5-14-2021

Genomic Epidemiology Analysis of SARS-CoV-2 Transmission Dynamics in Belarus

Ayotomiwa Ezekiel Adeniyi
Georgia State University, tomiwakiels@gmail.com

Genomic Epidemiology Analysis of SARS-CoV-2 Transmission Dynamics in Belarus

by

Ayotomiwa Ezekiel Adeniyi

Under the Direction of Pavel Skums, PhD

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2021

ABSTRACT

The recent outbreak of the COVID-19 virus has brought about the need for the implementation of Non-Pharmaceutical Interventions (NPIs) by governments and public health departments to fight the spread of the virus. Belarus, a small country in the eastern part of Europe implemented a small amount of NPI protocols to control the spread of the virus.

This thesis focuses on the usage of genomic epidemiological methods to analyze the origin and spread of SARS-CoV-2 in Belarus. Results showed that the number of case introductions and exports to other countries are indeed substantial. It was detected that the Belarussian SARS-CoV-2 genetic diversity originated from at least eighteen separate introductions from different geographical regions.

INDEX WORDS: COVID-19, Non-Pharmaceutical Interventions, Genomic Epidemiology, SARS-CoV-2, Belarus, Virus

Genomic Epidemiology of SARS-CoV-2 Transmission Dynamics in Belarus

by

Adeniyi Ezekiel Ayotomiwa

Committee Chair: Pavel Skums

Committee: Alex Zelikovskiy

David Campo

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

May 2021

DEDICATION

I would like to dedicate this page to my parents Mr and Mrs Adeniyi for their unrelenting support and for motivating me throughout the years of my academic pursuit.

ACKNOWLEDGEMENTS

I would like to acknowledge and thank my research advisor Dr. Pavel Skums for accepting me to be a part of his group and for his patience in guiding me through my thesis project.

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

The republic of Belarus, a country in eastern Europe with an approximate population of 9.5million when compared with surrounding neighboring countries has weaker socio-economic and political ties (Dastanka, 2015). Movements into and outside of the country is very low due to the need to obtain visa permits to travel to the European Union, and other countries in the world. However, the healthcare system and the country's Human Development Index (HDI) when compared to its neighboring countries is very high (United Nations, 2020).

Figure 1.0: Belarus when compared to its neighboring countries has weaker socio-economic and political ties (Dastanka, 2015)

The first confirmed case of the SARS-CoV-2 disease in Belarus occurred on February 28, 2020 from a person who arrived the country from Iran (Ministry of Health Belarus, 2020). There have been more than 300,000 confirmed cases by March 13, 2021. Health Officials in Belarus instituted liberal NPI methods when compared with other vhHDI countries (Aslund, 2020). One NPI method imposed by other countries involved imposing a mandatory 14days self-isolation

protocol for individuals arriving from countries with a high incidence of the virus outbreak or who have been identified as close contacts of individuals with the virus. Other common NPI methods involved practicing social distancing measures such as staying six feet apart in public places, avoiding large crowds, wearing masks in public places and enforcing remote learning and education. (World Health Organization, 2021)

In Belarus, no large-scale quarantine measures, lockdown or social distancing was practiced. Thus, making the country of Belarus open while the world battled with controlling the spread of the virus. Given the uniqueness of the experience that the people of Belarus had, understanding the epidemiological dynamics of the COVID-19 pandemic in Belarus could give a better insight into the present public health situation in the country. It also provides a better insight into assessing the impact of different NPI strategies around the globe.

Nevertheless, it is challenging to perform epidemiological analysis due to the limited amount of available data as it was not until the last quarter of 2020 that the official country level counts that included daily case incidence was released.

In this work, we combined Whole Genome Sequencing (WGS) genomic data and epidemiological data into carrying out the study of SARS-CoV-2 transmission dynamics in Belarus. The absence of a large number of reliable epidemiological data, integrated genomic and epidemiological analysis allowed us to provide a plausible scenario for the emergence and spread of SARS-CoV-2 in the country of Belarus. The obtained results also gave an insight into the effect of limited NPIs implemented during the first wave of the coronavirus epidemic.

