SARS-CoV-2 transmission dynamics in Belarus revealed by genomic and incidence data analysis.

Alina Nemira[1], Ayotomiwa Ezekiel Adeniyi[1], Elena L. Gasich[3], Kirill Y. Bulda[3], Leonid N. Valentovich[4], Anatoly G. Krasko[3], Olga Glebova[1], Alexander Kirpich[2,*], and Pavel Skums[1,*,°]

[1]Department of Computer Science, Georgia State University, Atlanta, Georgia, USA
[2]Department of Population Health Sciences, School of Public Health, Georgia State University, Atlanta, Georgia, USA
[3]Republican Research and Practical Center for Epidemiology and Microbiology, Minsk, Belarus
[4]Institute of Microbiology, National Academy of Sciences of Belarus, Minsk, Belarus

Abstract

Since the emergence of COVID-19, a series of non-pharmaceutical interventions (NPIs) has been implemented by governments and public health authorities worldwide to control and curb the ongoing pandemic spread. From that perspective, Belarus is one of a few countries with a relatively modern healthcare system, where much narrower NPIs have been put in place. Given the uniqueness of this Belarusian experience, the understanding its COVID-19 epidemiological dynamics is essential not only for the local assessment, but also for a better insight into the impact of different NPI strategies globally. In this work, we integrate genomic epidemiology and surveillance methods to investigate the emergence and spread of SARS-CoV-2 in the country. The observed Belarusian SARS-CoV-2 genetic diversity originated from at least eighteen separate introductions, at least five of which resulted in ongoing domestic transmissions. The introduction sources represent a wide variety of regions, although the proportion of regional virus introductions and exports from/to geographical neighbors appears to be higher than for other European countries. Phylodynamic analysis indicates a moderate reduction in the effective reproductive number $\mathcal{R}_e$ after the introduction of limited NPIs, with the reduction magnitude generally being lower than for countries with large-scale NPIs. On the other hand, the estimate of the Belarusian $\mathcal{R}_e$ at the early epidemic stage is comparable with this number for the neighboring ex-USSR country of Ukraine, where much broader NPIs have been implemented. The actual number of cases by the end of May, 2020 was predicted to be 2-9 times higher than the detected number of cases.

Keywords: COVID-19, SARS-CoV-2, Belarus, genomic epidemiology, phylodynamics, effective reproduction number

° Corresponding author: pskums@gsu.edu

* The authors contributed equally.

1   Introduction

The Republic of Belarus is a country in Eastern Europe with a population of approximately 9.5 million. In comparison to other ex-USSR and Central European countries, it is characterized by weaker socio-economic and political ties with the neighboring European Union [20, 52] and lower outward population mobility [9]. At the same time, Belarus has a relatively modern healthcare system [45], and the country's Human Development Index (HDI) is categorized as "very high" (vhHDI) [13].

The COVID-19 epidemic has reached Belarus later than most of Western European countries and approximately at the same time as its neighbors. The first confirmed imported case reported on February 28, 2020 was a person who arrived from Iran [5, 3]. Since then, there was a steady increase in the number of officially reported laboratory-confirmed cases that has surpassed $300,000$ on March 13, 2021.

The major feature of the COVID-19 pandemic in Belarus is the significantly narrower scope of non-pharmaceutical interventions (NPIs) in comparison to other vhHDI countries [17]. The implemented NPIs included a mandatory 14-day self-isolation for individuals who were arriving from abroad or were identified as close contacts of individuals with confirmed COVID-19; some social distancing measures such as the increase in frequency of public transportation operations to reduce crowding; remote teaching and delaying the class starting times at schools and higher education institutions [6]. No large-scale quarantines, lockdowns or other strict social distancing measures have ever been administered. The other widely practiced measures such as mask regimen and border closures were not mandated until November and December of 2020, respectively.

Given the uniqueness of the Belarusian experience, understanding of COVID-19 epidemiological dynamics in this country is essential not only for assessment of its past and current public health situation, but also for a better insight into the impact of different NPI strategies around the globe. However, the development of such understanding has been impeded by the limited amount of available data. Until the last quarter of 2020 the only available data have been the officially reported country-level counts that included daily incidence, numbers of conducted diagnostic tests, and COVID-related mortality. Such statistics are prone to biases and underreporting [31, 53]. While these drawbacks are well-known and common for all countries, they have a potential to be exacerbated in Belarus due to limited testing capacities provided by a handful of national-level laboratories [6].

In the meantime, whole-genome sequencing (WGS) data analyzed using genomic epidemiology methods provides a complementary and independent source of information. WGS SARS-CoV-2 data have already been used to study transmission histories and epidemiological dynamics in a variety of countries and administrative regions [35, 27, 15, 47, 37, 23, 28, 24, 22]. For Belarus sufficiently representative genomic dataset has become available only in the late 2020, when the limited sequencing data produced outside of the country on behalf of the World Health Organization (WHO) were extended by the locally produced sequences.

In this paper, we combined WGS genomic data and epidemiological data to carry out the first study of SARS-CoV-2 transmission dynamics in Belarus. In the absence of significant amounts of reliable epidemiological statistics, the integrated genomic and incidence analysis allowed to fill the information gap and provide a plausible picture of the emergence and spread of SARS-CoV-2 in the country. The obtained results also gave

insight into the effect of limited NPIs during the first epidemic wave.

2 Methods

2.1 Data

The SARS-CoV-2 genomic data for analysis were downloaded from GISAID [48] on March 15, 2021. The Belarusian dataset consists of 41 full-length genomes sampled between March 2020 and February 2021. One sequence was obtained from a citizen of Azerbaijan who was tested in Belarus to be allowed to return to his home country. This sequence is marked as Azerbaijanian in GISAID, but is considered as Belarusian here. The Ukrainian dataset that was analyzed for comparison purposes consists of 116 sequences. Daily numbers of new cases and conducted tests were collected from the official Telegram channel of the Ministry of Health of the Republic of Belarus [10]

2.2 Global phylogenetic analysis

For the phylogeny reconstruction, we utilized the SARS-CoV-2-specific phylogenetic inference pipeline implemented in Nextstrain [29]. The sequences from Belarus were analyzed together with $12,064$ background sequences from the global SARS-CoV-2 population. To obtain a representative sample with those background sequences, a country-specific Nextstrain context subsampling was used [29]. The sequences were aligned using MAFFT [34], and a maximum likelihood (ML) phylogenetic tree was constructed using IQ-TREE [39] under Hasegawa-Kishino-Yano (HKY)+Γ nucleotide substitution model with a gamma-distributed site rate variation [30].

