Reconstructing Viral Evolutionary and Transmission Histories

Reconstructing Viral Evolutionary and Transmission Histories

by

Alina Nemira

Under the Direction of Committee Chair's Alex Zelikovsky, Ph.D.

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

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2025

ABSTRACT

The COVID-19 pandemic has posed unprecedented challenges to global public health, sparking intense investigation into its transmission dynamics, control strategies, and viral evolution. This work explores the epidemiological, genomic, and evolutionary aspects of SARS-CoV-2 and Hepatitis C Virus (HCV), shedding light on the role of non-pharmaceutical interventions (NPIs), viral spread, and intra-host viral diversity.

Phylogenetic and phylodynamic analyses were used to study the early course of COVID-19 in Ukraine. Results show that even short delays in implementing NPIs could have significantly increased case numbers and overwhelmed healthcare systems. Genomic analysis revealed multiple independent introductions of SARS-CoV-2 into the country, with evidence of sustained local transmission. Estimates of the basic reproduction number from both case data and genomic sequences support these conclusions and suggest that travel restrictions had limited impact on virus importation.

In Belarus, where NPIs were minimal, genomic surveillance and phylodynamic modeling were employed to understand SARS-CoV-2 dynamics. The cumulative number of infections by mid-2020 was likely underreported by a factor of approximately four. At least 18 separate introductions of the virus were detected, followed by considerable regional transmission. Although limited interventions modestly reduced the effective reproduction number, its trajectory remained comparable to neighboring regions with stricter measures, highlighting the influence of sociodemographic and regional factors in shaping epidemic outcomes.

To investigate intra-host viral evolution, epistatic interactions among single-nucleotide variants (SNVs) in Hepatitis C Virus populations were analyzed. A novel metric, epistatic density, was introduced to quantify co-occurrence of positively linked SNV pairs within viral genomes. Analysis of populations from acute and chronic HCV infections revealed higher and more structured epistatic density in chronic cases, indicating differential selective pressures over the course of infection. Epistatic density features also enabled successful classification of viral populations by infection stage.

INDEX WORDS: Infectious disease, Virus, Phylogenetics, Phylodynamics, Genomic epidemiology

Reconstructing Viral Evolutionary and Transmission Histories

by

Alina Nemira

Committee Chair: Alex Zelikovsky

Committee: Alexander Kirpich

Murray Patterson

Pavel Skums

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2025

DEDICATION

To my precious mother, whose kindness and love know no bounds.

To my father, who taught me to be strong in my life no matter what.

To my grandfather, with whom my adventures never end.

To my uncle, whose wonderful sense of humor brightens any day.

And to my grandmother on my mother's side, whose love and spirit continue to guide me.

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Pavel Skums and Dr.Alex Zelikovsky, my scientific advisors, for providing me with careful guidance and support throughout my Ph.D. journey. Their mentorship has not only helped me grow as a professional but also as a collaborator and an individual. I would like to extend my gratitude to Dr.Murray Patterson for his invaluable advises and guidance.

I had the pleasure of collaborating with my labmates, and I would like to express my gratitude to Dr. Zahra Tayebi, Dr.Roya Hosseini, Dr.Mark Grinshpon, Dr.Sergey Knyazev, Dr.Daniel Novikov, Dr.Kiril Kuzmin, Dr. Akshay Juyal, Dr. Seth Sims, Dr. Bikram Sahoo, Ryan Alomair, Hafsa Farooq, Aytomiwa Ezekiel Adeniyi, Maryam Kafikang, Okene Ekene.

I would like to extend my gratitude to my colleagues from GSU Public Health department, Dr.Alexander Kirpich and Dr. Aleksandr Shishkin, for their collaboration and participation in our bioinformatic projects.