2 CHAPTER 2

2.1 Data Collection

The Belarussian SARS-CoV-2 genomic data was downloaded from the publicly available repository of GISAID (Yuelong Shu and John McCauley, 2017) on March 15, 2021. This data contained 41 full length genomes of samples collated from March 2020 to February 2021.

2.2 Global Phylogenetic analysis

For the phylogeny reconstruction, we used the SARS-CoV-2 specific phylogenetic inference pipeline as implemented on Nextstrain (Hadfield et al 2018). The available sequences from Belarus were analyzed together with 12,064 background sequences from the global SARS-CoV-2 population. In order to obtain a representative sample with background sequences, a country specific built in Nextstrain context subsampling was used (Hadfield et al, 2018). The sequences were aligned using MAFFT (Katoh and Standley, 2013) and a maximum likelihood (ML) phylogenetic tree was constructed using IQ-TREE (Nguyen et al, 2015) under Hasegawa-Kishino-Yano (HKY) + Γ nucleotide substitution model with a gamma distributed site rate variation. (Masami et al, 1985)

In the resulting time labelled tree, the ancestral geolocation traits were inferred using a so-called ‘mugration model’ (Sagulenko et al 2018). In this model, the countries of origin of the tree nodes were considered as discrete traits and the spread of the virus between countries considered as a general time reversible process. The human mobility statistics data provided by the European Commission Knowledge Center on Migration and Demography (KCMD) (Ettore et al, 2019) was used to estimate the transition rates between traits. Although global travels were

affected by COVID-19 related restrictions, the data is assumed to be representative of the relative density of human mobility between countries in lockdown-quarantine.

The transition states were assumed to be proportional to the normalized average numbers of inter-country trips. The resulting transition rate matrix was used to estimate the maximum joint likelihood internal node traits using dynamic programming algorithm (Tal Pupko et al 2000). The Belarussian clades were then defined as those having their most recent common ancestors (MRCA) with Belarus trait and intra-Belarussian clusters were inferred as the maximal subsets of Belarussian sequences inside these clades.

2.3 Intra-Country Phylodynamic Analysis

We followed a general analytic pipeline that has been adopted by similar country level studies (for example Jemma L Geoghegan et al, 2020 and Alessia Lai et al, 2020) with modifications tailored towards specifics applicable to the Belarussian data. We evaluated the temporal signal by constructing a 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) (Andrew Rambaut et al, 2016).

For the next step we used BEAST (v.2.6.3) software (Remco Bouckaert et al, 2014) to fit a Coalescent Bayesian Skyline Model to the full set of Belarussian 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 model with a prior mean equal to $8*10^{-4}$ mutation site/year and the standard deviation of $5*10^{-4}$(Kristian G Andersen et al 2020 and Jemma L Geoghegan et al 2020) where the distribution density was parameterized using the corresponding shape and rate parameters. Four segments were assumed for the effective

population size that corresponds to the growth and decline periods of the first and second Covid-19 epidemic waves.

Next, we used Markov Chain Monte Carlo (MCMC) method to sample the posterior distribution according to our model parameters. We used $3*10^7$ MCMC iterations sampling every $3*10^3$ iterations and an initial 10% 'burn-in' for the iterations. The MCMC sampling quality was assessed using Tracer (v.1.7.1) (A. J. Andrew Rambaut et al 2018) and was accepted provided that all parameters have an effective sampling size greater than 300. The obtained maximum clade credibility (MCC) tree was annotated using TreeAnnotator (v.1.8.4) (Bouckaert and Heled, 2013). We confirmed the reliability of the intra Belarussian clusters by verifying that they correspond to the monophyletic clades in the MCC tree. Thereafter we estimated the times to the most recent common ancestor of each clusters.