In the resulting time-labelled tree, ancestral geolocation traits have been inferred using so-called "mugration model" [46]. In this model, countries of origin of the tree nodes are considered as discrete traits, and the virus spread between countries is considered as a general time reversible process. We augmented this model by incorporating the human mobility statistics provided by European Commission Knowledge Center on Migration and Demography (KCMD) [44] via KCMD Dynamic Data Hub [9]. Even though global travel has been affected by COVID-19-related restrictions, this statistics are still assumed to representatively reflect the *relative* density of human mobility between countries even in quarantine settings. Specifically, the transition rates between traits were assumed to be proportional to the normalized average numbers of inter-country trips. The resulting transition rate matrix has been used to estimate the maximum joint likelihood traits of internal nodes using the dynamic programming algorithm [40]. This trait inference algorithm has been implemented in Matlab (v. R2019b).

Belarusian clades were defined as those having the most recent common ancestors (MRCA) with "Belarus" trait, and intra-Belarusian lineages were inferred as the maximal subtrees inside these clades. Upon examination of the Belarusian clades, we joined two clusters that have the same estimated source trait and the MRCA at the tree distance of 4 from both of them. Finally, global lineages of sequences were determined using Pangolin SARS-CoV-2 Lineage Assigner [11].

2.3   Intra-country phylodynamic analysis

In this work we largely followed a general analytic pipeline adopted in other similar country-level studies (see e.g. [27, 36, 35, 47]), with several modification tailored for the specifics of the analyzed data. At first, temporal signal was evaluated by constructing an ML phylogeny under HKY+Γ nucleotide substitution model and by regressing root-to-tip genetic divergence against sampling dates using TempEst (v.1.5.3) [43]. Next, BEAST (v.2.6.3) [18] was used to fit the Coalescent Bayesian Skyline model to the full set of Belarusian sequences. As before, HKY+Γ nucleotide substitution model was used together with a strict molecular clock. The clock rate was assumed to follow a gamma (Γ) distribution with the mean equal to 8×10^{-4} mutations/site/year and the standard deviation of 5×10^{-4} [16, 27], where the distribution density was parametrized using the corresponding shape and rate parameters. Four segments were assumed for the effective population size that roughly corresponded to growth and decline periods of the first and second COVID-19 epidemic waves. The model parameters were sampled from the corresponding posterior distribution using Markov Chain Monte Carlo (MCMC) method with 3×10^{7} iterations, sampling every 3×10^{3} iterations and the initial 10% "burn-in" iterations. The MCMC sampling quality was assessed using Tracer (v.1.7.1) [41] and accepted if all parameters had effective sampling sizes (ESS) higher than 200. The obtained maximum clade credibility (MCC) tree was annotated using Tree Annotator (v.1.8.4) [33]. The reliability of intra-Belarusian clusters detected by the ML phylogenetic inference was re-confirmed by verifying their correspondence to monophyletic clades in the MCC tree. For each cluster, time to the most recent common ancestor (TMRCA) was estimated.

The effective reproduction number $\mathcal{R}_e$ and the sampling proportion have been estimated for the two best-sampled Belarusian transmission lineages with a total of 19 genomes (Supplemental Table S3) using Birth Death Skyline Serial (BDSKY) model [49] implemented in BEAST. The analyzed lineages were likely co-circulating over the same susceptible population (see Results). Thus, we used a linked model where both lineages evolve and are being sampled independently but share the substitution model parameters, the molecular clock rate and the effective reproduction number drawn from the same respective priors. Given the relative sparsity of available genomic data, this approach allows to use larger and more representative combined sample for the analysis. The same settings as above have been used for the substitution model, molecular clock and MCMC. Since BDSKY model is parameter-rich, we equipped it with the informative priors on several parameters. Specifically, the sampling proportions were assumed to have a $Beta(\alpha, \beta)$ distribution prior with parameters $\alpha = 1$ and $\beta = 9.99 \cdot 10^5$, thus reflecting the sparsity of Belarusian sequence sample (the proportion of sequenced cases from the total number of cases is assumed to vary between 10^{-6} and 10^{-3}). The prior for the origin of each cluster was assumed to be normally distributed with the mean equal to the time estimated using the Coalescent Bayesian Skyline. For the rate of becoming non-infectious, we assumed an infectious period of 10 days [27, 32, 35, 38]. Finally, we considered the models with one and two changes of the effective reproduction number $\mathcal{R}_e$ and the sampling proportion. The times of the parameters change were fixed to July 1, 2020 for the first model and May 1 and July 1, 2020 for the second model. The list of model parameters is reported in the Supplemental Table S1.

4

2.4   Inference of case counts

Here we used two complementary approaches. In the first approach, trees and BDSKY parameters sampled by BEAST were used to reconstruct cumulative case count trajectories using the particle filter algorithm implemented in EpiInf (v.7.3.0) [51]. In the second approach, we utilized the method of [53]. It quantifies the case counts underestimation from the numbers of confirmed cases and conducted tests up to a specified date in a semi-Bayesian way under the assumptions that the observed data are subject to sampling, reporting and diagnosis biases. The model [53] has been used with the default settings. The case count trajectories were inferred by taking 10^4 samples from model-defined prior distributions of testing probabilities for individuals with different severity of symptoms.

3   Results

SARS-CoV-2 genomic diversity Despite the sparse sampling, the observed Belarusian SARS-CoV-2 sequences belong to 11 genomic lineages (by the nomenclature of [42], Fig.1A). In particular, the genome that was sampled on February 23, 2021 belongs to B.1.1.7 lineage that emerged in the UK in November 2020 and had been rapidly spreading toward fixation [21]. The root-to-tip regression analysis demonstrated moderately strong temporal signal ($R^2 = 0.56$, $p < 10^{-6}$, Fig.1B).