Finally, I would like to express my appreciation to all the friends I made in the GSU Computer Science department and UCLA Computational Medicine department, who have accompanied me throughout this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Motivation

Throughout human history, numerous pandemics and outbreaks have significantly impacted populations worldwide. These include the Plague of Justinian, Antonine Plague, Black Death, Cocoliztli epidemic, 1520 Mexico smallpox epidemic, Third Plague pandemic, 1918-1922 Russia typhus epidemic, Spanish flu, 1957-1958 influenza pandemic, Hong Kong flu, and the ongoing HIV/AIDS and COVID-19 pandemics. These pandemics have resulted in millions of deaths globally.

Infectious diseases in humans, caused by viruses and bacteria, have profound impacts on public health, economies, and societies worldwide. Understanding their evolutionary and transmission history is crucial for developing effective strategies to combat these pathogens. Advancements in molecular biology and sequencing have revolutionized the investigation of outbreaks and pandemics. These technologies enable rapid identification and genetic characterization of pathogens, tracking their evolution and spread in real-time. Utilizing genomic data enhances our understanding of phylogenetics and transmission dynamics, informs public health responses, and aids in developing targeted treatments and vaccines.

Bioinformatics and genomic epidemiology have played pivotal roles in the global response to COVID-19. The sequencing of SARS-CoV-2 genomes has enabled researchers to track the virus's evolution in real-time, identify mutations associated with increased transmissibility or immune escape, and monitor the spread of different variants across regions. These insights have informed public health decisions, such as travel restrictions, lockdown measures, and vaccine distribution strategies. Moreover, genomic data have facilitated the development of diagnostics, therapeutics, and vaccines, demonstrating the power of integrating genomic information into public health practice.

Despite these advancements, significant gaps remain in our understanding of the factors that drive the emergence and spread of infectious diseases like COVID-19. This research aims to address these gaps by investigating the evolutionary history and dynamics of SARS-CoV-2 and other pathogens. By leveraging bioinformatics tools and genomic data, this study seeks to uncover the mechanisms underlying pathogen evolution, transmission patterns, and the impact of public health interventions.

1.2 Problem Formulations

1. Investigating early stage of the COVID-19 pandemic in Ukraine using epidemiological and genomic data

   Given: a GISAID dataset of SARS-CoV-2 sequences from patients in Ukraine, sampled between April 24 - August 7, 2020.

   Find: estimations of the basic reproduction number R_0 of the epidemic, identify transmission clusters, and track possible introductions of the virus into the country.

2. SARS-CoV-2 transmission dynamics in Belarus in 2020 revealed by genomic and incidence data analysis

   Given: a GISAID dataset of SARS-CoV-2 sequences from patients in Belarus, sampled between 2018 and May 2022.

Find: estimations the effective reproduction number R_e and the rate of becoming non-infectious δ, and compare the epidemiological dynamics of Belarus with those of other countries to assess the impact of different non-pharmaceutical interventions.

3. Epistatic density of viral variants in acute and chronic HCV patients

   Given: a sequence dataset from HCV-infected patients with annotated infection type (acute vs. chronic)

   Find: differences between sequences of HCV patients with acute and chronic infection.

1.3 Contributions

1. Contributions to Chapter 1

   - collected and integrated COVID-19 incidence data and high-quality SARS-CoV-2 genomic sequences from Ukraine

   - constructed and analyzed phylogenetic trees to investigate global and Ukrainian SARS-CoV-2 lineages

   - identified transmission clusters within Ukraine and estimated their times to the most recent common ancestors (TMRCAs)

2. Contributions to Chapter 2

   - estimated the effective reproduction number R_e and the rate of becoming non-infectious δ for SARS-CoV-2 transmission in Belarus

- inferred the number and sources of viral introductions into Belarus using phylogenetic and phylodynamic analyses

- compared the epidemiological dynamics of Belarus with those of other countries to evaluate the impact of different non-pharmaceutical intervention strategies

3. Contributions to Chapter 3

- implemented a novel method to calculate linked SNVs and constructed coordinated substitution networks using pipelines