The two largest intra-Belarussian transmission lineages were estimated for their effective reproductive number R_e using Birth Death Skyline Serial (BDSKY) model (Tanja Stadler et al 2013) implemented in BEAST. These lineages appeared to cover multiple geographical regions and are most probably spread in parallel the same susceptible population. We then used a linked phylogenetic model where both lineages evolve and are sampled independently but share the same substitution model parameters, molecular clock rate and effective reproduction number drawn from the same respective priors.

Given the relative sparsity of available genomic data, the approach allowed us to use a larger and more representative combined sample for analysis. The sample proportions were assumed to have a Beta (α,β) distribution prior with parameters α=1 and β= $9.99 *10^5$ which reflects the sparsity of the Belarussian data sampling and the possible relation between the magnitude of the officially reported number of cases and the number of sampled sequences.

Finally, to compensate for the data sparsity, the model was equipped with informative priors on several parameters. Specifically, the prior for the origin of each cluster was assumed to be log-normally distributed with $\sigma = 0.1$ and μ selected in such a way that the mean time interval before the latest sequence sampling time equals the TMRCA estimated using coalescent Bayesian skyline.

3 CHAPTER 3

3.1 Results

Despite the sparsity of our sample as mentioned earlier, the observed genomic diversity of SARS-CoV-2 sequences from Belarus was relatively high with 13 lineages according to the genomic nomenclature (Andrew Rambaut et al, 2020) observed (Fig.2.Panel A). In particular, the genome that was sampled on February 23, 2021 belongs to B.1.1.7 lineage that emerged from the United Kingdom in November, 2020 and has since been spreading towards fixation (Nicholas G Davies et al 2021). The root-to-tip regression analysis of the available sequences demonstrated moderately strong temporal signal ($R^2 = 0.56, p < 10^{-6,}$ *Fig 2.Panel B)*

From analysis, we identified 18 distinct intra-Belarussian clades that most likely corresponds to separate introductions of the virus into the country. The inference about the origin of SARS-CoV-2 into countries is usually complicated since it was a global pandemic and close genomic variants can be observed in multiple geographic locations. Therefore, results from such inferences should be treated with caution.

Nevertheless, the phylogenetic analysis of the Belarussian Genomic sequences provides a reasonable and consistent transmission history which agrees with available epidemiological data records. For example, the first detected and confirmed SARS-CoV-2 case was from an individual who arrived from Iran (Ministry of Health Belarus, 2020) and the phylogenetic tree analysis results summarized in Figure 2 reaffirms this position. This agreement also holds for the second detected case of SARS-CoV-2 in Belarus from an individual who travelled from Italy to Belarus (Ministry of Health Belaurs, 2020). The first introduction produced at least one secondary case as confirmed by the phylogenetic tree. Both detected lineages were however not sampled until after March, 2020 (fig 2. Panel B). This can be attributed to the timely isolation of the infected

individuals and their first order contacts (Belarus, Ministry of Health 2020). Insufficient sampling could also be a possibility. In general, SARS-CoV-2 importations into the country can be attributed to a mixture of regional and global transmissions.