SARS-CoV-2 transmission history We identified 18 distinct intra-Belarusian clades that most likely correspond to separate introductions of SARS-CoV-2 into the country. The inference of between-country importations of SARS-CoV-2 is usually complicated, since during the global pandemic close genomic variants can be observed in multiple geographic locations. Therefore, the results of such inference should be treated with caution. With that in mind, we note that the inferred transmission history agreed with the travel records for those cases when they were available. In particular, the first confirmed SARS-CoV-2 case was the individual who arrived from Iran [3], and the phylogenetics reaffirmed that. The agreement also held for the second detected case brought by the travelled from Italy [1]. The first introduction produced at least one secondary case as indicated by the tree; however, both lineages were not sampled after March, 2020 (Fig.2B). This can be attributed to the timely isolation of those individuals and their first order contacts [2]. In general, SARS-CoV-2 importations into the country could be attributed to a mixture of regional and global transmissions. As illustrated in Fig.1B the most frequent alleged virus introduction sources were the neighboring countries of Russia (5 introductions) and Poland (3 introductions).

Five SARS-CoV-2 introductions (28%) are associated with clusters of two or more sequences, and thus are hypothesized to establish intra-country transmission lineages. Three largest transmission lineages are paraphyletic and may indicate virus re-export from Belarus to other countries. Even though some alleged export cases could be sampling artefacts, those of them involving large lineages are more reliable. Such cases include two SARS-CoV-2 introductions to the neighboring country of Latvia in June, 2020 (95% CI: May 31, 2020 - June 26, 2020) and in October, 2020 (95% CI: October 1, 2020 - October 22, 2020) that established substantial Latvian transmission lineages (Supplemental Fig.S2)

Effective reproduction number and the effect of NPIs. Most observed clades originated between March and July of 2020, and the majority of their times to MRCA fall
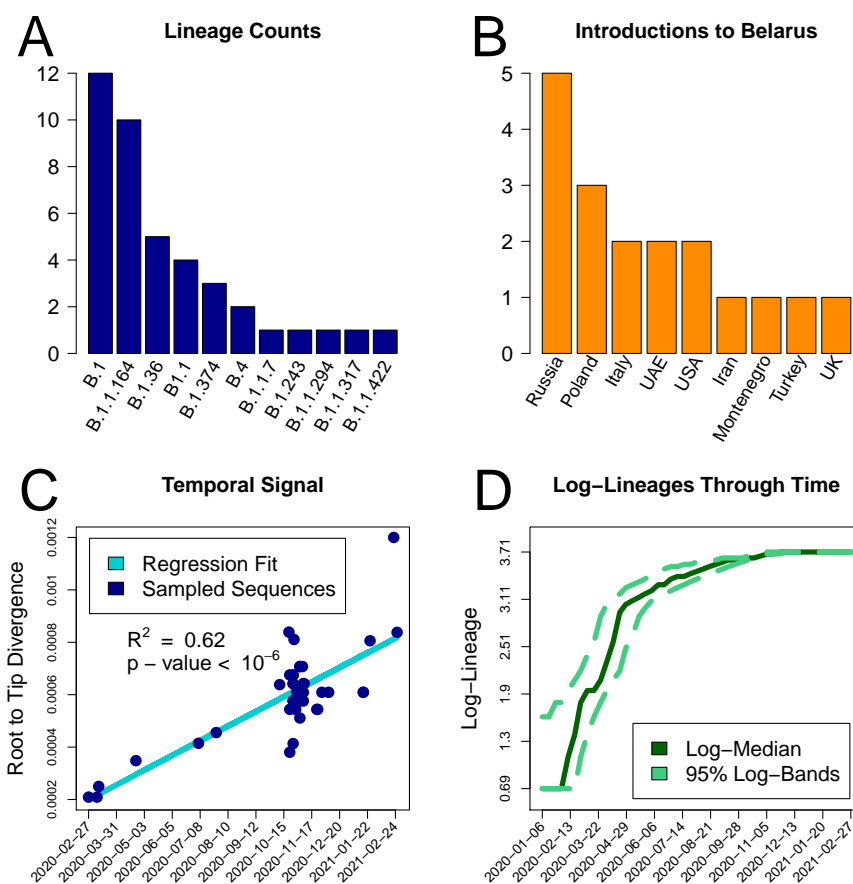
Figure 1: Lineages summaries: A) Abundances of genomicthe c lineages. B) Estimated introduction sources to Belarus. C) Temporal signal. D) Lineages-through-time (on logarithmic scale).

into April (Fig. 1D, Fig.2B and Supplementary Table S4). The majority of branching events also belonged to the same time period. It implies that, despite sequencing being performed mostly in late 2020 – early 2021, the phylodynamic analysis of currently available Belarusian genomes allows us to reliably assess only the first epidemic wave prior to July 2020. Another reason to choose July, 1 for the endpoint of our phylodynamics analysis is the dynamics of the daily percentage of positive tests. The WHO criterion for influenza-like-illnesses (ILI) assumes that epidemic is "under control" if the percentage of positive tests is below 5% for at least two weeks [14]. According to the officially reported data, Belarus reached this state with respect to the first COVID-19 wave by the end of June, 2020 (Fig.4D), even though the reported incidence peaked several weeks earlier.

