Holmes EC. Phylogeny and the origin of HIV-1. Nature. 2001. pp. 1047–1048. doi:10.1038/35074179

118. Huang A, Hogan JW, Istrail S, DeLong A, Katzenstein DA, Kantor R. Global analysis of sequence diversity within HIV-1 subtypes across geographic regions. Future Virol. 2012;7: 505.

119. Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, Ilunga W, et al. Unprecedented Degree of Human Immunodeficiency Virus Type 1 (HIV-1) Group M Genetic Diversity in the Democratic Republic of Congo Suggests that the HIV-1 Pandemic Originated in Central Africa. J Virol. 2000;74: 10498.

120. Kalish ML, Wolfe ND, Ndongmo CB, McNicholl J, Robbins KE, Aidoo M, et al. Central African hunters exposed to simian immunodeficiency virus. Emerg Infect Dis. 2005;11: 1928–1930.

121. de Sousa JD, Müller V, Lemey P, Vandamme A-M. High GUD incidence in the early 20 century created a particularly permissive time window for the origin and initial spread of epidemic HIV strains. PLoS One. 2010;5: e9936.

122. Chitnis A, Rawls D, Moore J. Origin of HIV type 1 in colonial French Equatorial Africa? AIDS Res Hum Retroviruses. 2000;16. doi:10.1089/088922200309548

123. Sauter D, Kirchhoff F. Key Viral Adaptations Preceding the AIDS Pandemic. Cell Host Microbe. 2019;25: 27–38.

124. Le Tortorec A, Willey S, Neil SJD. Antiviral Inhibition of Enveloped Virus Release by Tetherin/BST-2: Action and Counteraction. Viruses. 2011;3: 520.

125. Perez-Caballero D, Zang T, Ebrahimi A, McNatt MW, Gregory DA, Johnson MC, et al. Tetherin inhibits HIV-1 release by directly tethering virions to cells. Cell. 2009;139: 499.

126. McNatt MW, Zang T, Bieniasz PD. Vpu binds directly to tetherin and displaces it from nascent virions. PLoS Pathog. 2013;9. doi:10.1371/journal.ppat.1003299

127. Sauter D, Schindler M, Specht A, Landford WN, Münch J, Kim K-A, et al. Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains. Cell Host Microbe. 2009;6: 409–421.

128. Jia B, Serra-Moreno R, Neidermyer W, Rahmberg A, Mackey J, Fofana IB, et al. Species-specific activity of SIV Nef and HIV-1 Vpu in overcoming restriction by tetherin/BST2. PLoS Pathog. 2009;5. doi:10.1371/journal.ppat.1000429

129. Kmiec D, Iyer SS, Stürzel CM, Sauter D, Hahn BH, Kirchhoff F. Vpu-Mediated Counteraction of Tetherin Is a Major Determinant of HIV-1 Interferon Resistance. MBio. 2016;7. doi:10.1128/mBio.00934-16

130. Sauter D, Unterweger D, Vogl M, Usmani SM, Heigele A, Kluge SF, et al. Human Tetherin Exerts Strong Selection Pressure on the HIV-1 Group N Vpu Protein. PLoS Pathog. 2012;8: e1003093.

131. Kluge SF, Mack K, Iyer SS, Pujol FM, Heigele A, Learn GH, et al. Nef Proteins of Epidemic HIV-1 Group O Strains Antagonize Human Tetherin. Cell Host Microbe. 2014;16: 639.

132. Mack K, Starz K, Sauter D, Langer S, Bibollet-Ruche F, Learn GH, et al. Efficient Vpu-Mediated Tetherin Antagonism by an HIV-1 Group O Strain. J Virol. 2017;91. doi:10.1128/JVI.02177-16

133. Sauter D, Hué S, Petit SJ, Plantier J-C, Towers GJ, Kirchhoff F, et al. HIV-1 Group P is unable to antagonize human tetherin by Vpu, Env or Nef. Retrovirology. 2011;8: 1–9.

134. Exline CM, Yang SJ, Haworth KG, Rengarajan S, Lopez LA, Droniou ME, et al. Determinants in HIV-2 Env and tetherin required for functional interaction. Retrovirology. 2015;12. doi:10.1186/s12977-015-0194-0

135. Le Tortorec A, Neil SJ. Antagonism to and intracellular sequestration of human tetherin by the human immunodeficiency virus type 2 envelope glycoprotein. J Virol. 2009;83. doi:10.1128/JVI.01515-09

136. Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA, et al. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. N Engl J Med. 1981;305. doi:10.1056/NEJM198112103052401

137. Friedman-Kien AE. Disseminated Kaposi's sarcoma syndrome in young homosexual men. J Am Acad Dermatol. 1981;5: 468–471.

138. Barré-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, et al. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science. 1983;220: 868–871.

139. Pape JW, Liautaud B, Thomas F, Mathurin JR, St Amand MM, Boncy M, et al. Characteristics of the acquired immunodeficiency syndrome (AIDS) in Haiti. N Engl J Med. 1983;309. doi:10.1056/NEJM198310203091603

140. Cliff AD, Smallman-Raynor MR. The aids pandemic: Global geographical patterns and local spatial processes. Geogr J. 1992;158: 182.

141. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, et al. Timing the ancestor of the HIV-1 pandemic strains. Science. 2000;288. doi:10.1126/science.288.5472.1789

142. Zhu T, Korber BT, Nahmias AJ, Hooper E, Sharp PM, Ho DD. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature. 1998;391: 594–597.

143. Dearlove B, Tovanabutra S, Owen CL, Lewitus E, Li Y, Sanders-Buell E, et al. Factors influencing estimates of HIV-1 infection timing using BEAST. PLoS Comput Biol. 2021;17: e1008537.

144. Gilbert MTP, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. The emergence of HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci U S A. 2007;104: 18566–18570.

145. Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science. 2014;346: 56–61.

146. Salemi M, Strimmer K, Hall WW, Duffy M, Delaporte E, Mboup S, et al. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. FASEB J. 2001;15: 276–278.

147. Gryseels S, Watts TD, Kabongo Mpolesha J-M, Larsen BB, Lemey P, Muyembe-Tamfum J-J, et al. A near full-length HIV-1 genome from 1966 recovered from formalin-fixed paraffin-embedded tissue. Proc Natl Acad Sci U S A. 2020;117: 12222–12229.

148. Bbosa N, Kaleebu P, Ssemwanga D. HIV subtype diversity worldwide. Curr Opin HIV AIDS. 2019;14: 153–160.

149. Gartner MJ, Roche M, Churchill MJ, Gorry PR, Flynn JK. Understanding the mechanisms driving the spread of subtype C HIV-1. EBioMedicine. 2020;53: 102682.

150. Pitchenik AE, Fischl MA, Dickinson GM, Becker DM, Fournier AM, O'Connell MT, et al. Opportunistic infections and Kaposi's sarcoma among Haitians: evidence of a new acquired immunodeficiency state. Ann Intern Med. 1983;98: 277–284.

151. Li WH, Tanimura M, Sharp PM. Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol. 1988;5: 313–330.

152. Jackson RO. The failure of categories: Haitians in the united nations organization in the Congo, 1960–64. J Haitian Stud. 2014;20: 34–64.

153. Worobey M, Watts TD, McKay RA, Suchard MA, Granade T, Teuwen DE, et al. 1970s and "Patient 0" HIV-1 genomes illuminate early HIV/AIDS history in North America. Nature. 2016;539: 98–101.

154. Paraskevis D, Pybus O, Magiorkinis G, Hatzakis A, Wensing AM, van de Vijver DA, et al. Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. Retrovirology. 2009;6: 49.

155. Magiorkinis G, Angelis K, Mamais I, Katzourakis A, Hatzakis A, Albert J, et al. The global spread of HIV-1 subtype B epidemic. Infect Genet Evol. 2016;46: 169–179.

156. Beyrer C, Baral SD, van Griensven F, Goodreau SM, Chariyalertsak S, Wirtz AL, et al. Global epidemiology of HIV infection in men who have sex with men. Lancet. 2012;380: 367–377.

157. Junqueira DM, Almeida SE de M. HIV-1 subtype B: Traces of a pandemic. Virology. 2016;495: 173–184.

158. Tongo M, Harkins GW, Dorfman JR, Billings E, Tovanabutra S, de Oliveira T, et al. Unravelling the complicated evolutionary and dissemination history of HIV-1M subtype A lineages. Virus Evol. 2018;4. doi:10.1093/ve/vey003

159. Gao F, Robertson DL, Morrison SG, Hui H, Craig S, Decker J, et al. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. J Virol. 1996;70. doi:10.1128/JVI.70.10.7013-7029.1996

160. Delatorre E, Couto-Fernandez JC, Guimarães ML, Vaz Cardoso LP, de Alcantara KC, Stefani MM de A, et al. Tracing the origin and northward dissemination dynamics of HIV-1 subtype C in Brazil. PLoS One. 2013;8: e74072.

161. Carr JK, Salminen MO, Albert J, Sanders-Buell E, Gotte D, Birx DL, et al. Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants. Virology. 1998;247. doi:10.1006/viro.1998.9211

162. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V. Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull. 2001;58: 19–42.

163. Désiré N, Cerutti L, Le Hingrat Q, Perrier M, Emler S, Calvez V, et al. Characterization update of HIV-1 M subtypes diversity and proposal for subtypes A and D sub-subtypes reclassification. Retrovirology. 2018;15: 80.

164. Palm AA, Esbjörnsson J, Månsson F, Kvist A, Isberg PE, Biague A, et al. Faster progression to AIDS and AIDS-related death among seroincident individuals infected with recombinant HIV-1 A3/CRF02_AG compared with sub-subtype A3. J Infect Dis. 2014;209. doi:10.1093/infdis/jit416

165. Kaleebu P, Ross A, Morgan D, Yirrell D, Oram J, Rutebemberwa A, et al. Relationship between HIV-1 Env subtypes A and D and disease progression in a rural Ugandan cohort. AIDS. 2001;15. doi:10.1097/00002030-200102160-00001

166. Kyeyune F, Nankya I, Metha S, Akao J, Ndashimye E, Tebit DM, et al. Treatment failure and drug resistance is more frequent in HIV-1 subtype D versus subtype A-infected Ugandans over a 10-year study period. AIDS. 2013;27. doi:10.1097/QAD.0b013e3283610ec7

167. Abraha A, Nankya IL, Gibson R, Demers K, Tebit DM, Johnston E, et al. CCR5- and CXCR4-tropic subtype C human immunodeficiency virus type 1 isolates have a lower level of pathogenic fitness than other dominant group M subtypes: implications for the epidemic. J Virol. 2009;83. doi:10.1128/JVI.02051-08

168. Kiguoya MW, Mann JK, Chopera D, Gounder K, Lee GQ, Hunt PW, et al. Subtype-Specific Differences in Gag-Protease-Driven Replication Capacity Are Consistent with Intersubtype Differences in HIV-1 Disease Progression. J Virol. 2017;91. doi:10.1128/JVI.00253-17

169. Konings FA, Burda ST, Urbanski MM, Zhong P, Nadas A, Nyambi PN. Human immunodeficiency virus type 1 (HIV-1) circulating recombinant form 02_AG (CRF02_AG) has a higher in vitro replicative capacity than its parental subtypes A and G. J Med Virol. 2006;78. doi:10.1002/jmv.20572

170. Matthews PC, Koyanagi M, Kløverpris HN, Harndahl M, Stryhn A, Akahoshi T, et al. Differential clade-specific HLA-B*3501 association with HIV-1 disease outcome is linked to immunogenicity of a single Gag epitope. J Virol. 2012;86: 12643–12654.

171. Amornkul PN, Karita E, Kamali A, Rida WN, Sanders EJ, Lakhi S, et al. Disease progression by infecting HIV-1 subtype in a seroconverter cohort in sub-Saharan Africa. AIDS. 2013;27. doi:10.1097/QAD.0000000000000012

172. Venner CM, Nankya I, Kyeyune F, Demers K, Kwok C, Chen PL, et al. Infecting HIV-1 Subtype Predicts Disease Progression in Women of Sub-Saharan Africa. EBioMedicine. 2016;13. doi:10.1016/j.ebiom.2016.10.014

173. Touloumi G, Pantazis N, Pillay D, Paraskevis D, Chaix ML, Bucher HC, et al. Impact of HIV-1 subtype on CD4 count at HIV seroconversion, rate of decline, and viral load set point in European seroconverter cohorts. Clin Infect Dis. 2013;56. doi:10.1093/cid/cis1000

174. Faria NR, Vidal N, Lourenco J, Raghwani J, Sigaloff KCE, Tatem AJ, et al. Distinct rates and patterns of spread of the major HIV-1 subtypes in Central and East Africa. PLoS Pathog. 2019;15: e1007976.

175. Leitner T, Escanilla D, Franzén C, Uhlén M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc Natl Acad Sci U S A. 1996;93: 10864.

176. Sl KP, Weaver S, Aj LB, Wertheim JO. HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. Mol Biol Evol. 2018;35. doi:10.1093/molbev/msy016

177. Pilon R, Leonard L, Kim J, Vallee D, De Rubeis E, Jolly AM, et al. Transmission Patterns of HIV and Hepatitis C Virus among Networks of People Who Inject Drugs. PLoS One. 2011;6. doi:10.1371/journal.pone.0022245

178. Hué S, Pillay D, Clewley JP, Pybus OG. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc Natl Acad Sci U S A. 2005;102: 4425–4429.

179. Ragonnet-Cronin M, Hodcroft E, Hué S, Fearnhill E, Delpech V, Brown AJL, et al. Automated analysis of phylogenetic clusters. BMC Bioinformatics. 2013;14: 317.

180. Novitsky V, Moyo S, Lei Q, DeGruttola V, Essex M. Impact of Sampling Density on the

Extent of HIV Clustering. AIDS Res Hum Retroviruses. 2014;30: 1226.

181. Hassan AS, Pybus OG, Sanders EJ, Albert J, Esbjörnsson J. Defining HIV-1 transmission clusters based on sequence data. AIDS. 2017;31: 1211–1222.

182. Ragonnet-Cronin M, Ofner-Agostini M, Merks H, Pilon R, Rekart M, Archibald CP, et al. Longitudinal phylogenetic surveillance identifies distinct patterns of cluster dynamics. J Acquir Immune Defic Syndr. 2010;55: 102–108.

183. Chalmet K, Staelens D, Blot S, Dinakis S, Pelgrom J, Plum J, et al. Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. BMC Infect Dis. 2010;10: 262.

184. Pao D, Fisher M, Hué S, Dean G, Murphy G, Cane PA, et al. Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. AIDS. 2005;19: 85–90.

185. Dennis AM, Volz E, Frost SDW, Md. Mukarram Hossain A, Poon AFY, Rebeiro PF, et al. HIV-1 Transmission Clustering and Phylodynamics Highlight the Important Role of Young Men Who Have Sex with Men. AIDS Res Hum Retroviruses. 2018;34: 879.

186. Lubelchek RJ, Hoehnen SC, Hotton AL, Kincaid SL, Barker DE, French AL. Transmission clustering among newly diagnosed HIV patients in Chicago, 2008 to 2011: using phylogenetics to expand knowledge of regional HIV transmission patterns. J Acquir Immune Defic Syndr. 2015;68: 46–54.

187. Paraskevis D, Beloukas A, Stasinos K, Pantazis N, de Mendoza C, Bannert N, et al. HIV-1 molecular transmission clusters in nine European countries and Canada: association with demographic and clinical factors. BMC Med. 2019;17: 4.

188. Beyrer C, Sullivan P, Sanchez J, Baral SD, Collins C, Wirtz AL, et al. The increase in global HIV epidemics in MSM. AIDS. 2013;27: 2665–2678.

189. Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, Leigh Brown AJ, et al. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathog. 2009;5: e1000590.

190. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med. 2008;5: e50.

191. HIV in the United Kingdom: 2013 Report. [cited 13 Sep 2021]. Available: https://webarchive.nationalarchives.gov.uk/ukgwa/20181112133715mp_/https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/326601/HIV_annual_report_2013.pdf

192. Patel P, Borkowf CB, Brooks JT, Lasry A, Lansky A, Mermin J. Estimating per-act HIV transmission risk: a systematic review. AIDS. 2014;28. doi:10.1097/QAD.0000000000000298

193. Ragonnet-Cronin M, on behalf of the United Kingdom HIV Drug Resistance Database, Lycett SJ, Hodcroft EB, Hué S, Fearnhill E, et al. Transmission of Non-B HIV Subtypes in the United Kingdom Is Increasingly Driven by Large Non-Heterosexual Transmission Clusters. J Infect Dis. 2015;213: 1410–1418.

194. Ragonnet-Cronin ML, Shilaih M, Günthard HF, Hodcroft EB, Böni J, Fearnhill E, et al. A

Direct Comparison of Two Densely Sampled HIV Epidemics: The UK and Switzerland. Sci Rep. 2016;6: 1–9.

195. Poon AFY, Gustafson R, Daly P, Zerr L, Demlow SE, Wong J, et al. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. Lancet HIV. 2016;3: e231–8.

196. Centers for Disease Control (CDC). Update: transmission of HIV infection during invasive dental procedures--Florida. MMWR Morb Mortal Wkly Rep. 1991;40: 377–381.

197. Metzker ML, Mindell DP, Liu X-M, Ptak RG, Gibbs RA, Hillis DM. Molecular evidence of HIV-1 transmission in a criminal case. Proc Natl Acad Sci U S A. 2002;99: 14292–14297.

198. Albert J, Wahlberg J, Leitner T, Escanilla D, Uhlén M. Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 pol and gag genes. J Virol. 1994;68: 5918–5924.

199. Abecasis AB, Pingarilho M, Vandamme A-M. Phylogenetic analysis as a forensic tool in HIV transmission investigations. AIDS. 2018;32: 543.

200. Romero-Severson EO, Bulla I, Leitner T. Phylogenetically resolving epidemiologic linkage. Proc Natl Acad Sci U S A. 2016;113: 2690–2695.

201. Leitner T, Romero-Severson E. Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants. Nat Microbiol. 2018;3: 983–988.

202. Ratmann O, Grabowski MK, Hall M, Golubchik T, Wymant C, Abeler-Dörner L, et al. Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. Nat Commun. 2019;10: 1411.

203. edward_holmes, arambaut, trvrb, cupton, Kristian_Andersen, kihohong. Novel 2019 coronavirus genome. 11 Jan 2020 [cited 24 Sep 2021]. Available: https://virological.org/t/novel-2019-coronavirus-genome/319

204. Quick J. nCoV-2019 sequencing protocol. 2020 [cited 24 Sep 2021]. doi:10.17504/protocols.io.bbmuik6w

205. GISAID - Initiative. [cited 24 Sep 2021]. Available: https://www.gisaid.org/

206. COVID-19 Genomics UK Consortium. 11 Jan 2021 [cited 24 Sep 2021]. Available: https://www.cogconsortium.uk/

207. Lednicky JA, Tagliamonte MS, White SK, Elbadry MA, Alam MM, Stephenson CJ, et al. Emergence of porcine delta-coronavirus pathogenic infections among children in Haiti through independent zoonoses and convergent evolution. medRxiv. 2021. doi:10.1101/2021.03.19.21253391

208. Vlasova AN, Diaz A, Damtie D, Xiu L, Toh T-H, Lee JS-Y, et al. Novel Canine Coronavirus Isolated from a Hospitalized Pneumonia Patient, East Malaysia. Clin Infect Dis. 2021. doi:10.1093/cid/ciab456

209. Feng D, De Vlas SJ, Fang L, Han X, Zhao W, Sheng S, et al. The SARS epidemic in mainland China: bringing together all epidemiological data. Trop Med Int Health. 2009;14: 4.