- calculated epistatic density values and its statistics, and analyzed epistatic density distributions for acute and chronic HCV populations

- developed three methods for acute vs chronic classification and estimated performance of the classification model

1.4 Publications and Presentations

Journals:

1. Gankin Y, Nemira A, Koniukhovskii V, Chowell G, Weppelmann TA, Skums P, Kirpich A. Investigating the first stage of the COVID-19 pandemic in Ukraine using epidemiological and genomic data. Infect Genet Evol. 2021 Nov;95:105087. doi: 10.1016/j.meegid.2021.105087. Epub 2021 Sep 27. PMID: 34592415; PMCID: PMC8474758.

I have participated in collecting incidence and genomic data, built phylogenetic trees, analyzed global and Ukrainian trees, identified transmission clusters and their TMRCA, visualized results, and wrote and published the paper.

2. Nemira A, Adeniyi AE, Gasich EL, Bulda KY, Valentovich LN, Krasko AG, Glebova O, Kirpich A, Skums P. SARS-CoV-2 transmission dynamics in Belarus in 2020 revealed by genomic and incidence data analysis. Commun Med (Lond). 2021 Sep 16;1:31. doi: 10.1038/s43856-021-00031-1. PMID: 35602211; PMCID: PMC9053244.

I have participated in phylogenetic analysis, analyzed genomic data, and wrote the paper.

3. Skums P, Mohebbi F, Tsyvina V, Baykal PI, Nemira A, Ramachandran S, Khudyakov Y. SOPHIE: Viral outbreak investigation and transmission history reconstruction in a joint phylogenetic and network theory framework. Cell Syst. 2022 Oct 19;13(10):844-856.e4. doi: 10.1016/j.cels.2022.07.005. PMID: 36265470; PMCID: PMC9590096.

I have participated in performed a phylogenetic analysis, analyzed genomic data, and wrote the paper.

Poster:

4. "Constructing an Epistatic Network from Hepatitis C Protein Sequence Data using Algorithmic Approach" was presented at

Georgia State University Scientific Computing Day (SCD) 2023

Presentations:

5. Alina Nemira, A.E. Adeniyi, O. Glebova, A. Kirpich, P. Skums, L.N. Valentovich, E.L. Gasich, K.Y. Bulda, and A.G. Krasko, The tale of two Eastern European countries: genomic epidemiology analysis of SARS-CoV-2 transmission dynamics in Belarus and Ukraine

EPIDEMICS 8, 8th International Conference on Infection Disease Dynamics 2021

CHAPTER 2

INVESTIGATING THE FIRST STAGE OF THE COVID-19 PANDEMIC IN UKRAINE USING EPIDEMIOLOGICAL AND GENOMIC DATA

Published in *Infection, Genetics and Evolution* 95, 2021

2.1 Introduction

SARS-CoV-2 virus causing COVID-19 was first detected in December 2019 in the Chinese city of Wuhan [112][74][62][34][36], and has rapidly spread around the globe, prompting the World Health Organization (WHO) to declare a pandemic in March, 2020 [10], just three month after the first reported case. Despite having much lower case-fatality rate than other recent coronavirus pandemics such as the severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS), the novel coronavirus has claimed more lives just within a few months of introduction than both of those epidemics combined [75]. As of June 29, 2021 there were more than 182 million infections with over 3.9 million deaths [20]. In the absence of vaccines during the early pandemic period, non-pharmaceutical interventions (specifically, non-pharmaceutical epidemic mitigation interventions) were the only tools at the disposal of public health authorities to prevent and to mitigate the virus spread [35][33][79]. The strategies implemented and enforced by governments around the world were highly variable and included frequent sanitation of public spaces, enforced social distancing, wearing of masks, closure of schools, churches, and ban of mass gatherings [63][11][3].