Best-sampled transmission clusters are well-mixed and have representatives from at least two Belarusian administrative regions (Fig.2B). This fact and the relative homogeneity of the Belarusian demographical characteristics suggest that the corresponding viral lineages co-circulated over the same susceptible population. Thus, we estimated the effective reproduction number $\mathcal{R}_e$ for these lineages using a linked BDSKY model. The model with three segments shows a moderate decline of the median $\hat{\mathcal{R}}_e$ from $\hat{\mathcal{R}}_e = 1.95$ (95% highest posterior density (HPD) interval: $(1.03; 2.99)$) in March-April to $\hat{\mathcal{R}}_e = 1.59$ (95% HPD interval: $(0.82; 2.39)$) in May-June (Fig. 3B). The obtained HPD intervals, however, are rather wide due to the relatively small genome sample size. Thus, we also
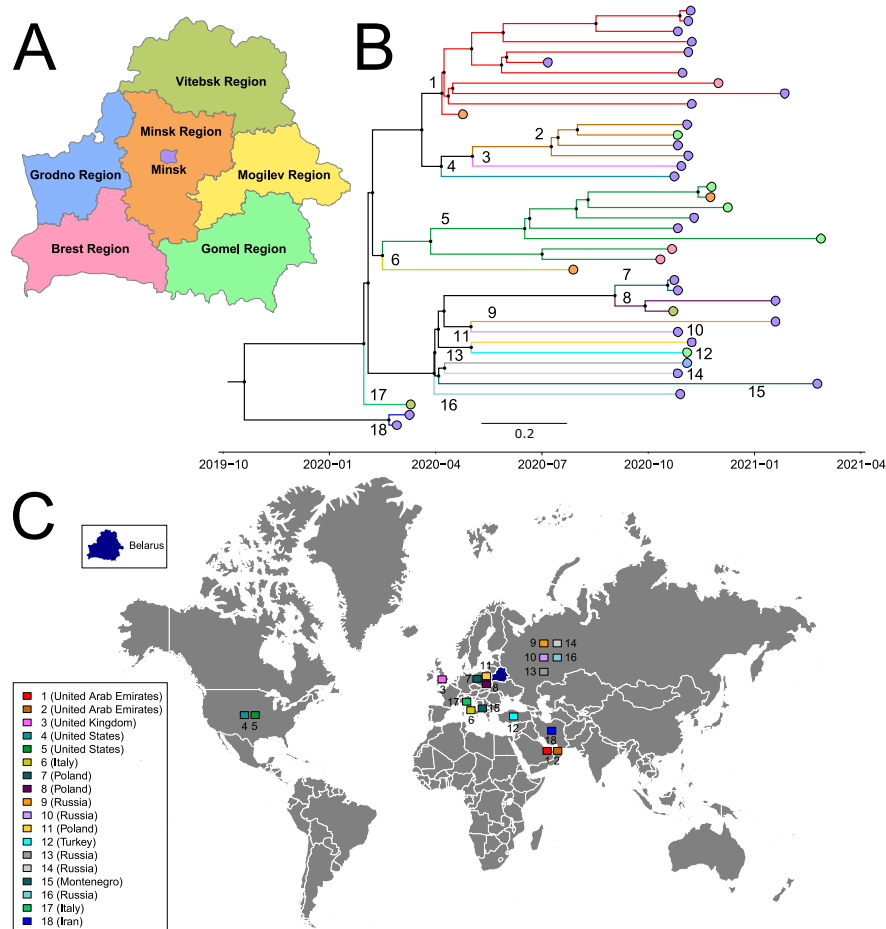
6

Figure 2: The annotated maximum clade credibility tree. A) Administrative regions of Belarus. B) The tree with the leaves color-coded by sampling regions (using the colors from panel A) and with the branches coded by the corresponding clusters. C) Cluster sources marked on the world map by their ids that correspond to panel B. Maps for figures were downloaded from Vemaps.com

estimated the median $\hat{\mathcal{R}}_e$ for the entire period of March-June, which turned out to be $\hat{\mathcal{R}}_e = 1.70$ (95% HPD interval: $(1.45; 1.96)$). The Kolmogorov-Smirnov test was used for the formal comparison of prior and posterior distribution samples for $\hat{\mathcal{R}}_e$ and resulted in $p < 10^{-10}$ for all of them.

In addition, we matched the estimate of the effective reproduction number $\hat{\mathcal{R}}_e$ for Belarus against that for Ukraine - the neighboring ex-USSR non-EU country with similar demographics. The major difference in COVID-19 epidemics between Belarus and Ukraine is the scope of NPIs, with Ukraine implementing much stricter lockdown and physical distancing policies [17]. The same Birth-Death Skyline Serial model was applied to two best-sampled Ukrainian clusters with the total of 28 sequences defined as in [25] (Supplemental Table S3). The median Ukrainian $\hat{\mathcal{R}}_e$ over the same tine period was estimated to be $\hat{\mathcal{R}}_e = 1.64$ (95% HPD interval: $(1.49; 1.81)$). This assessment agrees with the previous estimation based on Exponential Coalescent model [25] and appeared to be comparable to $\hat{\mathcal{R}}_e$ estimates for Belarus.

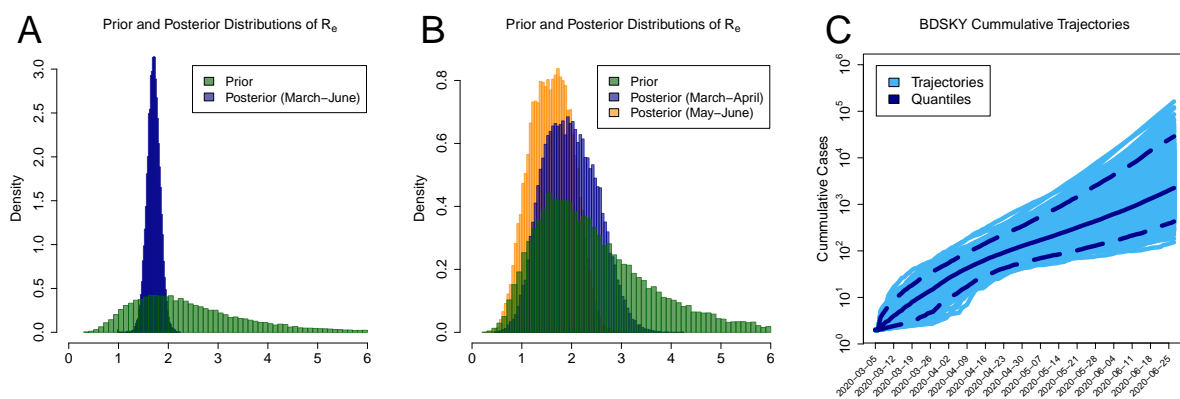**Cumulative incidence and case counts.** Cumulative case count trajectories for Be-

7

Figure 3: BDSKY model estimations. A), B) prior (green) and posteriors (blue and orange) distributions of the effective reproduction number estimate $\hat{\mathcal{R}}_e$ for Belarus during the first COVID-19 wave. C) The cumulative case count trajectories on $\log_{10}$ scale. Solid blue and dashed lines represents a median and 95% confidence intervals, respectively.

larus implied by the BDSKY model are reported on Fig.3C. The cumulative number of cases by July 1, 2020 falls into the 95% prediction interval: $(364; 17066)$. It should be kept in mind that these estimates apply only to two transmission lineages out of possibly many more.