210. Zumla A, Hui DS, Perlman S. Middle East respiratory syndrome. Lancet. 2015;386: 995–1007.

211. Oran DP, Topol EJ. Prevalence of Asymptomatic SARS-CoV-2 Infection : A Narrative Review. Ann Intern Med. 2020;173: 362–367.

212. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 2020;323: 1239–1242.

213. Bryce C, Grimes Z, Pujadas E, Ahuja S, Beasley MB, Albrecht R, et al. Pathophysiology of SARS-CoV-2: the Mount Sinai COVID-19 autopsy experience. Mod Pathol. 2021;34: 1456–1467.

214. Robba C, Battaglini D, Pelosi P, Rocco PRM. Multiple organ dysfunction in SARS-CoV-2: MODS-CoV-2. Expert Rev Respir Med. 2020;14: 865–868.

215. Parasher A. COVID-19: Current understanding of its Pathophysiology, Clinical presentation and Treatment. Postgrad Med J. 2021;97: 312–320.

216. Yuki K, Fujiogi M, Koutsogiannaki S. COVID-19 pathophysiology: A review. Clin Immunol. 2020;215: 108427.

217. Brian DA, Baric RS. Coronavirus genome structure and replication. Curr Top Microbiol Immunol. 2005;287: 1–30.

218. Robson F, Khan KS, Le TK, Paris C, Demirbag S, Barfuss P, et al. Coronavirus RNA Proofreading: Molecular Basis and Therapeutic Targeting. Mol Cell. 2020;79: 710–727.

219. Bosch BJ, van der Zee R, de Haan CAM, Rottier PJM. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. J Virol. 2003;77: 8801–8811.

220. Belouzard S, Chu VC, Whittaker GR. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. Proc Natl Acad Sci U S A. 2009;106: 5871.

221. Belouzard S, Millet JK, Licitra BN, Whittaker GR. Mechanisms of coronavirus cell entry mediated by the viral spike protein. Viruses. 2012;4: 1011–1033.

222. Cavanagh D, Davis PJ, Pappin DJ, Binns MM, Boursnell ME, Brown TD. Coronavirus IBV: partial amino terminal sequencing of spike polypeptide S2 identifies the sequence Arg-Arg-Phe-Arg-Arg at the cleavage site of the spike precursor propolypeptide of IBV strains Beaudette and M41. Virus Res. 1986;4. doi:10.1016/0168-1702(86)90037-7

223. Millet JK, Whittaker GR. Host cell entry of Middle East respiratory syndrome coronavirus after two-step, furin-mediated activation of the spike protein. Proc Natl Acad Sci U S A. 2014;111: 15214–15219.

224. WHO – COVID19 Vaccine Tracker. [cited 25 Sep 2021]. Available: https://covid19.trackvaccines.org/agency/who/

225. Buddy Creech C, Walker SC, Samuels RJ. SARS-CoV-2 Vaccines. JAMA. 2021;325: 1318–1320.

226. Millet JK, Jaimes JA, Whittaker GR. Molecular diversity of coronavirus host cell entry receptors. FEMS Microbiol Rev. 2021;45. doi:10.1093/femsre/fuaa057

227. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science. 2020;367: 1444–1448.

228. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020;367: 1260–1263.

229. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. Nature. 2020;581: 221–224.

230. Wrobel AG, Benton DJ, Xu P, Roustan C, Martin SR, Rosenthal PB, et al. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat Struct Mol Biol. 2020;27: 763–767.

231. Peacock TP, Goldhill DH, Zhou J, Baillon L, Frise R, Swann OC, et al. The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. Nature Microbiology. 2021;6: 899–909.

232. Wu Y, Zhao S. Furin cleavage sites naturally occur in coronaviruses. Stem Cell Res. 2020;50: 102115.

233. Menachery VD, Graham RL, Baric RS. Jumping species—a mechanism for coronavirus persistence and survival. Curr Opin Virol. 2017;23: 1.

234. Lytras S, Hughes J, Martin D, de Klerk A, Lourens R, Kosakovsky Pond SL, et al. Exploring the natural origins of SARS-CoV-2 in the light of recombination. bioRxiv. 2021. p. 2021.01.22.427830. doi:10.1101/2021.01.22.427830

235. Hu B, Zeng L-P, Yang X-L, Ge X-Y, Zhang W, Li B, et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog. 2017;13: e1006698.

236. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;26: 450–452.

237. Liu P, Jiang J-Z, Wan X-F, Hua Y, Li L, Zhou J, et al. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLoS Pathog. 2020;16: e1008421.

238. Wrobel AG, Benton DJ, Xu P, Calder LJ, Borg A, Roustan C, et al. Structure and binding properties of Pangolin-CoV spike glycoprotein inform the evolution of SARS-CoV-2. Nat Commun. 2021;12: 1–6.

239. Dicken SJ, Murray MJ, Thorne LG, Reuschl A-K, Forrest C, Ganeshalingham M, et al. Characterisation of B.1.1.7 and Pangolin coronavirus spike provides insights on the evolutionary trajectory of SARS-CoV-2. bioRxiv. doi:10.1101/2021.03.22.436468

240. arambaut, Pinned A, Unpinned A, Globally AP, yhg, ssagi, et al. Phylodynamic Analysis. 29 Jan 2020 [cited 24 Sep 2021]. Available: https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356

241. Kristian_Andersen, OliverPybus. Clock and TMRCA based on 27 genomes. 26 Jan 2020 [cited 24 Sep 2021]. Available:

https://virological.org/t/clock-and-tmrca-based-on-27-genomes/347

242. Pekar J, Worobey M, Moshiri N, Scheffler K, Wertheim JO. Timing the SARS-CoV-2 index case in Hubei province. Science. 2021;372: 412–417.

243. Adam DC, Wu P, Wong JY, Lau EHY, Tsang TK, Cauchemez S, et al. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. Nat Med. 2020;26: 1714–1719.

244. Laxminarayan R, Wahl B, Dudala SR, Gopal K, Mohan B C, Neelima S, et al. Epidemiology and transmission dynamics of COVID-19 in two Indian states. Science. 2020;370: 691–697.

245. GISAID - Clade and lineage nomenclature aids in genomic epidemiology of active hCoV-19 viruses. [cited 24 Sep 2021]. Available: https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/

246. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature Microbiology. 2020;5: 1403–1407.

247. CDC. SARS-CoV-2 Variant Classifications and Definitions. 23 Sep 2021 [cited 24 Sep 2021]. Available: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html

248. Alteri C, Cento V, Piralla A, Costabile V, Tallarita M, Colagrossi L, et al. Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. Nat Commun. 2021;12: 1–13.

249. Nadeau SA, Vaughan TG, Scire J, Huisman JS, Stadler T. The origin and early spread of SARS-CoV-2 in Europe. Proc Natl Acad Sci U S A. 2021;118. doi:10.1073/pnas.2012008118

250. Worobey M, Pekar J, Larsen BB, Nelson MI, Hill V, Joy JB, et al. The emergence of SARS-CoV-2 in Europe and North America. Science. 2020;370: 564.

251. Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report. Int J Infect Dis. 2021;103: 234–241.

252. McLaughlin A, Montoya V, Miller RL, Mordecai GJ, Worobey M, Poon AFY, et al. Early and ongoing importations of SARS-CoV-2 in Canada. bioRxiv. medRxiv; 2021. doi:10.1101/2021.04.09.21255131

253. du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science. 2021;371: 708–712.

254. Gámbaro F, Behillil S, Baidaliuk A, Donati F, Albert M, Alexandru A, et al. Introductions and early spread of SARS-CoV-2 in France, 24 January to 23 March 2020. Eurosurveillance. 2020;25: 2001200.

255. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, et al. Introductions and early spread of SARS-CoV-2 in the New York City area. Science. 2020;369: 297–301.

256. Miller D, Martin MA, Harel N, Tirosh O, Kustin T, Meir M, et al. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. Nat Commun. 2020;11:

5518.

257. [No title]. [cited 24 Sep 2021]. Available: https://www.cogconsortium.uk/wp-content/uploads/2021/09/COG-UK-geo-coverage_2021-09-13_summary.pdf

258. Lemey P, Hong SL, Hill V, Baele G, Poletto C, Colizza V, et al. Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. Nat Commun. 2020;11: 1–14.

259. Rahman B, Sadraddin E, Porreca A. The basic reproduction number of SARS−CoV−2 in Wuhan is about to die out, how about the rest of the World? Rev Med Virol. [cited 24 Sep 2021]. doi:10.1002/rmv.2111

260. Estimating the reproductive number R0 of SARS-CoV-2 in the United States and eight European countries and implications for vaccination. J Theor Biol. 2021;517: 110621.

261. Reproductive number of the COVID-19 epidemic in Switzerland with a focus on the Cantons of Basel-Stadt and Basel-Landschaft. Swiss Med Wkly. 2020 [cited 24 Sep 2021]. doi:10.4414/smw.2020.20271

262. Phylodynamic Analyses of outbreaks in China, Italy, Washington State (USA), and the Diamond Princess. 13 Mar 2020 [cited 24 Sep 2021]. Available: https://virological.org/t/phylodynamic-analyses-of-outbreaks-in-china-italy-washington-state-usa-and-the-diamond-princess/439

263. Müller NF, Wagner C, Frazar CD, Roychoudhury P, Lee J, Moncla LH, et al. Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington State. Sci Transl Med. 2021;13. doi:10.1126/scitranslmed.abf0202

264. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness. Nature. 2020;592: 116–121.

265. Zhang L, Jackson CB, Mou H, Ojha A, Peng H, Quinlan BD, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat Commun. 2020;11: 6013.

266. Daniloski Z, Jordan TX, Ilmain JK, Guo X, Bhabha G, tenOever BR, et al. The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. 2021 [cited 24 Sep 2021]. doi:10.7554/eLife.65365

267. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. Cell. 2021;184: 64–75.e11.

268. Kepler L, Hamins-Puertolas M, Rasmussen DA. Decomposing the sources of SARS-CoV-2 fitness variation in the United States. Virus Evol. 2021;7. doi:10.1093/ve/veab073

269. Tracking SARS-CoV-2 variants. [cited 24 Sep 2021]. Available: https://www.who.int/activities/tracking-SARS-CoV-2-variants

270. Thomson EC, Rosen LE, Shepherd JG, Spreafico R, da Silva Filipe A, Wojcechowskyj JA, et al. Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. Cell. 2021;184: 1171–1187.e20.

271. Deng X, Garcia-Knight MA, Khalid MM, Servellita V, Wang C, Morris MK, et al. Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. Cell. 2021;184: 3426–3437.e8.

272. Jangra S, Ye C, Rathnasinghe R, Stadlbauer D, Personalized Virology Initiative study group, Krammer F, et al. SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. Lancet Microbe. 2021;2: e283–e284.

273. Greaney AJ, Loes AN, Crawford KHD, Starr TN, Malone KD, Chu HY, et al. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. Cell Host Microbe. 2021;29: 463–476.e6.

274. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JCC, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. 2020 [cited 24 Sep 2021]. doi:10.7554/eLife.61312

275. Collier DA, De Marco A, Ferreira IATM, Meng B, Datir RP, Walls AC, et al. Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. Nature. 2021;593: 136–141.

276. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell. 2020;182: 1295–1310.e20.

277. Lubinski B, Tang T, Daniel S, Jaimes JA, Whittaker GR. Functional evaluation of proteolytic activation for the SARS-CoV-2 variant B.1.1.7: role of the P681H mutation. bioRxiv. 2021. p. 2021.04.06.438731. doi:10.1101/2021.04.06.438731

278. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science. 2021;372. doi:10.1126/science.abg3055

279. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature. 2021;592: 438–443.

280. Curran J, Dol J, Boulos L, Somerville M, McCulloch H, MacDonald M, et al. Transmission characteristics of SARS-CoV-2 variants of concern Rapid Scoping Review. medRxiv. 2021; 2021.04.23.21255515.

281. Fisman DN, Tuite AR. Progressive Increase in Virulence of Novel SARS-CoV-2 Variants in Ontario, Canada. medRxiv. 2021; 2021.07.05.21260050.

282. Bager P, Wohlfahrt J, Fonager J, Albertsen M, Yssing Michaelsen T, Holten Møller C, et al. Increased Risk of Hospitalisation Associated with Infection with SARS-CoV-2 Lineage B.1.1.7 in Denmark. 2021 [cited 24 Sep 2021]. doi:10.2139/ssrn.3792894

283. Pascall DJ, Mollett G, Blacow R, Bulteel N, Campbell R, Campbell A, et al. The SARS-CoV-2 Alpha variant causes increased clinical severity of disease. medRxiv. 2021; 2021.08.17.21260128.

284. Funk T, Pharris A, Spiteri G, Bundle N, Melidou A, Carr M, et al. Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. Euro Surveill. 2021;26. doi:10.2807/1560-7917.ES.2021.26.16.2100348

285. Twohig KA, Nyberg T, Zaidi A, Thelwall S, Sinnathamby MA, Aliabadi S, et al. Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: a cohort study. Lancet Infect Dis. 2021. doi:10.1016/S1473-3099(21)00475-8

286. Wang P, Nair MS, Liu L, Iketani S, Luo Y, Guo Y, et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature. 2021;593: 130–135.

287. Wibmer CK, Ayres F, Hermanus T, Madzivhandila M, Kgagudi P, Oosthuysen B, et al. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. Nat Med. 2021;27: 622–625.

288. Planas D, Veyer D, Baidaliuk A, Staropoli I, Guivel-Benhassine F, Rajah MM, et al. Reduced sensitivity of infectious SARS-CoV-2 variant B.1.617.2 to monoclonal antibodies and sera from convalescent and vaccinated individuals. bioRxiv. 2021. p. 2021.05.26.445838. doi:10.1101/2021.05.26.445838

289. Emary KRW, Golubchik T, Aley PK, Ariani CV, Angus B, Bibi S, et al. Efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine against SARS-CoV-2 variant of concern 202012/01 (B.1.1.7): an exploratory analysis of a randomised controlled trial. Lancet. 2021;397: 1351–1362.

290. Heath PT, Galiza EP, Baxter DN, Boffito M, Browne D, Burns F, et al. Efficacy of the NVX-CoV2373 Covid-19 Vaccine Against the B.1.1.7 Variant. medRxiv. 2021; 2021.05.13.21256639.

291. Sheikh A, McMenamin J, Taylor B, Robertson C. SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. Lancet. 2021;397: 2461–2462.

292. Ella R, Reddy S, Blackwelder W, Potdar V, Yadav P, Sarangi V, et al. Efficacy, safety, and lot to lot immunogenicity of an inactivated SARS-CoV-2 vaccine (BBV152): a, double-blind, randomised, controlled phase 3 trial. medRxiv. 2021; 2021.06.30.21259439.

293. Barchuk A, Cherkashin M, Bulina A, Berezina N, Rakova T, Kuplevatskaya D, et al. Vaccine Effectiveness against Referral to Hospital and Severe Lung Injury Associated with COVID-19: A Population-Based Case-Control Study in St. Petersburg, Russia. medRxiv. 2021; 2021.08.18.21262065.

294. Roser M, Ritchie H. HIV / AIDS, https://ourworldindata.org/hiv-aids. Our World in Data. 2018.

295. In Russian. Extended HIV report, 2019. http://www.hivrussia.info/wp-content/uploads/2020/12/Byulleten-45-VICH-infektsiya-2019-g..pdf.

296. In Russian. HIV report, 2019. http://www.hivrussia.info/wp-content/uploads/2020/02/VICH-infektsiya-v-Rossijskoj-Federatsii-na-31.12.2019.pdf.

297. Murzakova A, Kireev D, Baryshev P, Lopatukhin A, Serova E, Shemshura A, et al. Molecular Epidemiology of HIV-1 Subtype G in the Russian Federation. Viruses. 2019;11. doi:10.3390/v11040348

298. Thomson MM, Vinogradova A, Delgado E, Rakhmanova A, Yakovlev A, Cuevas MT, et al.

Molecular epidemiology of HIV-1 in St Petersburg, Russia: predominance of subtype A, former Soviet Union variant, and identification of intrasubtype subclusters. J Acquir Immune Defic Syndr. 2009;51: 332–339.

299. Gashnikova NM, Astakhova EM, Gashnikova MP, Bocharov EF, Petrova SV, Pun'ko OA, et al. HIV-1 Epidemiology, Genetic Diversity, and Primary Drug Resistance in the Tyumen Oblast, Russia. Biomed Res Int. 2016;2016. doi:10.1155/2016/2496280

300. Dukhovlinova E, Masharsky A, Toussova O, Verevochkin S, Solovyeva T, Meringof M, et al. Two Independent HIV Epidemics in Saint Petersburg, Russia Revealed by Molecular Epidemiology. AIDS Res Hum Retroviruses. 2015;31: 608–614.

301. Dmitry N, Aleksey L, Marina M, Anatoly B, Sergey S, Ekaterina O, et al. Molecular Surveillance of HIV-1 Infection in Krasnoyarsk Region, Russia: Epidemiology, Phylodynamics and Phylogeography. Curr HIV Res. 2019;17: 114–125.

302. https://en.wikipedia.org/wiki/Oblast.

303. https://en.wikipedia.org/wiki/Federal_district.

304. Hunt M, Gall A, Ong SH, Brener J, Ferns B, Goulder P, et al. IVA: accurate de novo assembly of RNA virus genomes. Bioinformatics. 2015;31: 2374.

305. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120.

306. HIV Sequence Compendium 2018: HIV-1 Genomes. https://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/2018/hiv1dna.pdf.

307. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215. doi:10.1016/S0022-2836(05)80360-2

308. SMALT - Wellcome Sanger Institute. https://www.sanger.ac.uk/tool/smalt-0/.

309. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

310. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. 2019;20: 1–19.

311. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842.

312. Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40. doi:10.1093/nar/gks918

313. Neph S, Scott Kuehn M, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012;28: 1919.

314. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30: 3059–3066.

315. HIValign. https://www.hiv.lanl.gov/cgi-bin/VIRALIGN/viralign.cgi.

316. Stanford HIVDB. https://hivdb.stanford.edu/page/webservice/.

317. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One. 2010;5: e9490.

318. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4: vex042.

319. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2019;15: e1006650.

320. Louis du Plessis skylinetools: Utilities and distributions for (birth-death) skyline models. https://github.com/laduplessis/skylinetools.

321. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst Biol. 2018;67: 901–904.

322. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013;178: 1505–1512.

323. Hollingsworth TD, Anderson RM, Fraser C. HIV-1 Transmission, by Stage of Infection. J Infect Dis. 2008;198: 687–693.

324. Bobkov AF, Kazennova EV, Selimova LM, Khanina TA, Ryabov GS, Bobkova MR, et al. Temporal trends in the HIV-1 epidemic in Russia: predominance of subtype A. J Med Virol. 2004;74: 191–196.

325. The Global HIV Epidemics Among Men who Have Sex with Men. World Bank Publications; 2011.

326. Kazennova E, Laga V, Gromov K, Lebedeva N, Zhukova E, Pronin A, et al. Genetic Variants of HIV Type 1 in Men Who Have Sex with Men in Russia. AIDS Res Hum Retroviruses. 2017;33: 1061–1064.

327. Lebedev A, Lebedeva N, Moskaleychik F, Pronin A, Kazennova E, Bobkova M. Human Immunodeficiency Virus-1 Diversity in the Moscow Region, Russia: Phylodynamics of the Most Common Subtypes. Front Microbiol. 2019;10: 320.

328. Balabanova Y, Coker R, Atun RA, Drobniewski F. Stigma and HIV infection in Russia. AIDS Care. 2006;18. doi:10.1080/09540120600643641

329. Kelly J, Amirkhanian Y, Yakovlev A, Musatov V, Meylakhs A, Kuznetsova A, et al. Stigma reduces and social support increases engagement in medical care among persons with HIV infection in St. Petersburg, Russia. J Int AIDS Soc. 2014;17. doi:10.7448/IAS.17.4.19618

330. Katz IT, Ryu AE, Onuegbu AG, Psaros C, Weiser SD, Bangsberg DR, et al. Impact of HIV-related stigma on treatment adherence: systematic review and meta-synthesis. J Int AIDS Soc. 2013;16. doi:10.7448/IAS.16.3.18640

331. Tkatchenko-Schmidt E, Renton A, Gevorgyan R, Davydenko L, Atun R. Prevention of HIV/AIDS among injecting drug users in Russia: opportunities and barriers to scaling-up of

harm reduction programmes. Health Policy. 2008;85: 162–171.

332. Dukhovlinova E, Masharsky A, Vasileva A, Porrello A, Zhou S, Toussova O, et al. Characterization of the Transmitted Virus in an Ongoing HIV-1 Epidemic Driven by Injecting Drug Use. AIDS Res Hum Retroviruses. 2018;34. doi:10.1089/AID.2017.0313

333. Bobkov A, Garaev MM, Rzhaninova A, Kaleebu P, Pitman R, Weber JN, et al. Molecular epidemiology of HIV-1 in the former Soviet Union: analysis of env V3 sequences and their correlation with epidemiologic data. AIDS. 1994;8: 619–624.

334. In Russian. Витрина статистических данных. https://showdata.gks.ru/report/278930/.

335. Writing committee, affiliations, May MT, Justice AC, Birnie K, Ingle SM, et al. Injection drug use and Hepatitis C as risk factors for mortality in HIV-infected individuals: the Antiretroviral Therapy Cohort Collaboration. J Acquir Immune Defic Syndr. 2015;69: 348.

336. In Russian. Increased incidence of HIV-1 among IDUs in Oryol Oblast. http://aids-orel.ru/news/v-orlovskoy-oblasti-vpervye-za-mnogie-gody-zaregistrirovan-sushchestvennyy-rost-inficirovaniya-virusom.

337. Foley BT, Leitner T, Paraskevis D, Peeters M. Primate immunodeficiency virus classification and nomenclature: Review. Infect Genet Evol. 2016;46: 150–158.

338. Pineda-Peña AC, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, et al. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. Infect Genet Evol. 2013;19. doi:10.1016/j.meegid.2013.04.032

339. Drummond A, Forsberg R, Rodrigo AG. The Inference of Stepwise Changes in Substitution Rates Using Serial Sequence Samples. Mol Biol Evol. 2001;18: 1365–1371.

340. Continuum of HIV care - Monitoring implementation of the Dublin Declaration - 2018 progress report. https://www.ecdc.europa.eu/en/publications-data/continuum-hiv-care-monitoring-implementation-dublin-declaration-2018-progress. 4 Feb 2019.

341. Funk S, Gilad E, Watkins C, Jansen VAA. The spread of awareness and its impact on epidemic outbreaks. Proc Natl Acad Sci U S A. 2009;106: 6872.

342. Hué S, Clewley JP, Cane PA, Pillay D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. AIDS. 2004;18: 719.

343. Lemey P, Derdelinckx I, Rambaut A, Van Laethem K, Dumont S, Vermeulen S, et al. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. J Virol. 2005;79. doi:10.1128/JVI.79.18.11981-11989.2005

344. Gashnikova NM, Bogachev VV, Baryshev PB, Totmenin AV, Gashnikova MP, Kazachinskaya AG, et al. A rapid expansion of HIV-1 CRF63_02A1 among newly diagnosed HIV-infected individuals in the Tomsk Region, Russia. AIDS Res Hum Retroviruses. 2015;31. doi:10.1089/AID.2014.0375

345. Gashnikova NM, Zyryanova DP, Astakhova EM, Ivlev VV, Gashnikova MP, Moskaleva NV, et al. Predominance of CRF63_02A1 and multiple patterns of unique recombinant forms of CRF63_A1 among individuals with newly diagnosed HIV-1 infection in Kemerovo Oblast,

Russia. Arch Virol. 2017;162. doi:10.1007/s00705-016-3120-4

346. Rudometova NB, Shcherbakova NS, Shcherbakov DN, Mishenova EV, Delgado E, Ilyichev AA, et al. Genetic Diversity and Drug Resistance Mutations in Reverse Transcriptase and Protease Genes of HIV-1 Isolates from Southwestern Siberia. AIDS Res Hum Retroviruses. 2021;37: 716–723.

347. Shcherbakova NS, Shalamova LA, Delgado E, Fernández-García A, Vega Y, Karpenko LI, et al. Short communication: Molecular epidemiology, phylogeny, and phylodynamics of CRF63_02A1, a recently originated HIV-1 circulating recombinant form spreading in Siberia. AIDS Res Hum Retroviruses. 2014;30: 912–919.

348. Fischetti L, Opare-Sem O, Candotti D, Lee H, Allain JP. Higher viral load may explain the dominance of CRF02_AG in the molecular epidemiology of HIV in Ghana. AIDS. 2004;18. doi:10.1097/00002030-200405210-00017

349. Easterbrook PJ, Smith M, Mullen J, O'Shea S, Chrystie I, de Ruiter A, et al. Impact of HIV-1 viral subtype on disease progression and response to antiretroviral therapy. J Int AIDS Soc. 2010;13: 4.

350. In Russian. STUDY OF PROPERTIES OF SERUM NEUTRALIZING HIV-INFECTED PATIENTS WITH NON-PROGRESSORS DISEASES OF HIV ISOLATES OF DIFFERENT GENETIC VARIANTS. https://elibrary.ru/item.asp?id=23171667.

351. In Russian. НАЦИОНАЛЬНЫЕ РЕКОМЕНДАЦИИ ПО ДИСПАНСЕРНОМУ НАБЛЮДЕНИЮ И ЛЕЧЕНИЮ БОЛЬНЫХ ВИЧ-ИНФЕКЦИЕЙ 2015. http://nnoi.ru/uploads/files/HIV_2015.pdf.

352. In Russian. НАЦИОНАЛЬНЫЕ РЕКОМЕНДАЦИИ ПО ДИСПАНСЕРНОМУ НАБЛЮДЕНИЮ И ЛЕЧЕНИЮ БОЛЬНЫХ ВИЧ-ИНФЕКЦИЕЙ 2016. https://aidsyakutsk.ru/wp-content/uploads/2018/12/Protokoly-2016.pdf.

353. In Russian. АНАЛИЗ ЗАКУПОК АРВ-ПРЕПАРАТОВ В РОССИЙСКОЙ ФЕДЕРАЦИИ В 2019 ГОДУ. https://www.itpcru.org/wp-content/uploads/2020/04/itpcru-otchet-arv-preparaty-2019-28.04.20-final_prep.pdf.

354. In Russian. АНАЛИЗ ЗАКУПОК АРВ-ПРЕПАРАТОВ В РОССИЙСКОЙ ФЕДЕРАЦИИ В 2020 ГОДУ. https://www.itpcru.org/wp-content/uploads/2021/05/arvt-2020-final-28.05.21.pdf. [cited 11 Oct 2021]. Available: https://www.itpcru.org/wp-content/uploads/2021/05/arvt-2020-final-28.05.21.pdf

355. Coronavirus Update (Live): 11,965,661 Cases and 546,988 Deaths from COVID-19 Virus Pandemic - Worldometer. [cited 8 Jul 2020]. Available: https://www.worldometers.info/coronavirus/#countries

356. Coronavirus disease 2019 (COVID-19): Situation report, 42. (2020). [cited 8 Jul 2020]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200302-sitrep-42-covid-19.pdf?sfvrsn=224c1add_2

357. Coronavirus disease 2019 (COVID-19): Situation report, 47. 2020. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200307-sitrep-47-covi

d-19.pdf?sfvrsn=27c364a4_4

358. Coronavirus disease 2019 (COVID-19) : Situation report, 53. 2020. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200313-sitrep-53-covid-19.pdf?sfvrsn=adb3f72_2

359. The Moscow Times. Russia Closes Far East Border Over Coronavirus. The Moscow Times; 30 Jan 2020 [cited 8 Jul 2020]. Available: https://www.themoscowtimes.com/2020/01/30/russia-closes-far-east-border-over-coronavirus-a69100

360. Russia restricts air travel with China from February 1 due to coronavirus. In: TASS [Internet]. TASS; [cited 8 Jul 2020]. Available: https://tass.com/economy/1115335

361. In Russian. Принято решение о временном ограничении движения через пункты пропуска на отдельных участках государственной границы Российской Федерации с Китайской Народной Республикой. [cited 9 Jul 2020]. Available: http://government.ru/docs/38879

362. In Russian. Принят ряд решений в целях предупреждения проникновения на территорию России коронавирусной инфекции с территории Китайской Народной Республики. [cited 9 Jul 2020]. Available: http://government.ru/docs/38900/

363. In Russian. Принято решение о временном ограничении въезда граждан иностранных государств с территории Китайской Народной Республики в воздушных пунктах пропуска через государственную границу Российской Федерации. [cited 9 Jul 2020]. Available: http://government.ru/docs/38912/

364. In Russian. Принято решение о временной приостановке пропуска через государственную границу Российской Федерации граждан Китайской Народной Республики, въезжающих для осуществления трудовой деятельности, в частных, учебных и туристических целях. [cited 9 Jul 2020]. Available: http://government.ru/docs/38996/

365. In Russian. Принято решение о временном ограничении въезда иностранных граждан с территории Республики Корея в воздушных пунктах пропуска через государственную границу Российской Федерации. [cited 9 Jul 2020]. Available: http://government.ru/docs/39041/

366. In Russian. Временно ограничен въезд иностранных граждан с территории Исламской Республики Иран в воздушных пунктах пропуска через государственную границу Российской Федерации. [cited 9 Jul 2020]. Available: http://government.ru/docs/39043

367. Why are there so few reported COVID-19 cases in Russia? — Meduza. In: Meduza [Internet]. [cited 9 Jul 2020]. Available: https://meduza.io/en/feature/2020/03/06/why-are-there-so-few-reported-covid-19-cases-in-russia

368. First two persons infected with coronavirus identified in Russia. In: TASS [Internet]. TASS; [cited 9 Jul 2020]. Available: https://tass.com/society/1115101

369. One imported coronavirus case confirmed in Russia. In: TASS [Internet]. TASS; [cited 8 Jul 2020]. Available: https://tass.com/society/1125627

370. Covid-19: Global summary. In: Covid-19 [Internet]. 7 Jul 2020 [cited 9 Jul 2020]. Available: https://epiforecasts.io/covid/posts/global/

371. Coronavirus disease 2019 (COVID-19) : Situation report, 61. (2020). [cited 8 Jul 2020]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200321-sitrep-61-covid-19.pdf?sfvrsn=ce5ca11c_2

372. In Russian. Объемы перевозок через аэропорты России. [cited 10 Jul 2020]. Available: https://favt.ru/dejatelnost-ajeroporty-i-ajerodromy-osnovnie-proizvodstvennie-pokazateli-aeroportov-obyom-perevoz/

373. In Russian. Статистика аэропорта Домодедово. [cited 10 Jul 2020]. Available: https://business.dme.ru/company/finance/operacionnye-rezul_taty-/

374. In Russian. Новости аэропорта Внуково. [cited 10 Jul 2020]. Available: http://corp.vnukovo.ru/press/news/

375. In Russian. Показатели аэропорта Пулково. [cited 9 Jul 2020]. Available: https://pulkovoairport.ru/about/performance/

376. In Russian. № 12-УМ от 05.03.2020 «О введении режима повышенной готовности». [cited 9 Jul 2020]. Available: https://www.mos.ru/authority/documents/doc/43503220/

377. In Russian. Коронавирус. Дополнительные меры 14.03.2020. [cited 9 Jul 2020]. Available: https://www.sobyanin.ru/koronavirus-dopolnitelnye-mery-14-03-2020

378. In Russian. Коронавирус. Запрет проведения массовых мероприятий и другие ограничительные меры 16.03.2020. [cited 9 Jul 2020]. Available: https://www.sobyanin.ru/koronavirus-ogranichitelnye-mery-16-03-2020

379. In Russian. Постановление правительства Санкт-Петербурга от 13 марта 2020 года № 121 “О мерах по противодействию распространению в Санкт-Петербурге новой коронавирусной инфекции (COVID-19).” Российская газета. [cited 9 Jul 2020]. Available: https://rg.ru/2020/03/13/spb-post121-reg-dok.html

380. In Russian. Принято решение о временной приостановке пропуска через государственную границу Российской Федерации иностранных граждан и лиц без гражданства, прибывающих с территории Итальянской Республики для обучения и трудовой деятельности, а также в частных, туристических и транзитных целях. [cited 9 Jul 2020]. Available: http://government.ru/docs/39140/

381. In Russian. Принято решение о временном ограничении въезда в Российскую Федерацию иностранных граждан и лиц без гражданства, в том числе прибывающих с территории Республики Беларусь, а также граждан Республики Беларусь. [cited 9 Jul 2020]. Available: http://government.ru/docs/39179/

382. In Russian. Объемы перевозок через аэропорты МАУ за январь-апрель 2020 года. [cited 10 Jul 2020]. Available: http://www.aex.ru/docs/2/2020/5/29/3074/

383. In Russian. Пассажиропоток аэропорта “Шереметьево” в мае возрос на 39%. [cited 10 Jul 2020]. Available: http://www.aex.ru/news/2020/6/23/213922/

384. Coronavirus disease 2019 (COVID-19) : Situation report, 51. [cited 10 Jul 2020]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf?sfvrsn=1ba62e57_10

385. Coronavirus disease 2019 (COVID-19) : Situation report, 94. [cited 10 Jul 2020]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200423-sitrep-94-covid-19.pdf?sfvrsn=b8304bf0_4)

386. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DKW, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance. 2020;25: 2000045.

387. ARTIC nanopore protocol for nCoV2019 novel coronavirus. [cited 10 Jul 2020]. Available: https://github.com/joshquick/artic-ncov2019

388. Oxford Nanopore Technologies. [cited 9 Jul 2020]. Available: https://github.com/nanoporetech

389. Porechop. [cited 8 Jul 2020]. Available: https://github.com/rrwick/Porechop

390. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100.

391. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27: 2987–2993.

392. Jvarkit : Java utilities for Bioinformatics. [cited 8 Jul 2020]. Available: http://lindenb.github.io/jvarkit/

393. Robinson P, Jtel TZ. Integrative genomics viewer (IGV): Visualizing alignments and variants. Computational Exome and Genome Analysis. 2017. pp. 233–245. doi:10.1201/9781315154770-17

394. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780.

395. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32: 268–274.

396. Phylogenetic tree of SARS-CoV-2 sequences used in the analysis, provided in newick format. 5 Dec 2020 [cited 17 Dec 2020]. Available: 10.6084/m9.figshare.13337432.v1

397. Pybus O, Rambaut A, du Plessis L, Zarebski A, Kraemer M, Jayna Raghwani, et al. Preliminary analysis of SARS-CoV-2 importation & establishment of UK transmission lineages. In: Virological [Internet]. 9 Jun 2020 [cited 8 Jul 2020]. Available: https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507/

398. GADM. [cited 8 Jul 2020]. Available: https://gadm.org/download_country_v3.html

399. In Russian. Real-time data on COVID-19. [cited 8 Jul 2020]. Available: https://xn--80aesfpebagmfblc0a.xn--p1ai/information/

400. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 2016;33: 1635–1638.

401. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47: W256–W259.

402. FigTree. [cited 8 Jul 2020]. Available: http://tree.bio.ed.ac.uk/software/figtree/

403. GISAID - Initiative. [cited 8 Jul 2020]. Available: https://www.gisaid.org/

404. Rambaut A, Holmes EC, Hill V, O'Toole Á, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. bioRxiv. 2020. p. 2020.04.17.046086. doi:10.1101/2020.04.17.046086

405. GISAID - Next hCoV-19 App. [cited 8 Jul 2020]. Available: https://www.gisaid.org/epiflu-applications/next-hcov-19-app/

406. In Russian. The head of the gas company in Yakutia was hospitalized for suspected viral infection. In: RBC [Internet]. [cited 8 Jul 2020]. Available: https://www.rbc.ru/society/18/03/2020/5e71fc479a7947187644b347

407. Two residents of Chechnya infected with coronavirus during Hajj. In: Caucasian Knot [Internet]. [cited 8 Jul 2020]. Available: https://www.eng.kavkaz-uzel.eu/articles/50388/

408. Most of Moscow's New Coronavirus Patients Younger Than 40. The Moscow Times; 30 Mar 2020 [cited 8 Jul 2020]. Available: https://www.themoscowtimes.com/2020/03/30/most-of-moscows-new-coronavirus-patients-younger-than-40-a69797

409. In Russian. Confined Space War on Coronavirus. The history of the clinic closed for quarantine in St. Petersburg. In: BBC News Russia [Internet]. BBC News Русская служба; 28 May 2020 [cited 9 Jul 2020]. Available: https://www.bbc.com/russian/features-52813191

410. Russian doctors, nurses face more risks as virus cases grow. In: AP NEWS [Internet]. Associated Press; 28 Apr 2020 [cited 9 Jul 2020]. Available: https://apnews.com/b4950726aea5b4ec33a0f4d8fa76cb40

411. Rambaut A. Phylodynamic Analysis | 176 genomes | 6 Mar 2020. In: Virological [Internet]. 29 Jan 2020 [cited 10 Jul 2020]. Available: https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356

412. da Silva Candido D, Claro IM, de Jesus JG, de Souza WM, Moreira FRR, Dellicour S, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. medRxiv. 2020; 2020.06.11.20128249.