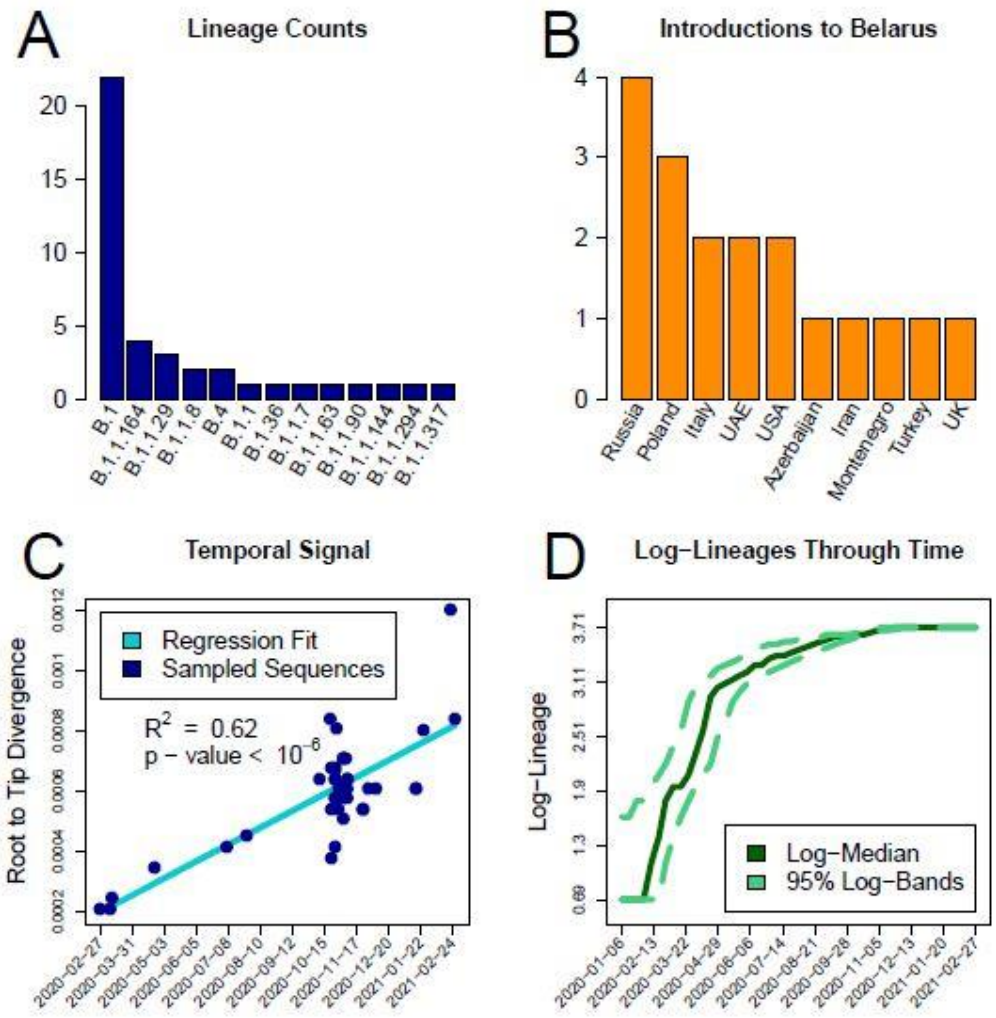

Figure 2: The summaries of the available genomic samples include: A) the counts of each lineage, B) estimated introduction sources to Belarus, C) temporal change in distance to the tree root, and D) lineages through time in logarithmic scale

As illustrated in figure 2, the most frequent viral introduction sources from the model were from two neighboring countries Russia (4 introductions) and Poland (3 introductions) followed by Italy, USA and UAE (2 introductions each). The hypothesized relationships between

sampling regions within Belarus, detected clusters and introductory routes that have been identified from analysis are summarized in figure 3.