The results of the complementary analysis based on the number of officially reported cases $D(t)$ and conducted tests $T(t)$ over time are presented in Fig.4. The model [53] was designed for the initial phase of the epidemics with the exponential growth of $D(t)$. Therefore, we calibrated and used it to estimate the cumulative number of infections $C(t)$ for the time interval from April 1 (the first date when the number of conducted tests was available) to May 16, 2020 (officially reported peak of the first wave) with a 15-day increments. The obtained results suggested a substantial underestimation of the cumulative number of cases through the study period (Fig.4B). In particular, on $t^* =$ May 16, 2020 the model predicted $C(t^*) = 118,521$ cases (95% PI: $(54,057; 249,000)$) while the reported number was $D(t^*) = 28,681$. Hence, 76% of infections occurred by that date were supposedly undetected (95% PI: $(47\%; 88\%)$). The model-inferred case detection rate $D(t)/C(t)$ increases over time as more tests are being conducted (Fig.4C).

4 Discussion

In this paper we presented the first detailed study of COVID-19 epidemic in Belarus using the officially reported incidence data, testing data and genomic data collected between March, 2020 and February, 2021. The reported results significantly expand our understanding of COVID-19 dynamics and effects of limited NPIs in Belarus, and reflect several key epidemiological issues that it shares with other countries around the globe.

First, the analysis revealed the diverse history of transmissions of SARS-CoV-2 into, from and inside the country. It identified 18 introductions within 13 genomic lineages, but this estimate is most likely a lower bound on the real number of introductions, since only a very small fraction of all SARS-CoV-2 genomic diversity has been sampled. In contrast to most Western European and North American countries [27, 15, 47, 37, 23, 28, 24, 38, 22], the larger portion of estimated transmission links was with geographic neighbors. It
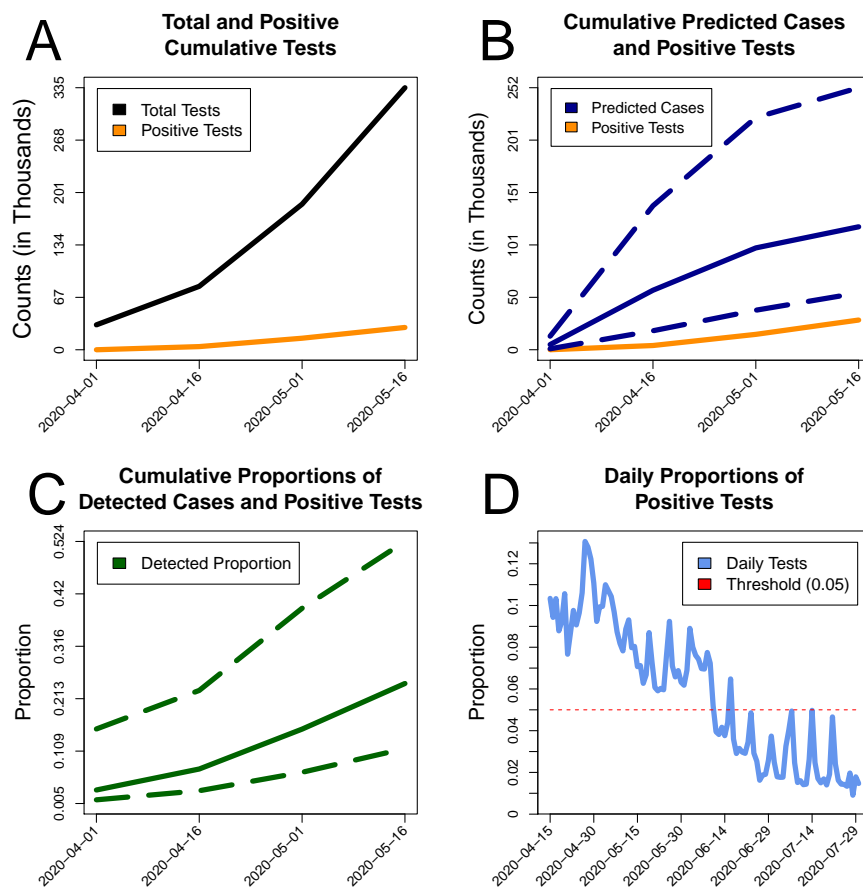
Figure 4: The summary of counts data analysis. A) Input data: officially reported cumulative numbers of cases $D(t)$ (orange) and conducted tests $T(t)$ (black). B) The cumulative numbers of officially reported cases ($D(t)$, orange) and the counts that were inferred by the model ($C(t)$, blue). C) Model-based case detection rate $D(t)/C(t)$ (green). D) The daily proportion of positive tests (light blue) together with the suggested WHO threshold of 0.05 (red). In panels B and C, solid blue and green lines represent median estimates across 10^4 model runs, while dashed lines depict 2.5th and 97.5th percentiles.

is not entirely surprising, given the comparatively lower outward mobility of Belarusian population. It is also worth mentioning that much stricter travel restrictions implemented by the Belarus' neighbors failed to stop the flow of SARS-CoV-2 across the borders in both directions. Furthermore, approximately half of estimated introductions did not appear directly across the border, which emphasize that Belarus, like most countries in the world, is a part of global interconnected environment and as such, affects and is affected by epidemiological developments in other countries.