413. Lemieux J. Introduction and spread of SARS-CoV-2 in the greater Boston area. In: Virological [Internet]. 4 Jun 2020 [cited 10 Jul 2020]. Available: https://virological.org/t/introduction-and-spread-of-sars-cov-2-in-the-greater-boston-area/503

414. KEMRI-CGMRC, Kilifi, KEMRI-CVR, Nairobi, The National Public Health Laboratory-National Influenza Centre (NPHL-NIC). Introduction and local transmission of SARS-CoV-2 cases in Kenya. In: Virological [Internet]. 2 Jun 2020 [cited 10 Jul 2020]. Available: https://virological.org/t/introduction-and-local-transmission-of-sars-cov-2-cases-in-kenya/497

415. Miller D, Martin MA, Harel N, Kustin T, Tirosh O, Meir M, et al. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. medRxiv. 2020; 2020.05.21.20104521.

416. Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. Cell. 2020;181: 997–1003.e9.

417. Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K, et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature. 2017;546: 401–405.

418. Kraemer MUG, Cummings DAT, Funk S, Reiner RC, Jr, Faria NR, et al. Reconstruction and prediction of viral disease epidemics. Epidemiol Infect. 2019;147. doi:10.1017/S0950268818002881

419. Villabona-Arenas CJ, Hanage WP, Tully DC. Phylogenetic interpretation during outbreaks requires caution. Nature Microbiology. 2020;5: 876–877.

420. In Russian. Entry of foreign citizens to the Russian Federation. [cited 8 Oct 2020]. Available: https://fedstat.ru/indicator/38479

421. In Russian. Departure of Russian citizens. [cited 8 Oct 2020]. Available: https://fedstat.ru/indicator/38480

422. Olsen SJ, Chen MY, Liu YL, Witschi M, Ardoin A, Calba C, et al. Early Introduction of Severe Acute Respiratory Syndrome Coronavirus 2 into Europe. Emerg Infect Dis. 2020;26: 1567–1570.

423. Böhmer MM, Buchholz U, Corman VM, Hoch M, Katz K, Marosevic DV, et al. Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series. The Lancet Infectious Diseases. 2020. doi:10.1016/s1473-3099(20)30314-5

424. Evidence for Limited Early Spread of COVID-19 Within the United States, January–February 2020. MMWR Morb Mortal Wkly Rep. 2020;69. doi:10.15585/mmwr.mm6922e1

425. Stadler T. Phylodynamic Analyses of outbreaks in China, Italy, Washington State (USA), and the Diamond Princess. In: Virological [Internet]. 13 Mar 2020 [cited 8 Jul 2020]. Available: https://virological.org/t/phylodynamic-analyses-of-outbreaks-in-china-italy-washington-state-usa-and-the-diamond-princess/439

426. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance. 2020;25: 2000180.

427. Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. Int J Infect Dis. 2020;93: 201–204.

428. Sekizuka T, Itokawa K, Kageyama T, Saito S, Takayama I, Asanuma H, et al. Haplotype networks of SARS-CoV-2 infections in the Diamond Princess cruise ship outbreak. medRxiv. 2020; 2020.03.23.20041970.

429. Deng X, Gu W, Federman S, du Plessis L, Pybus OG, Faria N, et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science. 2020 [cited 10 Jul 2020]. doi:10.1126/science.abb9263

430. Report into a nosocomial outbreak of coronavirus disease 2019 (COVID−19) at Netcare St. Augustine's Hospital. [cited 10 Jul 2020]. Available:

https://www.krisp.org.za/manuscripts/StAugustinesHospitalOutbreakInvestigation_FinalReport_15may2020_comp.pdf

431. Juan Ángel Patiño-Galindo FG-C. The substitution rate of HIV-1 subtypes: a genomic approach. Virus Evolution. 2017;3. doi:10.1093/ve/vex029

432. Gytis Dudas et al. MERS-CoV spillover at the camel-human interface. Elife. 2018;7:e31257

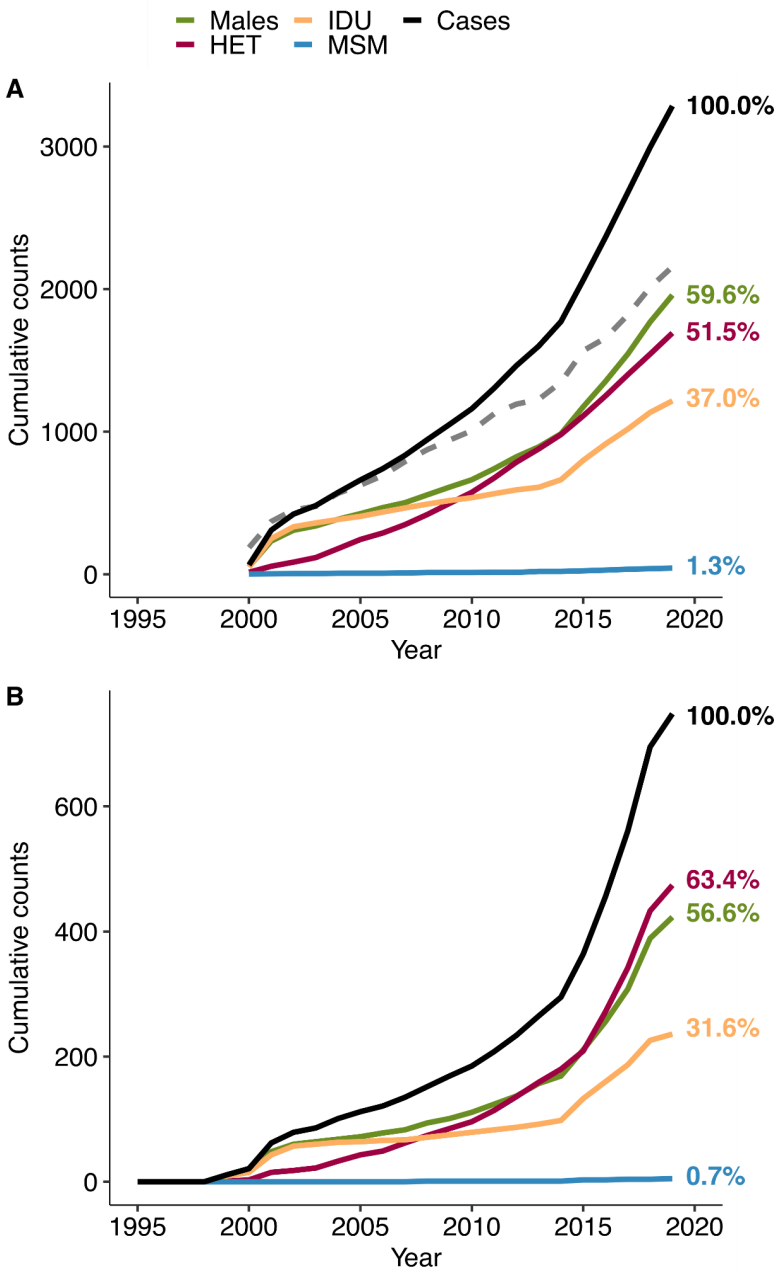
433. N. R. Faria et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. Science. 2018;6405:894-899

434. Alina Matsvay, Galya V. Klink et al. Genomic epidemiology of SARS-CoV-2 in Russia reveals recurring cross-border transmission throughout 2020. medRxiv 2021.03.31.21254115

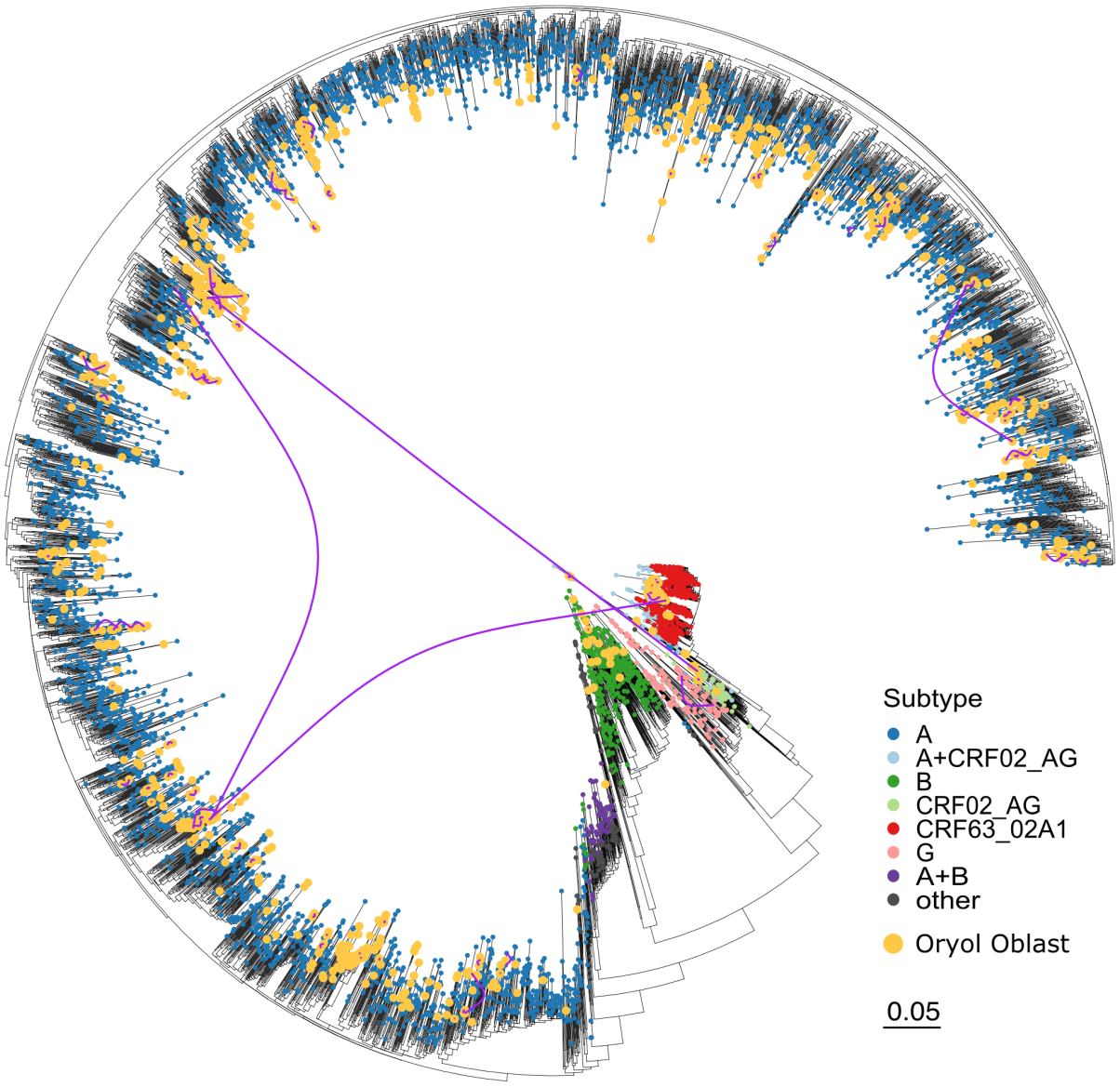
435. Galya V. Klink et al. Spread of endemic SARS-CoV-2 lineages in Russia. medRxiv 2021.05.25.21257695

APPENDIX A

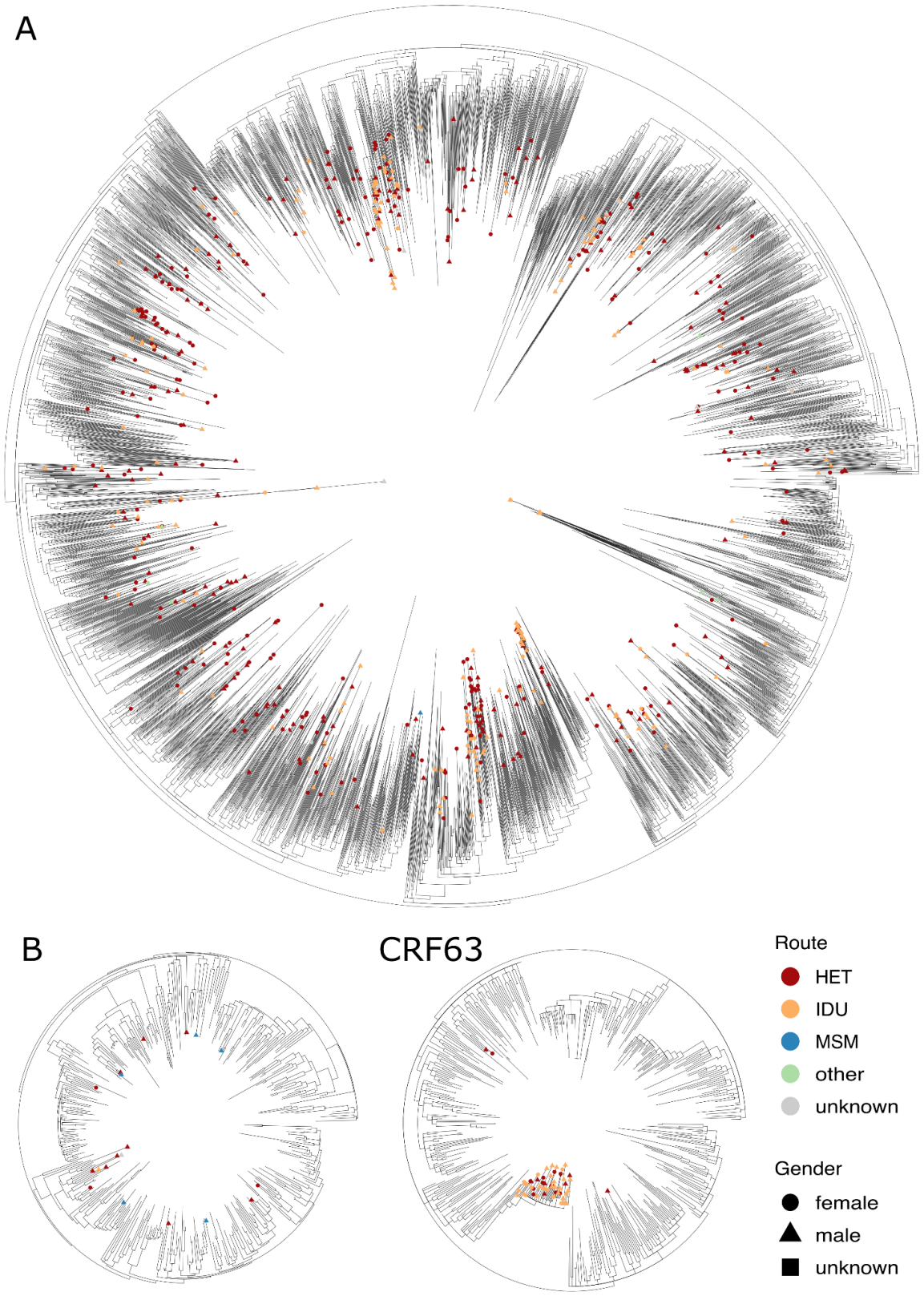
Supplementary Figures A



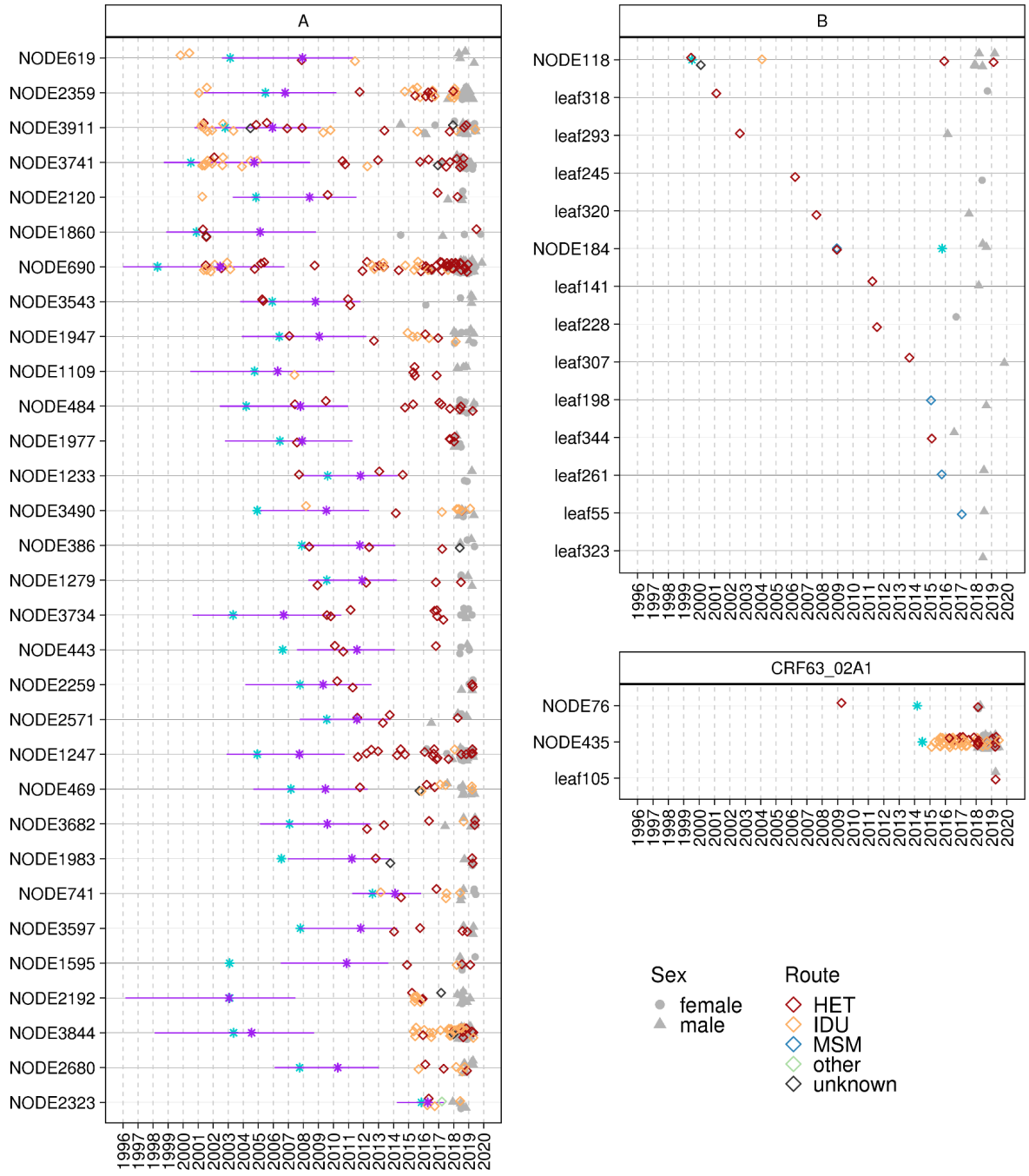
Supplementary Figure A-1. The statistics inferred from the sequenced dataset agrees with the official statistics on the Oryol Oblast. The plots show the cumulative number of different categories according to (A) the statistics on new cases provided by the Oryol Oblast AIDS center and (B) sequenced samples analyzed in this study (by the year of diagnosis). The dashed gray line on A shows the total number of HIV-1 positive people registered in Oryol Oblast every year and differs from the black line due to deaths and migrations.