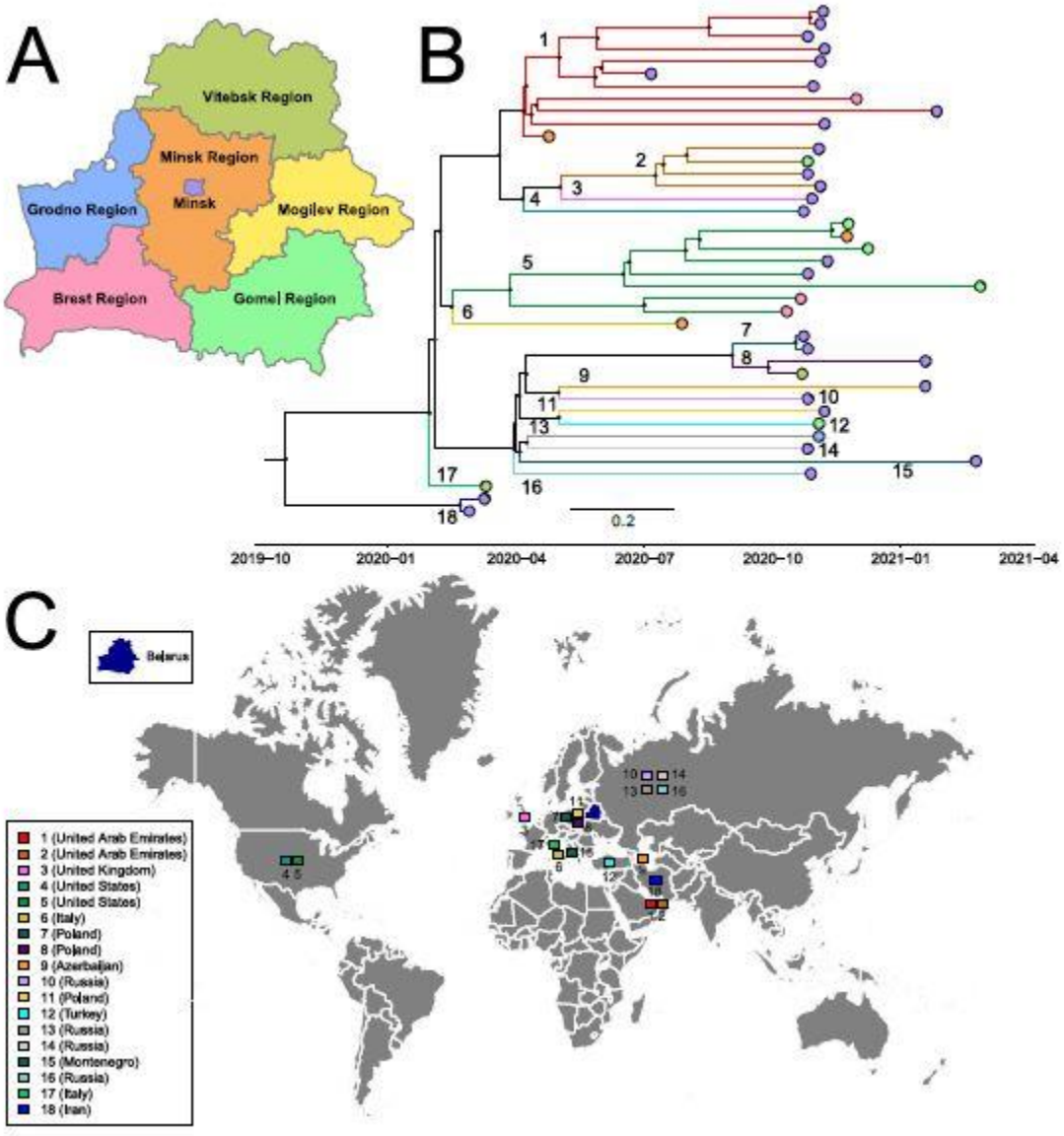

Figure 3: The summaries of the analyzed genomic sequences: A) Color-coded administrative regions of Belarus, B) genetic tree color-coded according to both: sampling by administrative regions (tree leaves or circles) and introductory clusters (lines or tree branches), and C) introductory clusters marked on the world map created by Vemaps.com

Five SARS-CoV-2 introductions (31.25%) were associated with clusters with two or more sequences and thus hypothesized to lead to ongoing intra-country transmission and resulting emergence of transmission lineages. The three largest detected clusters are paraphyletic which indicates the virus export from Belarus to other countries following its initial introduction to Belarus. Although such paraphylies could be pure sampling artefacts, the presence of multiple such paraphylies can serve as an indicator of multiple virus exports.

The transmission clusters are well-mixed with all of them having representatives from at least two Belarussian administrative regions illustrated in Figure 3 Panel B. The relative homogeneity of the Belarussian demographic suggest that the transmission lineages circulated over the same susceptible population. It is also observed that the majority of the estimated corresponding branching events belong to the same time period while the smaller number of branching events over the second half of 2020 can be attributed to epidemiological reasons or sparse sampling. This implies that despite sequencing being performed mostly in the fall of 2020 and in early 2021, the phylodynamic analysis of currently available Belarussian genomes allows us to reliably assess only the first wave of the epidemic which took place before July 2020. The mean effective reproductive number estimate was $R_e = 1.59$ with a corresponding 95% highest posterior density (HPD) interval of 1.32 -1.91.