Second, the estimation of the effective reproduction number $\mathcal{R}_e$ allowed the preliminary assessment of the effect of limited NPIs implemented in the country during the first epidemic wave. These estimates should be interpreted only in comparison with similar estimates for other countries. The analysis suggests a moderate but statistically significant decrease of $\mathcal{R}_e$ after the NPIs were put in action (Fig. 3B). The magnitude of decrease, however, is lower in comparison to the countries with broader and stricter NPIs (Table S2). Furthermore, the estimated median effective reproduction number $\hat{\mathcal{R}}_e = 1.70$ (CI: $(1.45; 1.96)$) over the entire analyzed period for Belarus is comparable with the estimates

of $\mathcal{R}_e$ in developed countries *before the introduction of strict NPIs* [47, 37, 19, 38]. For example, for Victoria, Australia this value is 1.63 (CI: $(1.45; 1.8)$) [47]. On the other hand, the estimate of $\mathcal{R}_e$ for Belarus is also close to the estimate of $\mathcal{R}_e$ over the same time period for neighboring Ukraine, where the scope of implemented NPIs has been significantly broader. In our opinion, the latter fact is not entirely surprising and is more reflective of the reported extensive violations of lockdown and distancing measures in Ukraine and limited ability of authorities to control the epidemics [4, 7]. Similar estimate $\mathcal{R}_e = 1.76$ $(0.91; 2.71)$ has been also reported for Russia [35] which borders both Belarus and Ukraine. This comparison of three ex-USSR countries suggests that regional demographic and social specifics could be important factors for COVID-19 epidemiology along with NPIs. Study of such factors should be the subject of further investigation.

Third, the true number of infections by the end of May, 2020 is most likely ~ 4 (CI :$(2; 9)$) times higher than the detected number of cases, which is expected for respiratory diseases in general and for COVID-19 in particular [53, 50]. For example, according to observed seroprevalence of SARS-CoV-2 antibodies, in the USA the total number of COVID-19 infections in March-May, 2020 was probably between 6 to 24 times the number of reported cases [31].

It is important to highlight that the presented study has several limitations. The first of them is the scarcity of currently available genomic data, especially in comparison with most of other European countries. Our approach strives to compensates for it by utilizing informative priors and linked models for phylogenetic and phylodynamics inference. BD-SKY models are also sensitive enough and suitable for inference even for smaller genomic datasets. For example, the numbers of sequences and/or density of branching events in this study is similar to those in other studies [49, 38, 26], where meaningful estimates of $\mathcal{R}_e$ have been produced for several epidemics, including SARS-CoV-2. Nevertheless, the inference precision could have been higher, if more SARS-CoV-2 genomes have been available. For Belarus, however, significant expansion of the available genomic dataset in the near future is unlikely, and in our opinion the lack of other studies justifies the need to fill the knowledge gap and to report the results based on the existing data. We also hope that this study will serve as a trigger for further SARS-CoV-2 genomic epidemiology studies in Belarus and will encourage funding increase and the corresponding development and expansion of sequencing facilities for molecular surveillance.

The second limitation is that phylogeographic inference of introduction sources can be sensitive to sampling bias and can be affected by relatively slow accumulation of mutations in SARS-CoV-2 genomes [35, 28, 38]. In particular, even though no transmission links with Ukraine has been detected, it is likely that such links will emerge when more data from both countries will become available. Thus, SARS-CoV-2 phylogeography analysis should always be treated with a grain of salt, even though transmission history presented in this study is consistent enough and agrees with the travel records for those cases when they are available. The source inference for Belarus during the early pandemic can actually be more accurate than for some other regions, since Belarusian lineages were established after most of their source lineages were already sufficiently diversified. The incorporation of the global travel statistics into the "mugration model" of [46] may also have contributed towards the increase in transmission inference accuracy. Finally, for the case of Belarus, even if new data refine estimation of sources of some lineages, the obtained results are likely reflecting a true trend towards the higher prevalence of regional and neighbor-to-neighbor virus importations.

10

The third limitation is the sparsity of the SARS-CoV-2 incidence and testing data. In contrast to other countries [8, 12], Belarusian COVID-19 statistics are currently reported only for the entire country rather than for specific regions. The reported numbers of tests are not dichotomized into first-time tests and retests, those conducted by state or commercial laboratories, PCR and antibody tests. Furthermore, the sampling for testing is likely incomplete and biased towards individuals with COVID-19 symptoms and their close contacts and, for instance, persons who were tested upon arrival or prior to departure from the country. These issues may result in underestimation of the true number of cases, even though we are employing a method that is supposed to take them into account. For example, if a significant number of recovered individuals were tested at least twice, then the adjusted proportion of positive tests among those who are getting tested the first time will be higher and, consequently, the estimates of the number of cases will also increase. Furthermore, the aforementioned issues impede the development of stochastic agent-based models that otherwise can be used for high-precision analysis and forecasting. If (or when) more precise data will become available, it can be used to improve the precision and accuracy of our estimates.

In conclusion, this study demonstrates the power of SARS-CoV-2 surveillance using combined genomic and epidemiological data. For such resource-constrained countries as Belarus, it is vitally important to develop sequencing facilities, detailed statistics and analytical resources to the level already established in other countries. These facilities and resources should become integral parts of the national mechanism to respond to emergence, re-emergence and spread of SARS-CoV-2 and other pathogens.

5 Data availability

The sequences used in this study are available at GISAID [48]. The Matlab scripts, Nextstrain configuration files, BEAST 2 XML files used to perform the described analyses and the acknowledgements table with sequence accession numbers and names of researchers and laboratories who produced the sequences are available at `https://github.com/compbel/COVID-Belarus`.

6 Acknowledgements

7 Contributions

AN performed a phylogenetic analysis, analyzed genomic data and wrote the paper. AEA performed a phylogenetic analysis and analyzed genomic data. EG, KB, LV and AK prepared and handled genomic and associated epidemiological data, carried out the primary sequence processing. OG analyzed genomic data and wrote the paper. AK supervised the incidence data analysis, processed and analyzed incidence data, wrote the paper. PS designed and supervised the study, designed and implemented bioinformatics algorithms, performed a phylogenetic analysis, analyzed genomic data and wrote the paper.

References

[1] Ministry of Health of Republic of Belarus: Belarusian woman from Vitebsk has been tested positive for coronavirus (in Russian). `http://minzdrav.gov.by/ru/sobytiya/minzdrav-respubliki-belarus-informiruet/`, 2020. Accessed: 2021-03-31.

[2] Ministry of Health of Republic of Belarus: Six new coronavirus cases have been confirmed in Belarus (in Russian). `https://minzdrav.gov.by/ru/sobytiya/shest-sluchaev-koronavirusa-podtverzhdeno-v-belarusi/`, 2020. Accessed: 2021-03-31.