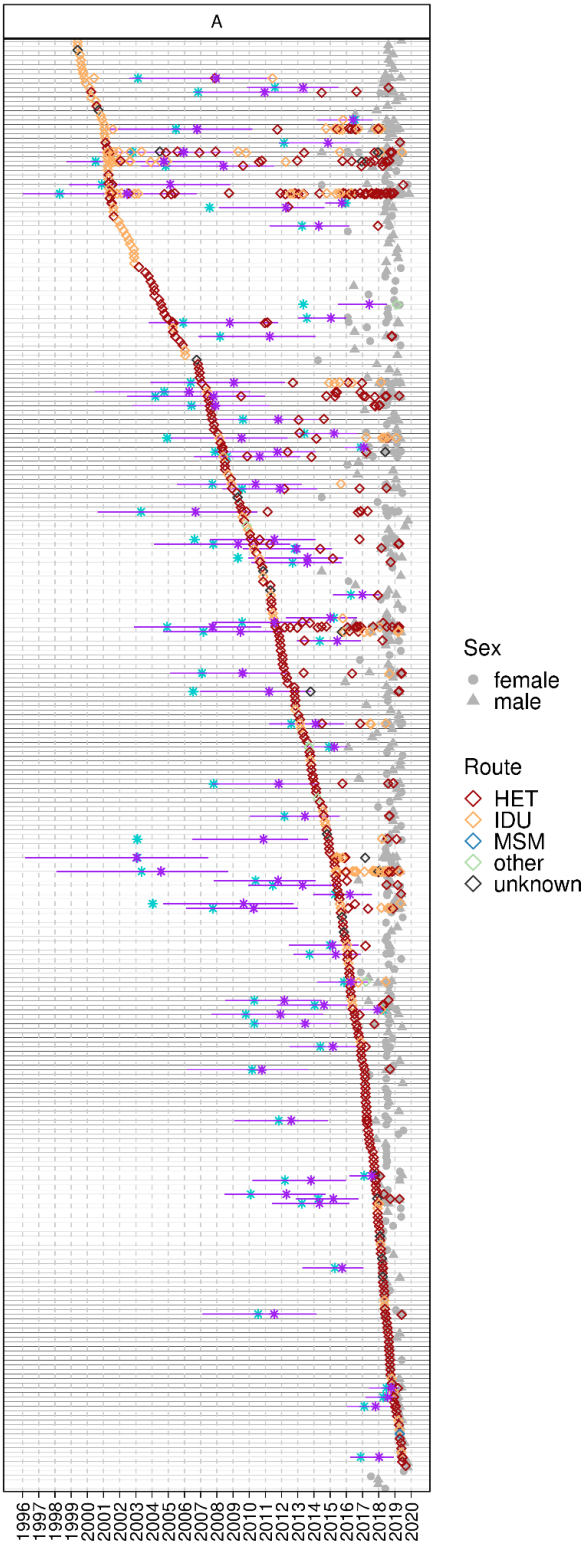
Supplementary Figure A-2. The complete phylogenetic tree of the combined Russian dataset, including multiple samples from repeatedly sequenced patients (joined with purple arcs).



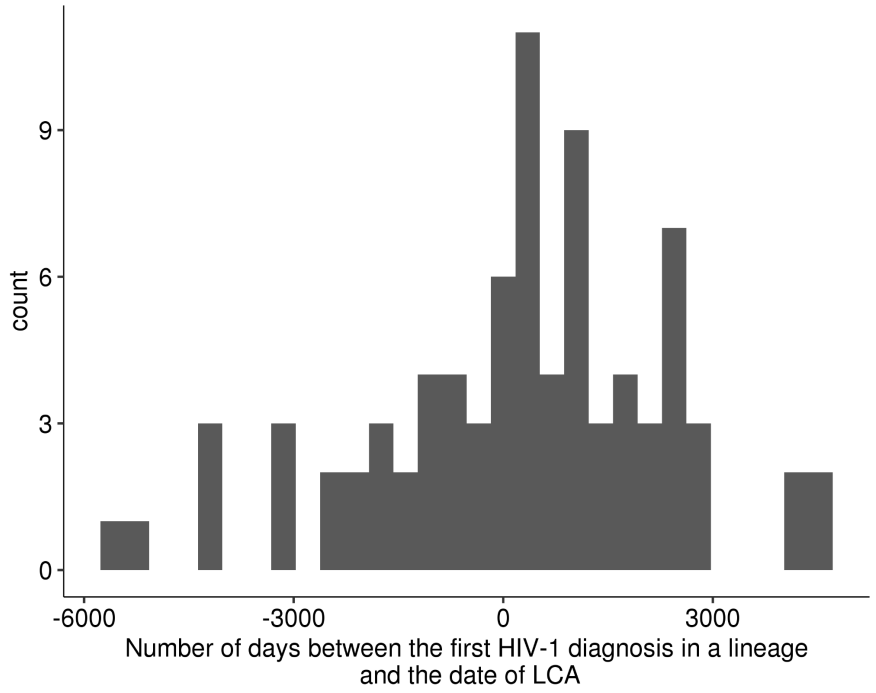
Supplementary Figure A-3. Phylogenetic trees reconstructed for subtypes A and B and CRF63. When more than one sample per patient was available, only the earliest sample was used.



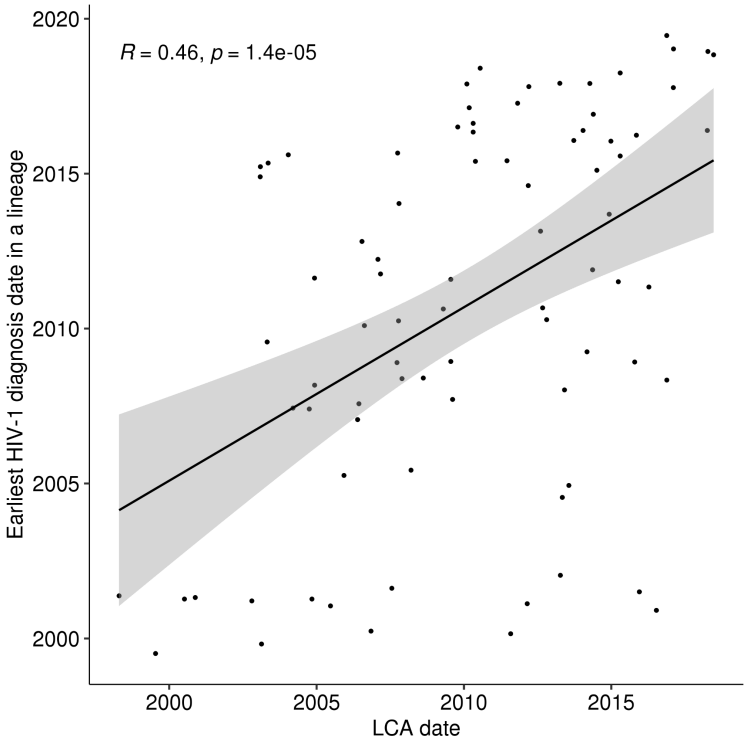
Supplementary Figure A-4. Introductions sorted by the earliest diagnosis. Legend as in Fig. 3.5.



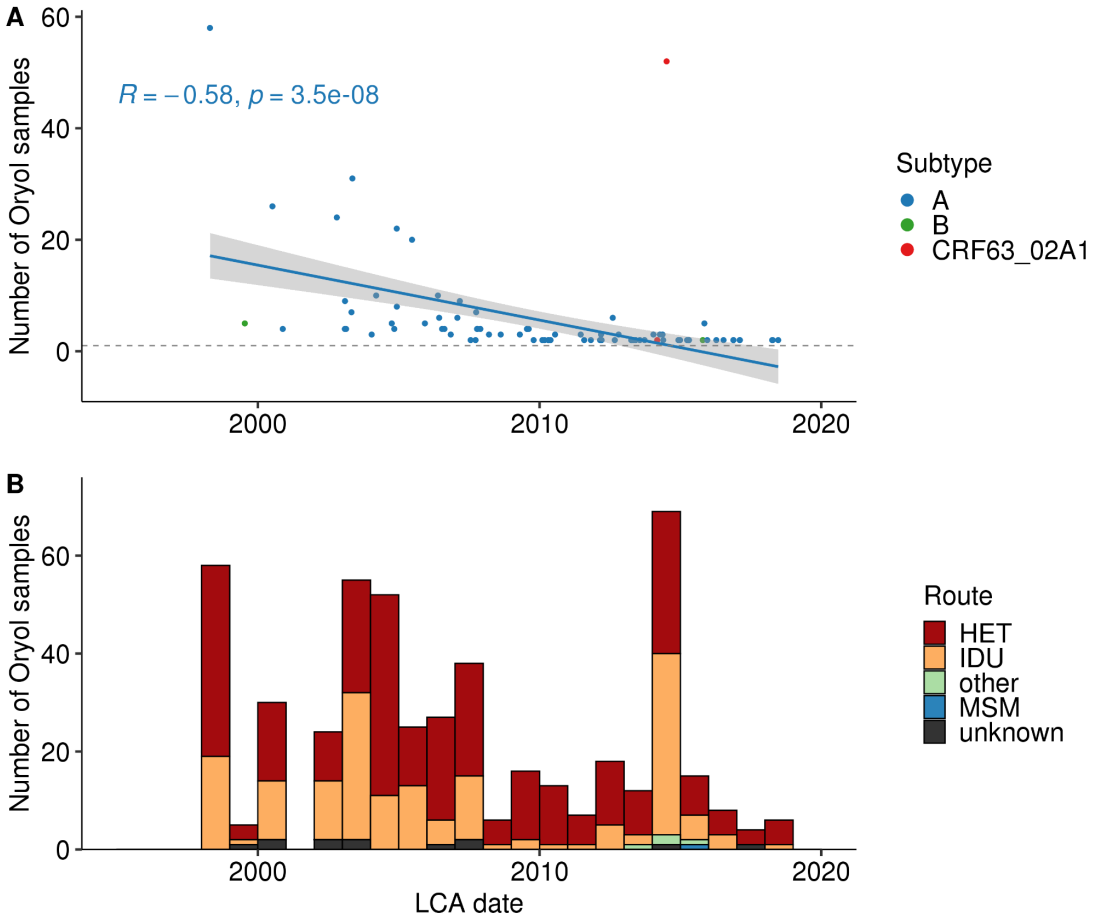
Supplementary Figure A-5. All subtype A introductions sorted by the earliest diagnosis. Legend as in Fig. 3.5.



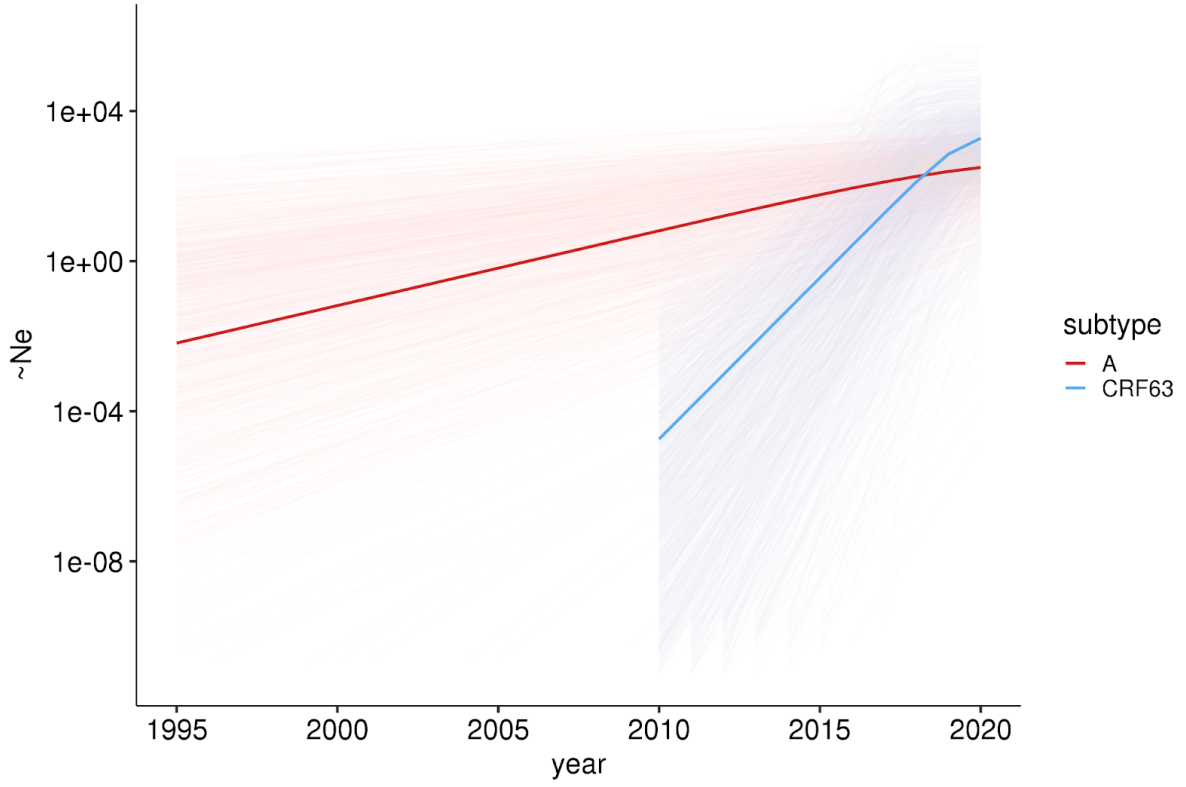
Supplementary Figure A-6. The number of days between the earliest HIV-1 diagnosis in a lineage and the inferred date of its LCA.



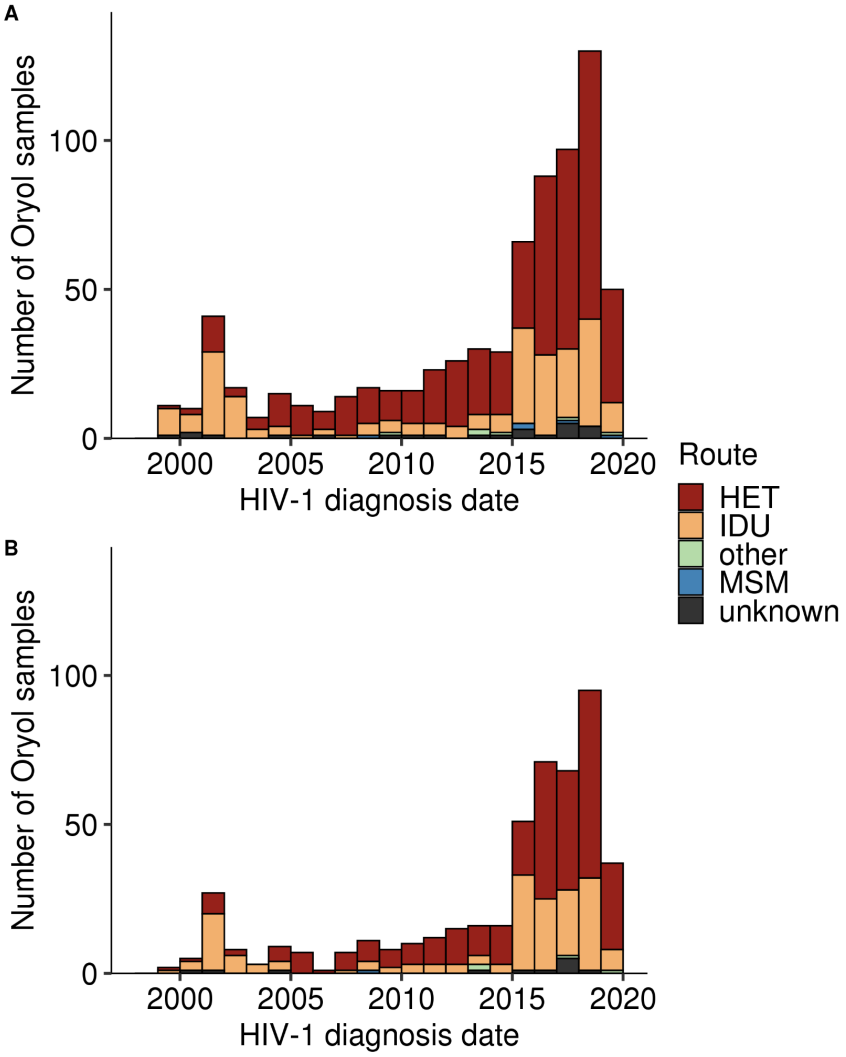
Supplementary Figure A-7. The inferred LCA date correlates with the date of the earliest diagnosis in a lineage.



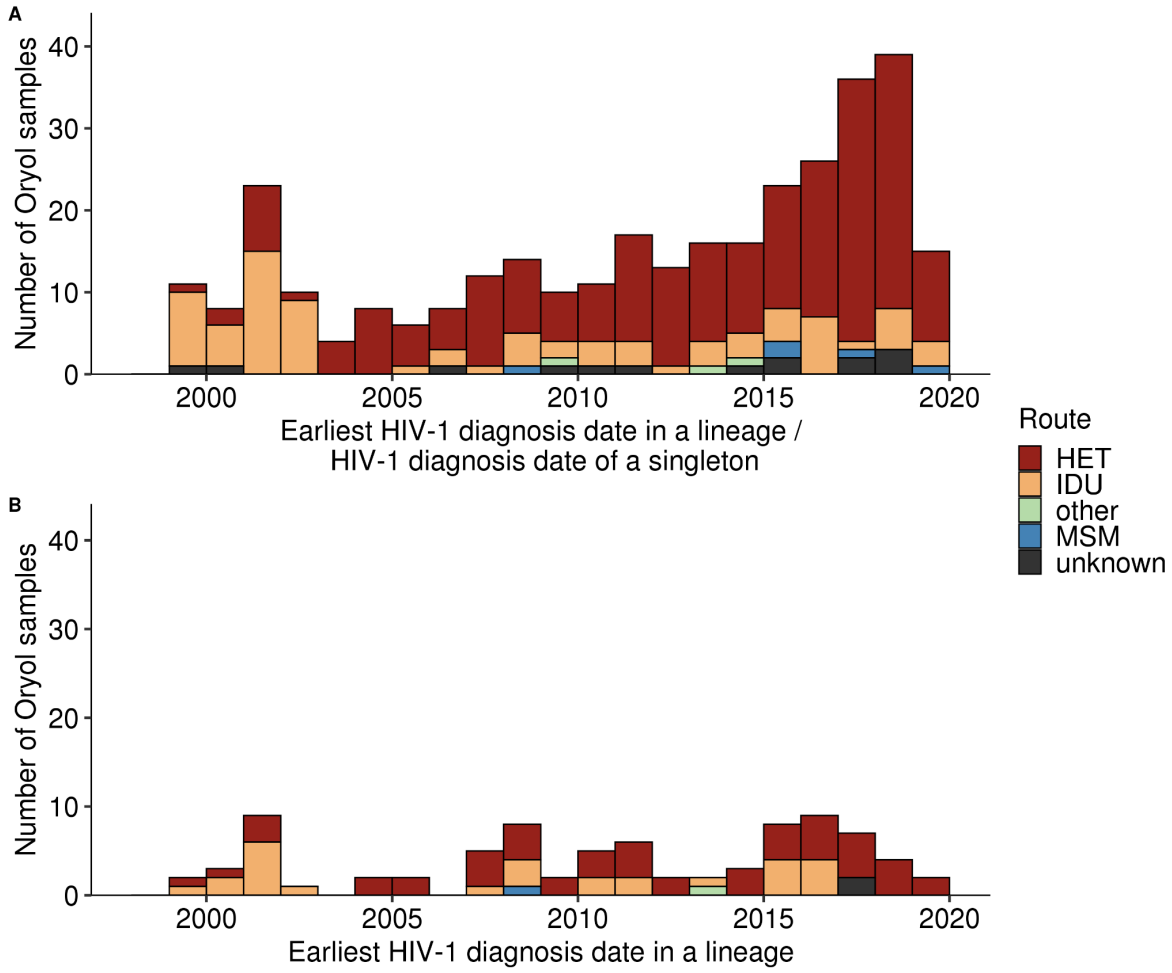
Supplementary Figure A-8. A. Lineages with earlier LCAs tend to carry more Oryol samples. The regression line is shown for subtype A; the regression analysis for all samples also shows a negative dependency (the slope p-value is 1e-04 for all lineages and 3.5e-08 for subtype A lineages). B. The distribution of samples by routes and the LCA date of their lineages.