Even well after a year since the epidemic started, fundamental questions regarding the effects of non-pharmaceutical interventions (NPIs) [41][66] and the genomic evolution of SARS-CoV-2 [101][111][72][31] during the introductory period remain. Additionally, recent modeling efforts aimed at shedding light on those questions have mostly focused on China [112][74][62][34], the rest of Asia [105][110][91], Western and Central Europe [93][110][47], and North America [53][45][96][46][36], largely neglecting Africa, the Middle East, and Eastern Europe.

In Eastern Europe, post-socialist economics and healthcare systems are inherently different from Western Europe. The available SARS-CoV-2 transmission models for Eastern Europe are based on relatively simple SIR or similar compartmental models [64][82][69] where individuals are assigned to groups and all individuals within a given group are expected to have the same characteristics. To the best of our knowledge no agent-based modeling studies have been conducted for Ukraine to evaluate the impact of spatial heterogeneities in key transmission drivers such as density of infected individuals and their geographic locations. Furthermore, the number of genomic epidemiology studies on the COVID-19 pandemic in Eastern Europe has been limited. In this paper, we sought to fill the knowledge gap for the Ukrainian epidemic [30], which provides a unique setting for studying the COVID-19 spread under the ex-USSR healthcare system, and with the epidemic mitigation policies similar to the rest of Europe.

The first confirmed case in Ukraine was reported on March 3, 2020 and was an individual who has recently traveled from Italy. The first death was reported on March 13, 2020 [5][30]. The Ukrainian government started to implement quarantine measure on March 12, 2020 [30][69] while the cases continued to rise possibly because of the delayed detections of existing infections and returns of infected Ukrainians from abroad [69] (Figure 3B). As a result, more strict measures have been implemented on April 6, 2020 which included the closure of schools, universities, shopping malls, and mandatory mask regiment in public places [69]. Those measures were slightly softened on April 24, 2020 and many services resumed even though some restrictions lasted till the end of June 2020 [30]. As a summary, Ukrainian officials took the epidemic very seriously from the beginning and started to implement the mitigation efforts and corresponding regulations almost immediately after multiple cases in the country have been detected. At the same time implementation of the proposed mitigation efforts did vary from region to region, and so did the compliance with those regulations [9][14].

2.2 Methods

2.2.1 Genomic Epidemiology Analysis

Sixty high-quality SARS-CoV-2 genomes from Ukraine sampled between April 24, 2020 and August 7, 2020 were extracted from GISAID [4]. These genomes were utilized to construct a maximum likelihood phylogeny using Nextstrain build for SARS-CoV-2 with the default country-specific subsampling settings [19]. The obtained timed phylogeny contained Ukrainian sequences together with a representative subsample of 6479 sequences from other geographic regions, and included inferred ancestral geographic traits of internal nodes. Using these traits, intra-country transmission clusters were identified as clades with the most recent common ancestors (MRCA) estimated as originating from Ukraine. For each cluster, confidence intervals for emergence times for MRCA and its parent were also obtained.

Next, a phylodynamic analysis of the three largest clusters and the entire Ukrainian

SARS-CoV-2 population was performed using BEAST v1.10.4 [100]. We used a strict molecular clock, HKY+Γ nucleotide substitution model, a tree prior with exponential growth coalescent. Priors for the parameters were defined in BEAUti v 1.10.4 and were the following: a) normal $\mathcal{N}(mean = 8.0\text{e-}4,\ st.dev = 2.0\text{e-}5)$ for the clock rate, b) log-normal $\mathcal{LN}(mean = 1.0,\ st.dev = 1.25)$ for the population size, c) double exponential (Laplace) distribution $\mathcal{DEXP}(\mu = 0,\ b = 100)$ for the growth rate, d) normal $\mathcal{N}(mean = 0,\ st.dev = 1)$ for the freqParameter, e) exponential $\mathcal{EXP}(mean = 0.5,\ offset = 0)$ for the gammaShape parameter, and f) log-normal $\mathcal{LN}(mean = 1.0,\ st.dev = 1.25)$ for the kappa parameter. The detailed parameters file is available in XML format at `https://github.com/alanira/COVID19-Ukraine-phylodynamics`. The parameters were estimated after 30,000,000 iterations of Markov Chain Monte Carlo (MCMC) sampling, with the initial 10% values discarded as burn-in. The results were accepted if the effective sample sizes were above 200 for all parameters. The estimated exponential growth rates were used to calculate the basic reproduction numbers $\mathcal{R}_0$ under the assumption that SARS-CoV-2 generation intervals (i.e. times between infection onset and onward transmission) were gamma-distributed [73]. We used the formula