[3] Ministry of Health of Republic of Belarus: The first imported case of the coronavirus has been registered in Belarus (in Belarusian). `http://minzdrav.gov.by/sobytiya/v-belarusi-zaregistrirovan-zavoznoy-sluchay-koronavirusa/`, 2020. Accessed: 2021-03-31.

[4] UN report: Impact of COVID-19 on human rights in Ukraine . `https://www.ohchr.org/Documents/Countries/UA/Ukraine_COVID-19_HR_impact_EN.pdf`, 2020. Accessed: 2021-03-31.

[5] WHO Coronavirus disease 2019 (COVID-19). Situation report 40. `http://hdr.undp.org/en`, 2020. Accessed: 2021-03-31.

[6] WHO COVID-19 Technical mission of experts to the Republic of Belarus: 8–11 April 2020. Executive summary. `https://www.euro.who.int/en/countries/belarus/publications/covid-19-technical-mission-of-experts-to-the-republic-of-belarus-811-april-2020.-executive-summary`, 2020. Accessed: 2021-03-31.

[7] Atlantic Council: Ukraine's local authorities and the Covid-19 pandemic. `https://www.atlanticcouncil.org/blogs/ukrainealert/ukraines-local-authorities-and-the-covid-19-pandemic/`, 2021. Accessed: 2021-04-02.

[8] COVID-19 Map - Johns Hopkins Coronavirus Resource Center. `https://coronavirus.jhu.edu/map.html`, 2021. Accessed: 2021-03-31.

[9] KCMD Dynamic Data Hub. `https://bluehub.jrc.ec.europa.eu/migration/app/`, 2021. Accessed: 2021-03-31.

[10] Ministry of Public Health of the Republic of Belarus - Official Telegram Channel (in Russian). `https://t.me/s/minzdravbelarus`, 2021. Accessed: 2021-03-31.

[11] pangolin.cog-uk.io. `https://pangolin.cog-uk.io/`, 2021. Accessed: 2021-03-31.

[12] The National Health Service of Ukraine (NHSU) - COVID-19 Dashboard. `https://nszu.gov.ua/e-data/dashboard/covid19`, 2021. Accessed: 2021-03-31.

[13] United Nations Human Development Reports. `http://hdr.undp.org/en`, 2021. Accessed: 2021-03-31.

[14] WHO. Public health criteria to adjust public health and social measures in the context of COVID-19. `https://www.who.int/docs/default-source/coronaviruse/situation-reports/20201012-weekly-epi-update-9.pdf`, 2021. Accessed: 2021-03-31.

[15] Claudia Alteri, Valeria Cento, Antonio Piralla, Valentino Costabile, Monica Tallarita, Luna Colagrossi, Silvia Renica, Federica Giardina, Federica Novazzi, Stefano Gaiarsa, et al. Genomic epidemiology of sars-cov-2 reveals multiple lineages and early spread of sars-cov-2 infections in lombardy, italy. *Nature communications*, 12(1):1–13, 2021.

[16] Kristian G Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes, and Robert F Garry. The proximal origin of sars-cov-2. *Nature medicine*, 26(4):450–452, 2020.

[17] Anders Åslund. Responses to the covid-19 crisis in russia, ukraine, and belarus. *Eurasian Geography and Economics*, pages 1–14, 2020.

[18] Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, 2014.

[19] Gonché Danesh, Baptiste Elie, Yannis Michalakis, Mircea T Sofonea, Antonin Bal, Sylvie Behillil, Grégory Destras, David Boutolleau, Sonia Burrel, Anne-Geneviève Marcelin, et al. Early phylodynamics analysis of the covid-19 epidemic in france. *medRxiv*, 2020.

[20] Alena A Dastanka. Multilateralism in foreign policy of belarus: European and euarasian dimension. *Regional Formation and Development Studies*, 16(2):16–23, 2015.

[21] Nicholas G Davies, Sam Abbott, Rosanna C Barnard, Christopher I Jarvis, Adam J Kucharski, James D Munday, Carl AB Pearson, Timothy W Russell, Damien C Tully, Alex D Washburne, et al. Estimated transmissibility and impact of sars-cov-2 lineage b. 1.1. 7 in england. *Science*, 2021.

[22] Xianding Deng, Wei Gu, Scot Federman, Louis Du Plessis, Oliver G Pybus, Nuno R Faria, Candace Wang, Guixia Yu, Brian Bushnell, Chao-Yang Pan, et al. Genomic surveillance reveals multiple introductions of sars-cov-2 into northern california. *Science*, 369(6503):582–587, 2020.

[23] Louis du Plessis, John T McCrone, Alexander E Zarebski, Verity Hill, Christopher Ruis, Bernardo Gutierrez, Jayna Raghwani, Jordan Ashworth, Rachel Colquhoun, Thomas R Connor, et al. Establishment and lineage dynamics of the sars-cov-2 epidemic in the uk. *Science*, 371(6530):708–712, 2021.

[24] Joseph R Fauver, Mary E Petrone, Emma B Hodcroft, Kayoko Shioda, Hanna Y Ehrlich, Alexander G Watts, Chantal BF Vogels, Anderson F Brito, Tara Alpert, Anthony Muyombwe, et al. Coast-to-coast spread of sars-cov-2 during the early epidemic in the united states. *Cell*, 2020.

13

[25] Yuriy Gankin, Vladimir Koniukhovskii, Alina Nemira, Gerardo Chowell, Thomas A Weppelmann, Pavel Skums, and Alexander Kirpich. Investigating the first wave of the covid-19 pandemic in ukraine using epidemiological and genomic sequencing data. *medRxiv*, 2021.

[26] Lily Geidelberg, Olivia Boyd, David Jorgensen, Igor Siveroni, Fabrícia F Nascimento, Robert Johnson, Manon Ragonnet-Cronin, Han Fu, Haowei Wang, Xiaoyue Xi, et al. Genomic epidemiology of a densely sampled covid-19 outbreak in china. *Virus evolution*, 7(1):veaa102, 2021.