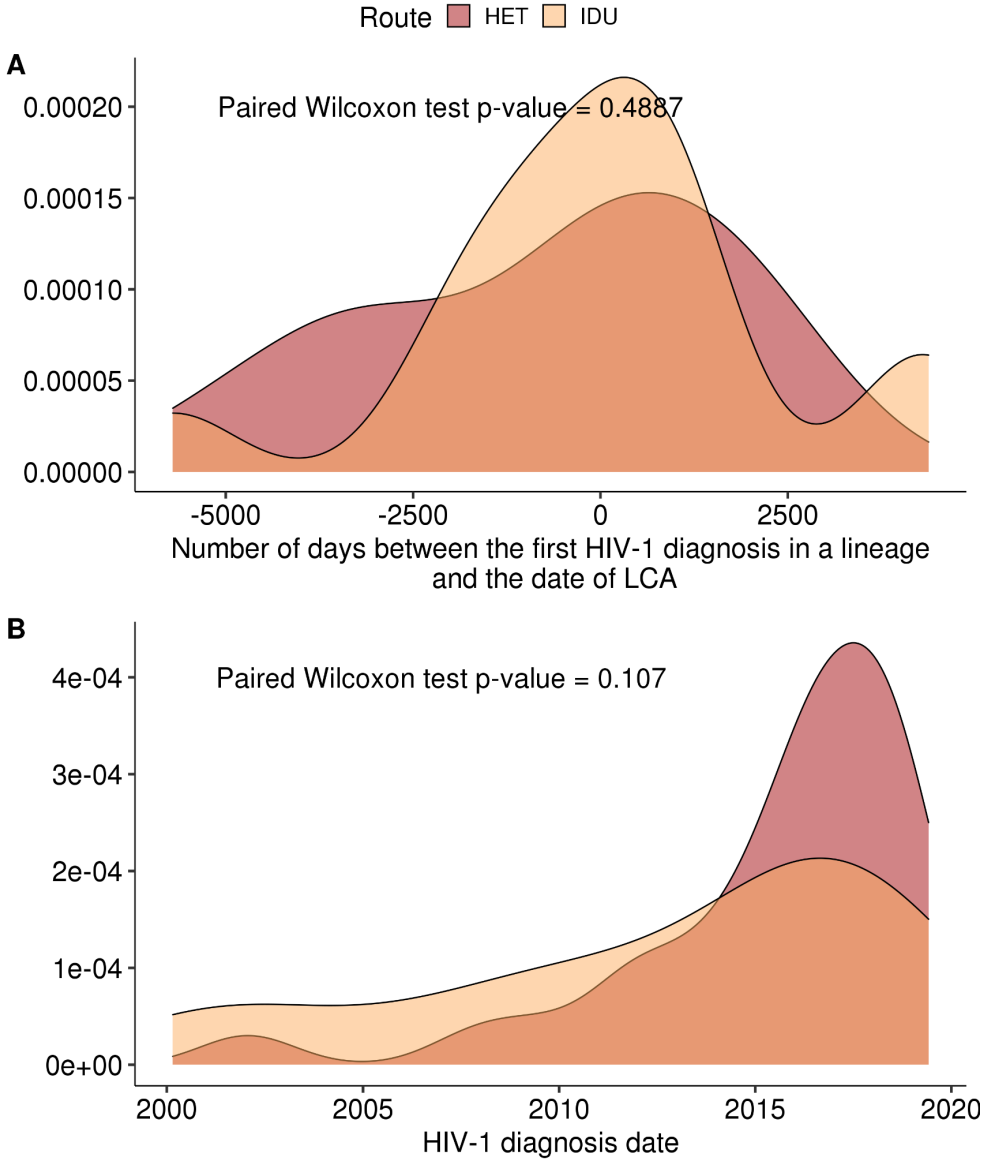
Supplementary Figure A-9. The logistic growth dynamics inferred in BEAST for the two largest transmission lineages. Shading corresponds to 1,000 randomly sampled trajectories during the MCMC run.



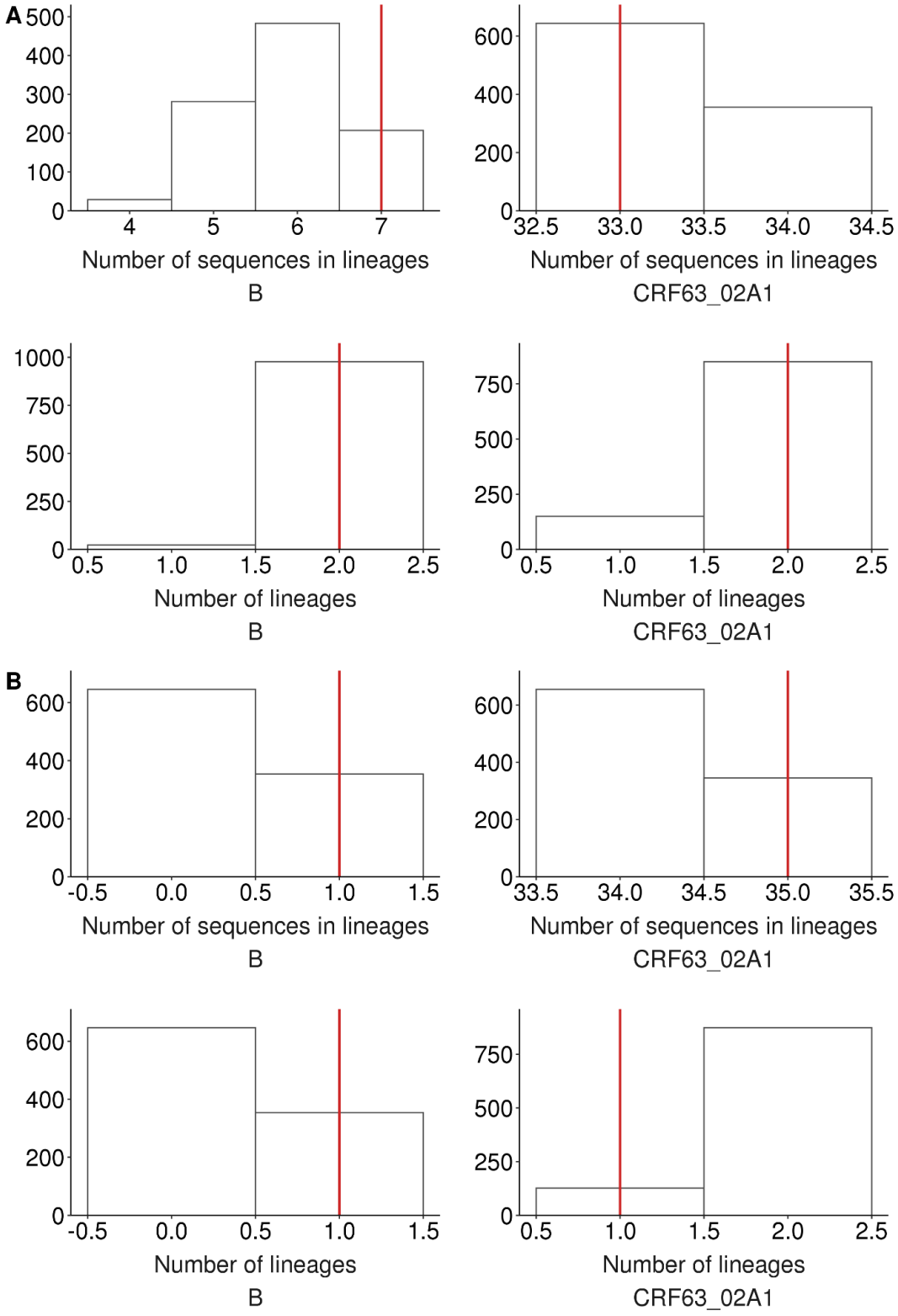
Supplementary Figure A-10. The distribution of samples by route and diagnosis date for all samples (A) and samples belonging to transmission lineages (B).



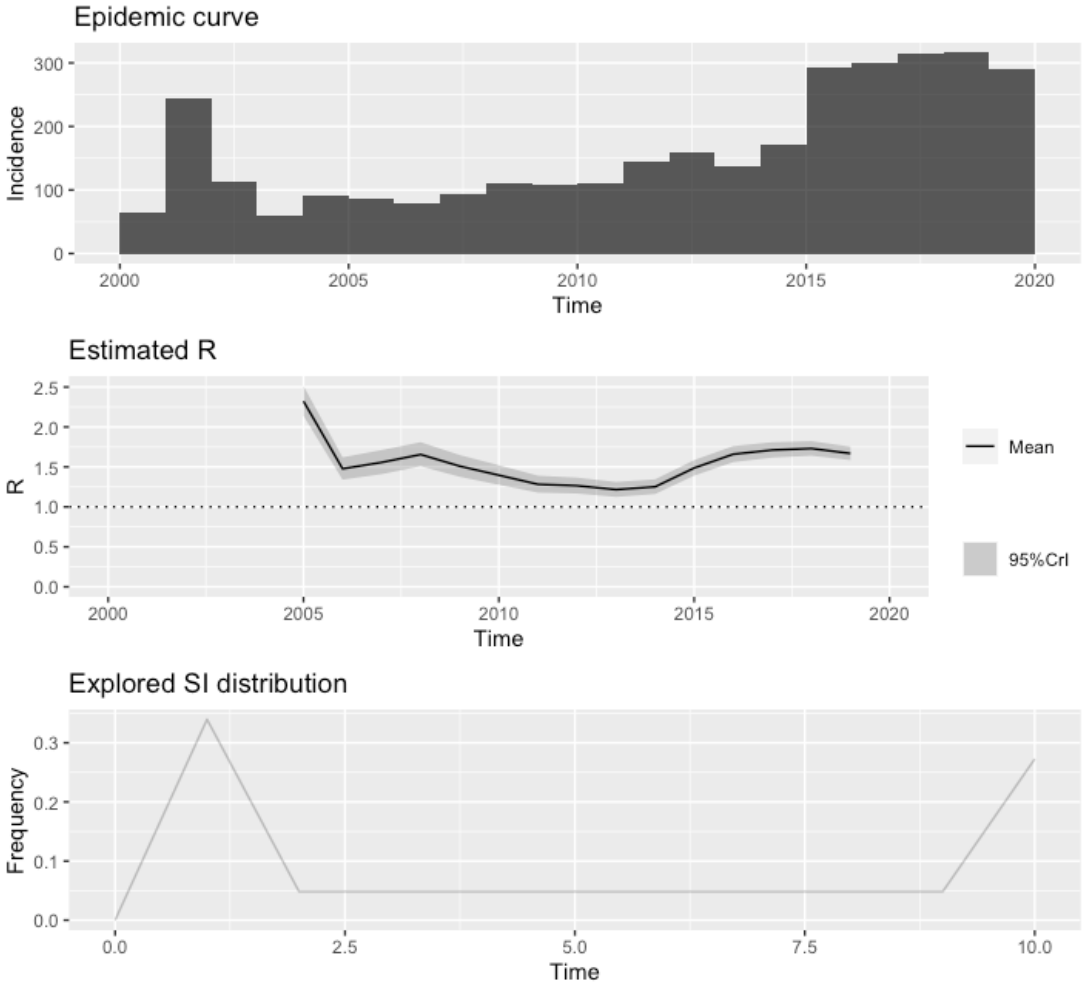
Supplementary Figure A-11. The distribution of the earliest sample per import by the route and the first diagnosis date for all imports (A) and transmission lineages (B).



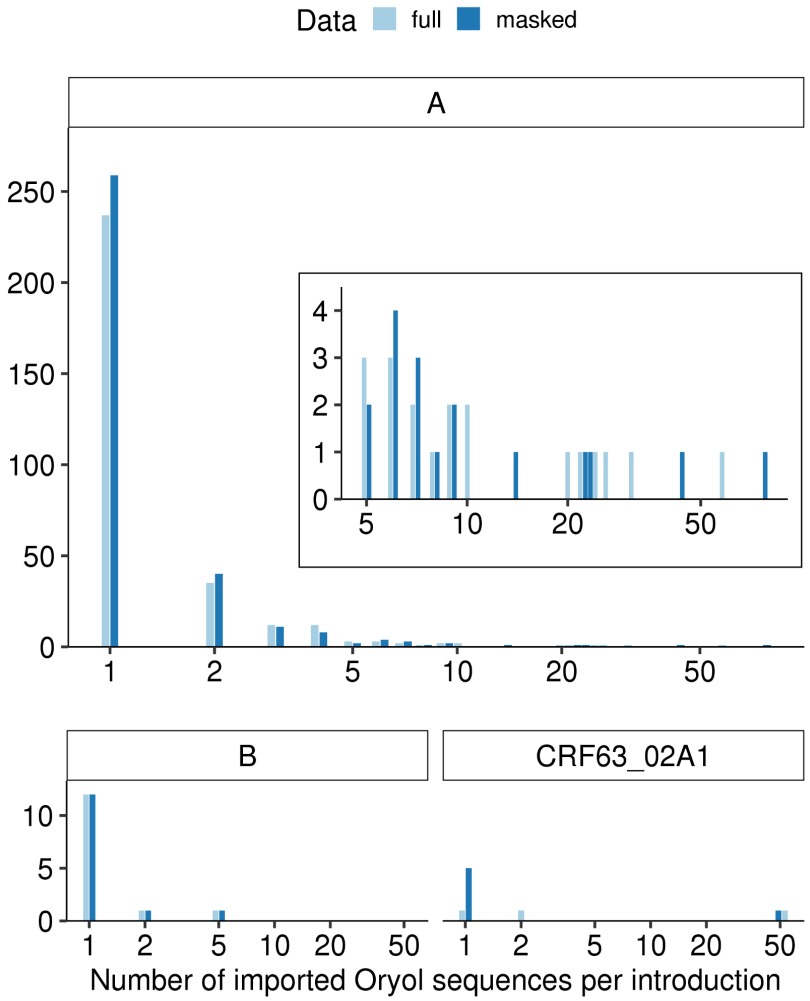
Supplementary Figure A-12. IDUs are not more likely to found a transmission lineage. A. The number of days between the first HIV-1 diagnosis in a lineage and the date of LCA in a matched-pair dataset sorted by the earliest diagnosis in a lineage (see 3.2.11). B. HIV-1 diagnosis date of HETs and IDUs in a matched-pair dataset sorted by the median date of diagnosis in a lineage.



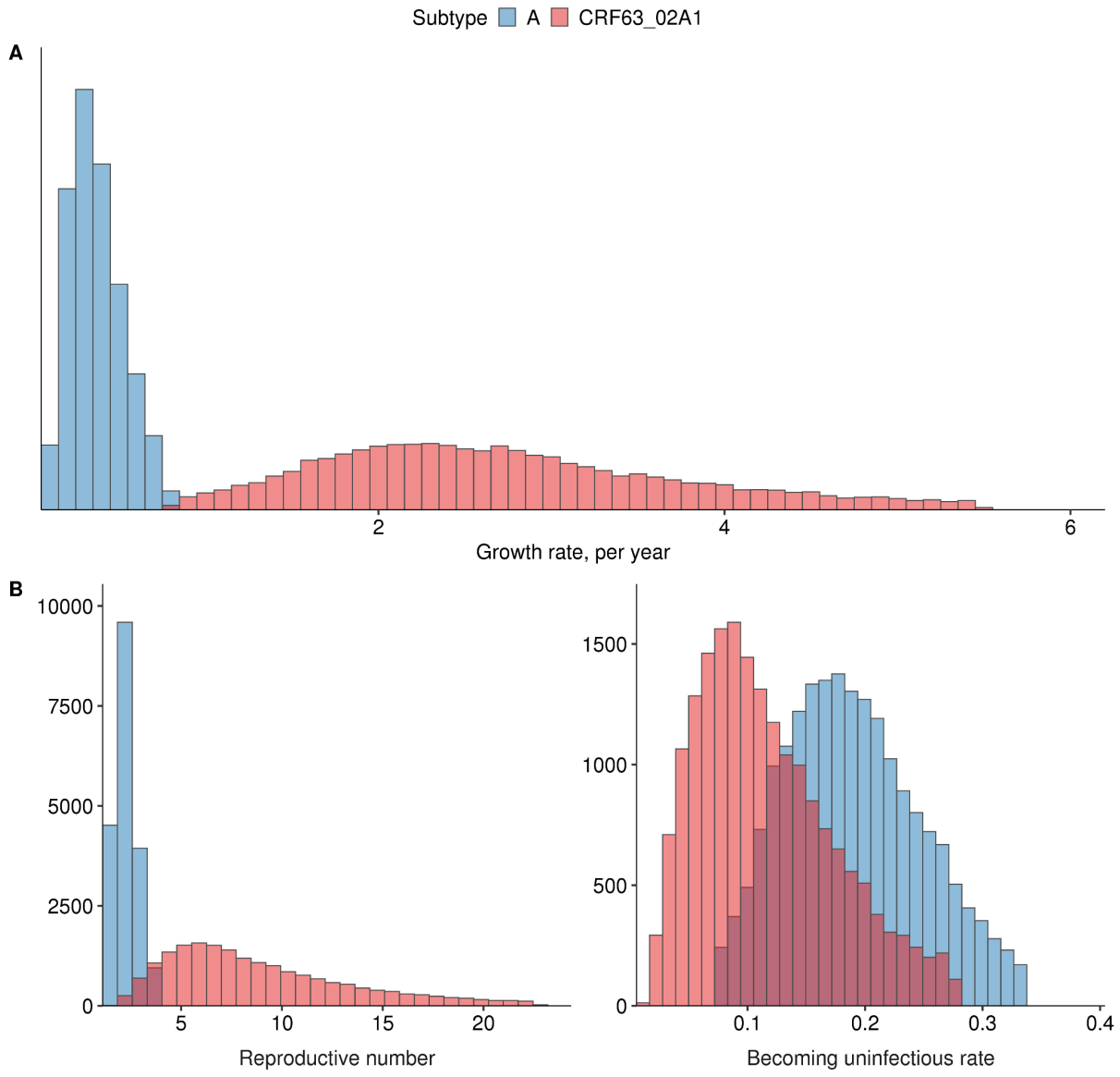
Supplementary Figure A-13. The expected distribution of the number of clustered sequences and the number of clusters carrying (A) males or (B) IDUs within subtypes B (left) and CRF63 (right).