$$\hat{\mathcal{R}}_0 = \left[1 + \frac{\hat{f}\hat{\sigma}^2}{\hat{\mu}}\right]^{\frac{\hat{\mu}^2}{\hat{\sigma}^2}}, \qquad (2.1)$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the mean and standard deviation of the aforementioned gamma distribution [1][2][104][55]. For these values, we used the estimates $\hat{\mu} = 5.20$ and $\hat{\sigma} = 1.72$ from [50] and $\hat{\mu} = 3.95$ and $\hat{\sigma} = 1.51$ from [50]. The formula (2.1) defines a strictly monotone transformation of $\hat{f}$, and, therefore, it also straightforwardly transforms the 95% highest posterior density (HPD) intervals for f into those for $\mathcal{R}_0$.

2.3 Results

Despite a sparse sampling, the observed genomic diversity of SARS-CoV-2 in Ukraine is substantial, indicating both multiple introductions of the virus and sustained intra-country evolution (Figure 2). This agrees well with the patterns observed in other countries [52], and emphasizes the contribution of global movement of people to the rapid spread of SARS-CoV-2. Specifically, Ukrainian sequences are distributed among eight lineages by the classification of [87] as follows: $B.1$ - 50.0 % of genomes, $B.1.1$ - 28.3 %, $B1.1.243$ - 8.3 %, $B.1.527$ - 5.0 %, $B.1.1.325$ - 3.3 %, and 1.7 % for each $B.1.131$, $B.1.1.194$, B. Similarly, by Nextstrain classification the distribution of lineages is: 19A - 1.7 %, 20A - 51.7 % and 20B - 46.7 % (Figure 2).

Seven Ukrainian clusters contain multiple sequences and jointly constitute 73.3% of all sampled genomes. Presence of these clusters and the corresponding intra-country lineages indicate sustained internal transmissions (Figure 3A and Figure A1-A7 in Appendix). For each such lineage, a confidence interval of its introduction time can be assessed by the union of the confidence intervals for inferred dates of its Ukrainian MRCA v and the non-Ukrainian parent of v.

We analyzed these introduction times relatively to the implementation time of the travel ban, that was established on March 16, 2020 [15] for foreign citizens and on March 17, 2020 [16] for all travelers with the exception of Ukrainian citizens returning from abroad. It turned