The Kolmogorov-Smirnov test was used for the formal comparison of both prior and posterior distribution samples presented in figure 3a. The obtained estimate us comparable with phylodynamics assessments of R_e estimated from developed countries before the introduction of travel restrictions and physical distancing measures (Torsten Seemann et al 2020, Danielle Miller et al, 2020, Sarah Nadeau et al 2021 and Gonche Danesh et al, 2020).

To further expand this analysis, we matched early effective reproductive numbers of Belarus against Ukraine, a neighboring ex-soviet country with similar demographic and comparable density of SARS-CoV-2 genome sampling. With Ukraine's implementation of stricter lockdown and physical distancing measures being the main difference between the two countries (Aslund, 2020). The same Birth-Death skyline serial model was used on two large Ukrainian clusters of 28 sequences. The reproductive number of the Ukrainian sequences was estimated as $R_e = 1.66$ *(95% HPD interval: 1.50 -1.84;* $p < 10^{-10}$, Kolmogorov-Smirnov test for prior and posterior distributions). These numbers agree with previous estimations based on Exponential Coalescent model (Yuriy Gankin et al 2021) and appeared to be very close to the R_e assessment for Belarus. This was not entirely surprising as it is more reflective of the impact of public health intervention measures in Ukraine where extensive violations of lockdown measures and inability of the government to enforce these measures have been widely reported.

4 CONCLUSION

In this thesis, we studied the SARS-CoV-2 epidemic in Belarus using officially reported incidence data, testing data and genomic data collected between March 2020 and February 2020. Reported findings provide new insights into insufficiently understood COVID-19 transmission dynamics and effects of NPIs in the country. This thesis also sheds light on several key epidemiological issues that Belarus shares with other countries around the world.

First, the analysis revealed the history of multiple transmissions of SARS-CoV-2 into and from the country. The analysis also identified several intra-country transmission lineages. We identified 18 introductions within 13 genomic lineages but this estimate is lower bound on the real number of introductions because only a small fraction of the overall SARS-CoV-2 genomic diversity was sampled. A significant portion of the estimated transmission links were with geographic neighbors. Links from USA transmission was expected given the relatively large number of Belarussians in diaspora living in the USA and the uninterrupted travels of the people with dual citizenships.

It is however important to highlight the limitations of this study. A major limitation of this study is the scarcity of currently available genomic data especially when compared with other European countries. The thesis approach compensates for this by utilizing informative priors for phylogenetic and phylodynamics inference as well as linked models. In addition to this, the Birth-Death Skyline model that was used for the phylodynamics analysis have been confirmed to be sensitive to smaller genomic datasets. We hope that this thesis serves as a trigger for increased SARS-CoV-2 genomic epidemiology studies in Belarus and also a way to encourage the development and funding of sequencing facilities for molecular surveillance of SARS-CoV-2 in the country.

WORKS CITED

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 mdical virology* (2020): 675-679.

Andrew Rambaut, Alexei J Drummond, Dong Xie, Guy Baele and Marc A Suchard. "Posterior summarization in bayesian phylogenetics using tracer 1.7." *Systematic Biology* (2018): 901.

Andrew Rambaut, Edward C Holmes, Aine O'Toole, Verity Hill, John T McCrone, Christopher ruis, Louis du Pleiss and Oliver G Pybus. "Exploring the temporal structure of heterchronous sequences using tempest (formerly path-o-gen)." *Virus evolution* (2016).

Andrew Rambaut, Edward C Holmes, Aine O'Toole, Verity Hill, John T McCrone, Christopher Ruis, Louis du Plessis and Oliver G Pybus. "A dynamic nomeclature proposal for sars-cov-2 lineages to assist genomic epidemiology." *Nature Microbiology* (2020): 1403-1407.

Aslund, Anders. "Responses to the COVID-19 crisis in Russia, Ukraine, and Belarus." *Eurasian Geography and Economics* (2020): 1-14.