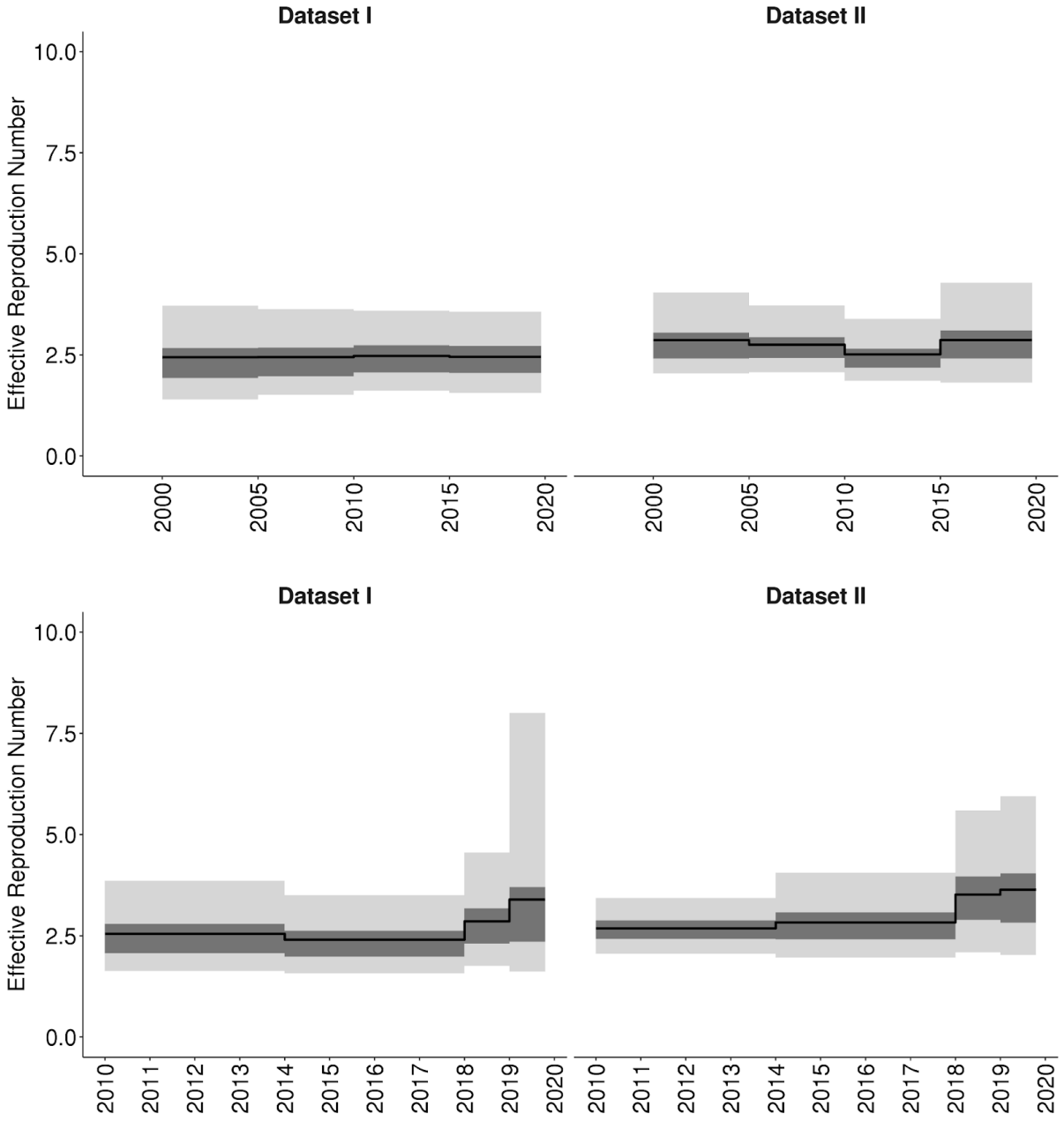
Supplementary Figure A-14. The reproduction number dynamics inferred by EpiEstim based on case counts.



Supplementary Figure A-15. Dependence of distribution of samples across transmission lineages on DRM masking.



Supplementary Figure A-16. Comparison of two clades with DRM masked. A. Logistic growth, B. BDSKY



Supplementary Figure A-17. Inferring the time-varying dynamics of R_e for subtype A introductions. Upper panel, R_e was allowed to change in 2005, 2010, and 2015. Lower panel, R_e was allowed to change at the same time as sampling proportion.

Supplementary Tables A

Supplementary Table A-1. Inferring prior on sampling proportion. Estimated fractions are used as mean values of prior distributions on sampling proportion.

Time period	Number of new cases	Number of recently diagnosed sequences	Fraction
Before 2014	1600	0	0.0001 (pseudocounted)
2014-2017	1078	8	0.007
2018	316	149	0.473
2019	290	71	0.246

Supplementary Table A-2. Priors used for multi-tree birth-death analysis. Reproduction number and sampling proportion are implemented in log-scale in this custom BEAST2 release.

Parameter	Prior
ucld.mean	Normal(X, 0.001); X = 0.00150 / 0.00080 for A / CRF63
Reproduction number	Normal(0,1)
Rate of becoming uninfectious	Lognormal(-1.5,1)
Sampling proportion before 2014	Normal(-10.00,0.1)
Sampling proportion in 2014-2017	Normal(-4.90,0.1)
Sampling proportion in 2018	Normal(-0.75,0.05)
Sampling proportion in 2019	Normal(-1.40,0.05)

Supplementary Table A-3. Priors used for BDSKY analysis of two clades.

Parameter	Prior
ucld.mean	Normal(Slope, 0.001); Slope = 0.00150 / 0.00164 for A / CRF63
Reproduction number	Lognormal(0,1.25)
Rate of becoming uninfectious	Lognormal(-1.5,1)

Sampling proportion before 2014	Normal(0.0001,0.001)
Sampling proportion in 2014-2017	Normal(0.007,0.001)
Sampling proportion in 2018	Normal(0.743,0.01)
Sampling proportion in 2019	Normal(0.246,0.01)

Supplementary Table A-4. Priors used for logistic growth analysis.

Parameter	Prior
ucld.mean	Normal(Slope, 0.001); Slope = 0.00150 / 0.00164 for A / CRF63
logistic.popSize	Lognormal(1,5)
logistic.growthRate	Exponential(1)
logistic.t50 (the number of years from the most recent sample till the inflection point, when the epidemic size is popSize/2)	Uniform(0,30) / Uniform(0,10) for A / CRF63

Supplementary Table A-5. Multitree birth-death estimates produced for subtype A. Median (95% HPD) values are provided.

	Full alignment	Alignment with masked DRM sites
Reproduction number	2.64 [2.05-3.39]	2.64 [2.07-3.46]
The rate of becoming uninfectious	0.20 [0.15-0.26]	0.20 [0.15-0.26]
ucld.mean	2.02e-03 [1.43-2.66]	2.06e-03 [1.50-2.70]

APPENDIX B

Supplementary Note B

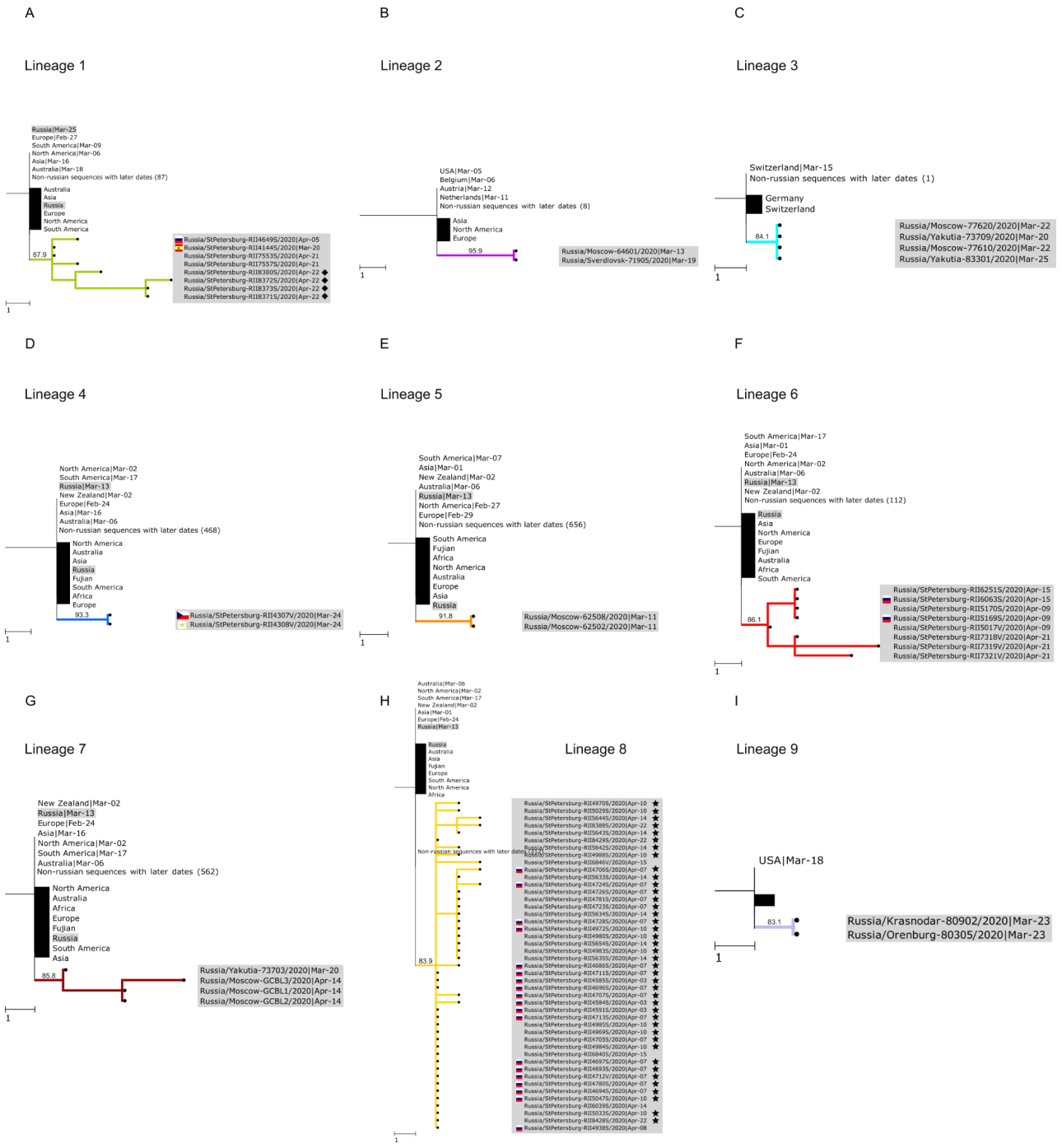
To estimate the number of introduction events, we used the following procedure. When a Russian lineage or singleton had no Russian sequences in their stem (e.g., Figs. 4.4c, 4.4d, 4.5a), we assumed that they originated from distinct introductions. There were 3 such Russian transmission lineages (lineages 2, 3, and 9) which together included 8 sequences; and 33 such singletons (Supplementary Figs. B-2,3), for a total of 36 introduction events.

Additionally, some of the Russian lineages descended from internal nodes with a mix of Russian and non-Russian sequences (e.g., Fig. 4.4a-b). Similarly, in a fraction of cases, a Russian singleton descended from an internal node with multiple sequences corresponding to it, such that some of them were Russian. 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) (Fig. 4.3).

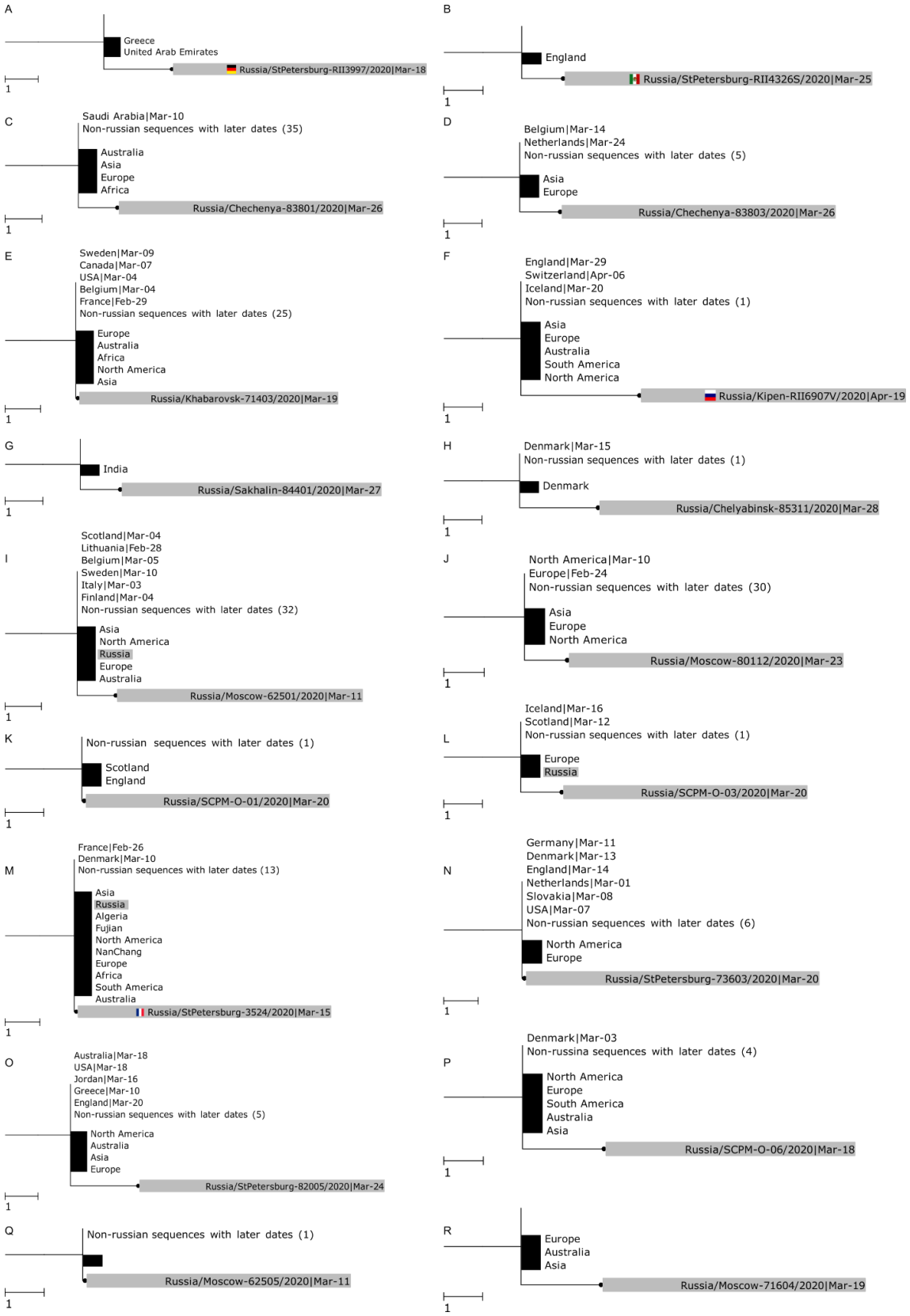
To address this, we used the following statistical procedure. We used the fact that for a fraction of samples, direct travel data were available: we had information on travel abroad or the absence of travel history of the sampled individuals. We assumed that these data are reflective of the fraction 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 traveled abroad, this lineage was introduced; and if some of the individuals had documented absence of travel (but none had traveled abroad), this lineage was not 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 B-1).

Using this procedure, we estimate that sequences among these three categories result from additional ~3 introductions yielding transmission lineages (three of the lineages 1, 4, 5, 6, 7, and 8 with 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 36+31=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.