[27] Jemma L Geoghegan, Xiaoyun Ren, Matthew Storey, James Hadfield, Lauren Jelley, Sarah Jefferies, Jill Sherwood, Shevaun Paine, Sue Huang, Jordan Douglas, et al. Genomic epidemiology reveals transmission patterns and dynamics of sars-cov-2 in aotearoa new zealand. *Nature communications*, 11(1):1–7, 2020.

[28] Ana S Gonzalez-Reiche, Matthew M Hernandez, Mitchell J Sullivan, Brianne Ciferri, Hala Alshammary, Ajay Obla, Shelcie Fabre, Giulio Kleiner, Jose Polanco, Zenab Khan, et al. Introductions and early spread of sars-cov-2 in the new york city area. *Science*, 369(6501):297–301, 2020.

[29] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

[30] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution*, 22(2):160–174, 1985.

[31] Fiona P Havers, Carrie Reed, Travis Lim, Joel M Montgomery, John D Klena, Aron J Hall, Alicia M Fry, Deborah L Cannon, Cheng-Feng Chiang, Aridth Gibbons, et al. Seroprevalence of antibodies to sars-cov-2 in 10 sites in the united states, march 23-may 12, 2020. *JAMA internal medicine*, 180(12):1576–1586, 2020.

[32] Xi He, Eric HY Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y Wong, Yujuan Guan, Xinghua Tan, et al. Temporal dynamics in viral shedding and transmissibility of covid-19. *Nature medicine*, 26(5):672–675, 2020.

[33] Joseph Heled and Remco R Bouckaert. Looking for trees in the forest: summary tree from posterior samples. *BMC evolutionary biology*, 13(1):1–11, 2013.

[34] Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[35] Andrey B Komissarov, Ksenia R Safina, Sofya K Garushyants, Artem V Fadeev, Mariia V Sergeeva, Anna A Ivanova, Daria M Danilenko, Dmitry Lioznov, Olga V Shneider, Nikita Shvyrev, et al. Genomic epidemiology of the early stages of the sars-cov-2 outbreak in russia. *Nature communications*, 12(1):1–13, 2021.

[36] Alessia Lai, Annalisa Bergna, Carla Acciarri, Massimo Galli, and Gianguglielmo Zehender. Early phylogenetic estimate of the effective reproduction number of sars-cov-2. *Journal of medical virology*, 92(6):675–679, 2020.

[37] Danielle Miller, Michael A Martin, Noam Harel, Omer Tirosh, Talia Kustin, Moran Meir, Nadav Sorek, Shiraz Gefen-Halevi, Sharon Amit, Olesya Vorontsov, et al. Full genome viral sequences inform patterns of sars-cov-2 spread into and within israel. *Nature communications*, 11(1):1–10, 2020.

[38] Sarah A Nadeau, Timothy G Vaughan, Jérémie Sciré, Jana S Huisman, and Tanja Stadler. The origin and early spread of sars-cov-2 in europe. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

[39] Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh. Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1):268–274, 2015.

[40] Tal Pupko, Itsik Pe, Ron Shamir, and Dan Graur. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular biology and evolution*, 17(6):890–896, 2000.

[41] Andrew Rambaut, Alexei J Drummond, Dong Xie, Guy Baele, and Marc A Suchard. Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systematic biology*, 67(5):901, 2018.

[42] Andrew Rambaut, Edward C Holmes, Áine O'Toole, Verity Hill, John T McCrone, Christopher Ruis, Louis du Plessis, and Oliver G Pybus. A dynamic nomenclature proposal for sars-cov-2 lineages to assist genomic epidemiology. *Nature microbiology*, 5(11):1403–1407, 2020.

[43] Andrew Rambaut, Tommy T Lam, Luiz Max Carvalho, and Oliver G Pybus. Exploring the temporal structure of heterochronous sequences using tempest (formerly path-o-gen). *Virus evolution*, 2(1):vew007, 2016.

[44] Ettore Recchi, Emanuel Deutschmann, and Michele Vespe. Estimating transnational human mobility on a global scale. *Robert Schuman Centre for Advanced Studies Research Paper No. RSCAS*, 30, 2019.

[45] Erica Richardson, Irina Malakhova, Irina Novik, Andrei Famenka, and WHO. *Belarus: health system review*. World Health Organization. Regional Office for Europe, 2013.

[46] Pavel Sagulenko, Vadim Puller, and Richard A Neher. Treetime: Maximum-likelihood phylodynamic analysis. *Virus evolution*, 4(1):vex042, 2018.

[47] Torsten Seemann, Courtney R Lane, Norelle L Sherry, Sebastian Duchene, Anders Gonçalves da Silva, Leon Caly, Michelle Sait, Susan A Ballard, Kristy Horan, Mark B Schultz, et al. Tracking the covid-19 pandemic in australia using genomics. *Nature communications*, 11(1):1–9, 2020.

[48] Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494, 2017.

[49] Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). *Proceedings of the National Academy of Sciences*, 110(1):228–233, 2013.

[50] Silvia Stringhini, Ania Wisniak, Giovanni Piumatti, Andrew S Azman, Stephen A Lauer, Hélène Baysson, David De Ridder, Dusan Petrovic, Stephanie Schrempft, Kailing Marcus, et al. Seroprevalence of anti-sars-cov-2 igg antibodies in geneva, switzerland (serocov-pop): a population-based study. *The Lancet*, 396(10247):313–319, 2020.

[51] Timothy G Vaughan, Gabriel E Leventhal, David A Rasmussen, Alexei J Drummond, David Welch, and Tanja Stadler. Estimating epidemic incidence and prevalence from genomic data. *Molecular biology and evolution*, 36(8):1804–1816, 2019.

[52] Stephen White, Ian McAllister, and Valentina Feklyunina. Belarus, ukraine and russia: East or west? *The British journal of politics and international relations*, 12(3):344–367, 2010.

[53] Sean L Wu, Andrew N Mertens, Yoshika S Crider, Anna Nguyen, Nolan N Pokpongkiat, Stephanie Djajadi, Anmol Seth, Michelle S Hsiang, John M Colford, Art Reingold, et al. Substantial underestimation of sars-cov-2 infection in the united states. *Nature communications*, 11(1):1–10, 2020.