Belarus, Ministry of Health of Republic of Belarus: Six new cases have been confirmed in. *https://mizdrav.gov.by/ru/sobytiya/shest-sluchaev-koronavirusa-podtverzhdeno-v-belarusi*. 2020. 2021 03 2021.

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

coronavirus, Ministry of Health of Republic of Belarus: Belarussian woman from Vitebsk has been tested positive for. *https://minzdrav.gov.by/ru/sobytiya/mizdrav-respuliki-belarus-informiruet/*. 2020. 31 03 2021.

Danielle Miller, Michael 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 srars-cov-2 spread into and within israel." *Nature communications* (2020): 1-10.

Dastanka, Alena A. "Multilateralism in foreign policy of belarus: European and euarasian dimension." *Regional Formation and Development Studies* (2015): 16-23.

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* (2019): 30.

Gonche Danesh, Baptiste Elie, Yannis Michalakis, Mircea T Sofonea, Antonin Bal, Sylvie Behillil, Gregory Destras, David Boutolleau, Sonia Burrel, Anne-Genevieve Marcelin et al. "Early phylodynamics analysis of the covid-19 epidemic in france." *medRxiv* (2020).

Hadfield, James, et al. "Nextstrain: real-time tracking of pathogen evolution." *Bioinformatics* (2018).

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 transmissin patterns and dynamics of sars-cov-2 in aotearoa new zealand." *Nature communications* (2020): 1-7.

Katoh, Kazutaka and Daron Standley. "Mafft multiple sequence alignment software version 7: improvements in performance and usability." *Molecular biology and evolution* (2013): 772-780.

Kristian G Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes and Robert F Garry. "The Proximal origin of sars-cov-2." *Nature medicine* (2020): 450-452.

Masami Hasegawa, Hirohisa Kishino and Taka-aki Yano. "Dating of the human-ape splitting by a molecular clock of mitochondiral dna ." *Journal of molecular evolution* (1985): 160-174.

Nations, United. "Human Development Report 2020." 2020. *http://hdr.undp.org/en.* 31 03 2021.

Nguyen, Lam-Tung, et al. "Iq-tree a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies." *Molecular biology and evolution* (2015): 268-274.

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 ." *Science* (2021).

Pavel Sagulenko, Vadim Puller and Richard Neher. "Treetime: Maximum-likelihood phylodynamic analysis." *Virus evolution* (2018).

Remco Bouckaert, Joseph Heled, Denise Kuhnert, Tim Vaugha, Chieh-Hsi Wu, Dong Xie, Marc Suchard, Andrew Rambaut and Alexei J Drummond. "Beast 2: a software platform for bayesian evolutionary analysis." *PLoS Comput Biol* (2014).

Sarah Nadeau, Timothy Vaughan, Jeremie Scire, Jana Huisman and Tanja Stadler. "The origin and early spread of sars-cov-2 in europe." *Proceedings of the National Academy of Sciences* (2021): 118.

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

Tanja Stadler, Denise Kuhnert, Sebastian Bonhoeffer and Alexei J Drummond. "Birth-death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv)." *Proceeding of the National Academy of Sciences* (2013): 228-233.

Torsten Seemann, Courtney R Lane, Norelle L Sherry, Sebastian Duchene, Anders Goncalves da Silva, Leon Caly, Michelle Sait, Susan Ballard, Kristy Horan, Mark Schultz et al. "Tracking the covid-19 pandemic in austalia using genomics ." *Nature communications* (2020): 1-9.

World Health Organization. *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. 31 03 2021.

Yuelong Shu and John McCauley. "GISAD: Global initiative on sharing all influenza data - from vision to reality." *Eurosurveillance* (2017).

Yuriy Gankin, Vladimir Koniukhovskii, Alina Nemira, Gerardo Chowell, Thomas Weppelmann, Pavel Skums and Alexander Kirpich. "Investigating the first wave of the covid-19 pandemic in ukraine using epidemiological and genomic sequencing data." *medRxviv* (2021).