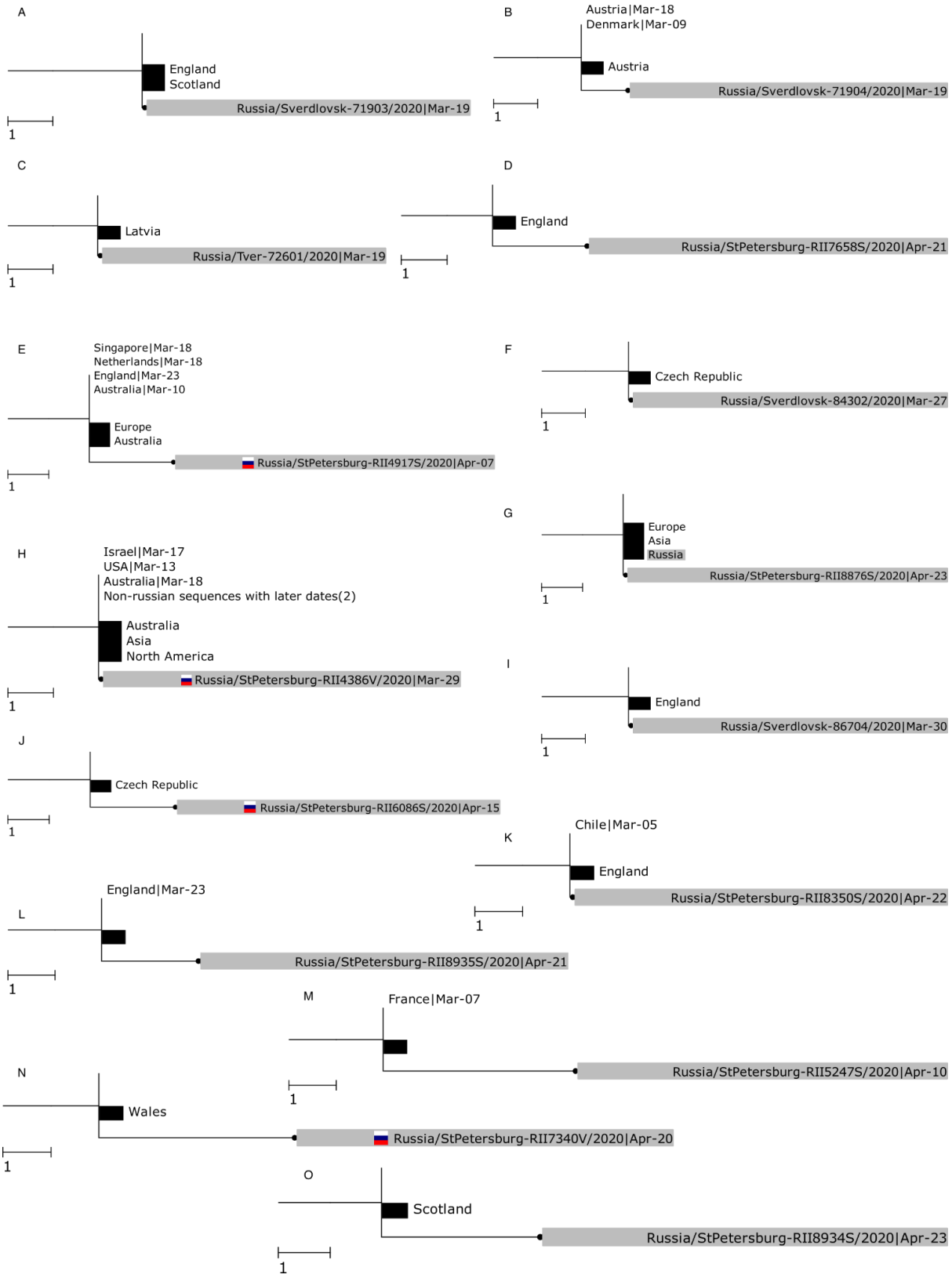
Supplementary Figures B



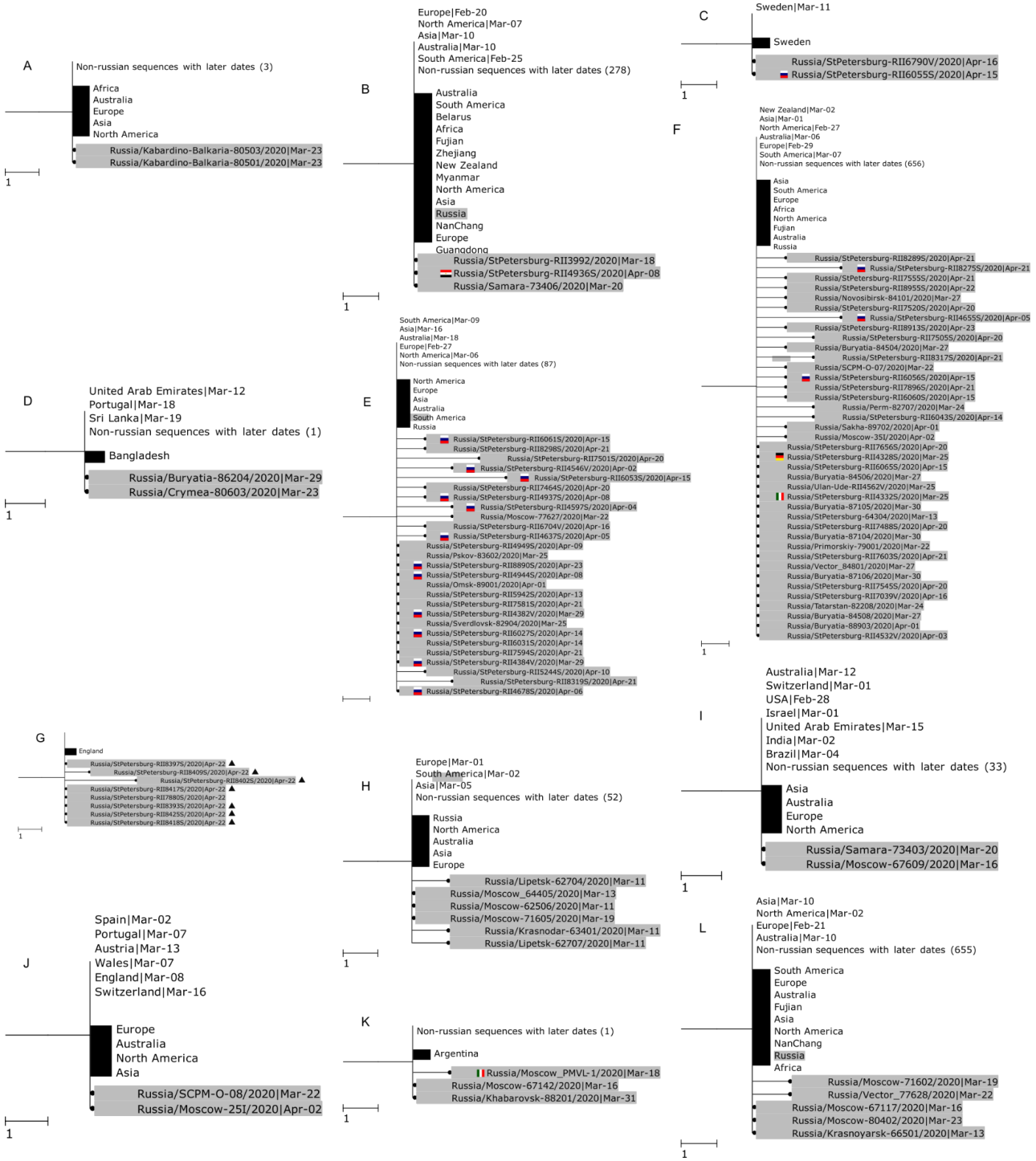
Supplementary Figure B-1. Russian transmission lineages and stem-derived transmission lineages. Each panel (A-I) shows an independent transmission lineage. Lineage colors are as in Figs. 4.2 and 4.4-4.6. Notation as in Fig. 4.4. Stars and diamonds in (A) and (I) indicate samples associated with the Vreden hospital (see Fig. 4.8).



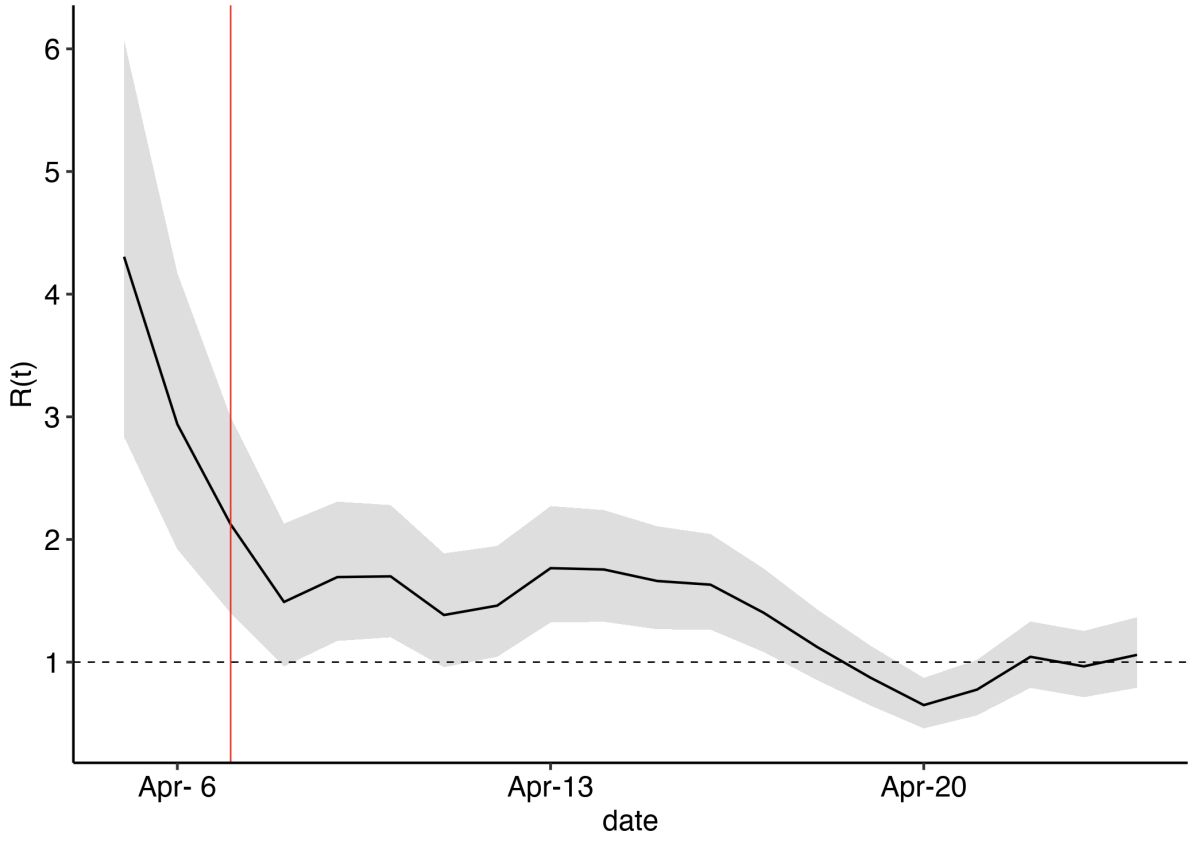
Supplementary Figure B-2. Russian singletons (continued on Supplementary Figure B-3). Each panel (A-O) shows an individual Russian singleton. Notation as in Figs. 4.4-4.5.



Supplementary Figure B-3. Russian singletons (continued from Supplementary Figure B-2). Each panel (A-R) shows an individual Russian singleton. Notation as in Figs. 4.4-4.5.



Supplementary Figure B-4. 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. 4.4-4.5. Triangles in (G) represent samples associated with the Vreden hospital outbreak (see Fig. 4.8).



Supplementary Figure B-5. Estimated R_e (mean (solid line) and 95% credible interval (shaded area) inferred from incidence data on moderately/severely ill patients in Vreden hospital. The red line marks the introduction of quarantine on Apr-7.

Supplementary Tables B

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

	lineages	sequences	travel history		Estimated number of imports
			yes	no	
Russian stem-derived transmission lineages	6	-	2 (50%)	2	=6*0.50=3.00
Russian stem-derived singletons	-	40	1 (14%)	6	=40*0.14=5.60
Russian stem clusters	-	61	4 (36%)	7	=61*0.36=21.96
Total					30.56

Supplementary Table B-2. Symptom onset dates for the 11 sequences for which these data are available. The 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 B-3. Vreden hospital samples per collection date. 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

Multi-rho birth-death skyline model. Supplementary Tables B-4,5,6 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 provides evidence for multiple introductions. We used Tracer v1.7.1(Rambaut et al., 2018) to summarise the results.

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

Parameter	Estimate	95% credible interval
TMRCA date	February 21	January 20 — March 21
clockRate	9.41E-4	[8.44E-4, 1.04E-3]

rho1	0.19	[0.03, 0.39]
rho2	0.44	[0.25, 0.64]
rho3	0.43	[0.21, 0.66]
rho4	0.37	[0.13, 0.63]
rho5	0.61	[0.21, 1.00]
reproductiveNumber1 before March 27	0.94	[0.54, 1.41]
reproductiveNumber2 March 27-April 8	3.00	[1.85, 4.25]
reproductiveNumber3 after April 8	1.76	[0.91, 2.71]

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

Parameter	Estimate	95% credible interval
TMRCA date	March 24	March 6 — April 1
clockRate	9.40E-4	[8.44E-4, 1.04E-3]
rho1	0.26	[0.05, 0.51]
rho2	0.48	[0.28, 0.69]
rho3	0.50	[0.25, 0.74]
rho4	0.48	[0.19, 0.77]
rho5	0.65	[0.22, 1.00]
reproductiveNumber1 before March 27	0.67	[0.08, 1.01]
reproductiveNumber2 March 27-April 8	3.70	[2.08, 5.48]
reproductiveNumber3 after April 8	1.70	[0.68, 2.84]

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

Parameter	Estimate	95% credible interval
TMRCA date	March 26	March 13 — April 2
clockRate	9.36E-4	[8.40E-4, 1.03E-3]
rho1	0.28	[0.05, 0.56]
rho2	0.52	[0.30, 0.74]
rho3	0.56	[0.30, 0.82]
rho4	0.56	[0.22, 0.88]
rho5	0.45	[0.03, 0.94]
reproductiveNumber1 before March 27	1.21	[0.40, 2.95]
reproductiveNumber2 March 27-April 8	3.64	[2.01, 5.43]
reproductiveNumber3 after April 8	1.85	[0.77, 3.06]

Multi-rho birth-death skyline model with independent tree models. We ran the multi-rho birth-death skyline model on groups 1, 2, and 3 assuming that each group has an independent tree model in order to address possible biases due to the population structure given the strong evidence of three independent introductions into the hospital similar to the analysis in (Vasylyeva et al., 2019) (Supplementary Table B-7). Though, groups 2 and 3 have only a few sequences representing each of them collected on a single date of April 22. We find parameter estimates for groups 2 and 3 seem to be misleading in this model.

Supplementary Table B-7. Phylodynamic parameter estimates for groups 1, 2, and 3 with independent tree models. The parameter estimates were obtained using BEAST2 with the multi-rho birth-death skyline model with three separate trees.

Parameter	Estimate	95% credible interval
TMRCA date 1	March 26	March 19 — April 2
TMRCA date 2	April 3	March 15 — April 17
TMRCA date 3	March 27	March 2 — April 15

clockRate	9.33E-4	[8.36E-4, 1.03E-3]
rho1.1	0.27	[0.04, 0.55]
rho1.2	0.53	[0.32, 0.76]
rho1.3	0.57	[0.29, 0.81]
rho1.4	0.57	[0.24, 0.89]
rho1.5	0.45	[0.01, 0.92]
rho2	0.54	[0.11, 1.00]
rho3	0.43	[7.82E-3, 0.92]
reproductiveNumber1.1 before April 8	3.44	[1.73, 5.48]
reproductiveNumber1.2 after April 8	1.89	[0.80, 3.13]
reproductiveNumber2.1 before April 8	1.15	[0.41, 2.65]
reproductiveNumber2.2 after April 8	2.13	[0.88, 3.63]
reproductiveNumber3.1 before April 8	1.11	[0.45, 2.42]
reproductiveNumber3.2 after April 8	1.76	[0.64, 3.09]

Serial birth-death skyline model. The fifty-two Vreden samples were collected on 5 distinct dates, with a substantial lag between subsequent collection dates (see Supplementary Table B-3). For some of the samples, the sample collection date could differ substantially from the symptoms onset date. We expect that there might be possible bias due to the loss of within-patient variation due to variant calling procedure. To address this possible bias, we used the symptoms onset date instead of the collection date in BEAST2 analysis and the tip-sampling for the rest of the samples. For 11 samples, the symptoms onset dates were known (Supplementary Table B-2). For each of the remaining 41 samples, we produced a posterior estimate of its symptoms onset date by using a uniform prior between March 31st and the collection date. The possible problem of this approach though is the assumption of the birth-death skyline model that sampling immediately results in the death of the lineage. In other words, the resulting model assumes that the patients were isolated immediately after showing the symptoms. We could not verify this assumption, and it is likely that the patients were isolated only after the Covid-19 was confirmed by PCR test. More detailed analysis

including simulations is needed to verify whether this approach is more or less reliable than the birth-death model with rho-sampling, which we used in the main text.

Supplementary Tables B-8,9,10 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 provides evidence for multiple introductions. We used Tracer (Rambaut et al., 2018) to summarise the results.

Supplementary Table B-8. Phylodynamic parameter estimates for groups 1, 2, and 3. The parameter estimates were obtained using BEAST2 with the serial birth-death skyline model.

Parameter	Estimate	95% credible interval
TMRCA date	February 4	January 1 — March 7
reproductiveNumber1 before March 27	0.92	[0.60, 1.16]
reproductiveNumber2 March 27-April 8	3.72	[2.48, 5.05]
reproductiveNumber3 after April 8	1.38	[0.48, 2.41]
samplingProportion1	0.0 (fixed)	--
samplingProportion2	0.79	[0.46, 1.00]
samplingProportion3	0.10	[0.01, 0.25]
samplingProportion4	0.01	[4.77E-8, 0.05]
clockRate	9.43E-4	[8.46E-4; 1.04E-3]

Supplementary Table B-9. Phylodynamic parameter estimates for groups 1 and 2. The parameter estimates were obtained using BEAST2 with the serial birth-death skyline model.

Parameter	Estimate	95% credible interval
TMRCA date	March 15	February 25 — March 31
reproductiveNumber1 before March 27	1.12	[0.46, 2.42]

reproductiveNumber2 March 27-April 8	3.96	[2.52, 5.49]
reproductiveNumber3 after April 8	1.30	[0.47, 2.26]
samplingProportion1	0.0 (fixed)	--
samplingProportion2	0.81	[0.50, 1.00]
samplingProportion3	0.15	[0.02, 0.35]
samplingProportion4	0.02	[1.40E-7, 0.06]
clockRate	9.40E-4	[8.44E-4; 1.04E-3]

Supplementary Table B-10. Phylodynamic parameter estimates for group 1. The parameter estimates were obtained using BEAST2 with the serial birth-death skyline model.

Parameter	Estimate	95% credible interval
TMRCA date	March 23	March 11 — March 30
reproductiveNumber1 before March 27	1.28	[0.42, 3.12]
reproductiveNumber2 March 27-April 8	4.02	[2.46, 5.69]
reproductiveNumber3 after April 8	1.30	[0.48, 2.26]
samplingProportion1	0.0 (fixed)	--
samplingProportion2	0.82	[0.52, 1.00]
samplingProportion3	0.15	[0.02, 0.35]
samplingProportion4	0.02	[3.35E-8, 0.05]
clockRate	9.38E-4	[8.42E-4; 1.03E-3]

Supplementary Table B-11. 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 (Pybus et al., 2020). Other priors are the same or similar to those used in the birth-death skyline analysis in (Stadler, 2020).

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	Effective reproductive number	Log Normal(0.8, 0.5)
	Date of infection origin	Uniform(0, 1000)
	Become uninfectious rate	36.5 (fixed)
multi-rho	Rho sampling probability	Beta(1, 1)
serial	Sampling proportion	Uniform(0, 1)

Supplementary References B

Pybus, O. et al. (2020) Preliminary analysis of SARS-CoV-2 importation & establishment of UK transmission lineages, Virological. Available at: https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507/ (Accessed: 8 July 2020).

Rambaut, A. et al. (2018) 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7', Systematic Biology, pp. 901–904. doi: 10.1093/sysbio/syy032.

Stadler, T. (2020) Phylodynamic Analyses of outbreaks in China, Italy, Washington State (USA), and the Diamond Princess, Virological. Available at: https://virological.org/t/phylodynamic-analyses-of-outbreaks-in-china-italy-washington-state-usa-and-the-diamond-princess/439 (Accessed: 8 July 2020).

Vasylyeva, T. I. et al. (2019) 'Tracing the Impact of Public Health Interventions on HIV-1 Transmission in Portugal Using Molecular Epidemiology', The Journal of infectious diseases, 220(2), pp. 233–243.