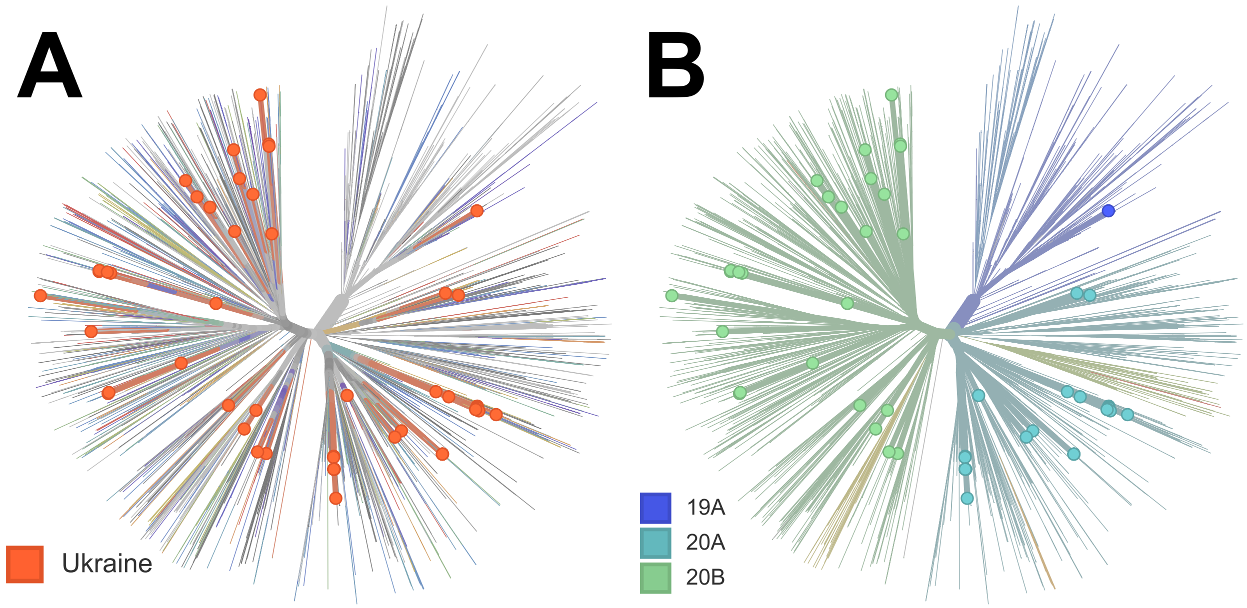

Figure 2.1 The global phylogenetic tree of SARS-CoV-2 genomes: A) distribution of Ukrainian SARS-CoV-2 genomes inside global SARS-CoV-2 population, B) the same tree with global SARS-CoV-2 lineages highlighted.

out that three out of seven lineages were most likely introduced into the country after the travel ban date, as indicated by their introduction confidence intervals (Figure 4). Similarly, a single lineage was likely imported before that date; for three remaining lineages the travel ban date falls into their confidence intervals, preventing us from the decisive conclusion, even though the date lies closer to the left ends of all intervals. Thus, the analysis support the hypothesis that the travel restrictions had limited effect on the virus importation control.

The estimates of the basic reproduction number $\mathcal{R}_0$ for three largest lineages are summarized in Table 2.1. All estimates are significantly above one, indicating sustained local transmission of SARS-CoV-2 during the first months of the epidemic in Ukraine.

Table 2.1 The estimates of the basic reproduction number $\mathcal{R}_0$ for three largest clusters together with the corresponding 95% confidence intervals (CI-s). The results are reported for two pairs of generation interval distribution parameters $\hat{\mu}$ and $\hat{\sigma}$ reported by two studies

Cluster	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mathcal{R}}_0$ **& 95% CI**	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mathcal{R}}_0$ **& 95% CI**
First	5.20	1.72	1.31 (1.12; 1.52)	3.95	1.51	1.23 (1.09; 1.37)
Second	5.20	1.72	1.47 (1.1; 1.98)	3.95	1.51	1.34 (1.07; 1.68)
Third	5.20	1.72	1.48 (1.16; 1.94)	3.95	1.51	1.35 (1.12; 1.65)

2.4 Discussion and Conclusion

In this study, we have detailed the epidemic characteristics of the first months of the COVID-19 pandemic in Ukraine and studied the effects of NPIs. We considered two complementary approaches based on the stochastic modeling applied to incidence data and genomic epidemiology methods applied to sequencing data. Different types of data reflect various aspects of the epidemics, and are prone to different biases. Therefore, such synthetic approach facilitates ubiquitous understanding of the early stages of the epidemic in Ukraine.

COVID-19 pandemic is characterized by a richness of available data, that allow to utilize agent-based modelling and genomic analysis at the finest possible resolution. In Ukraine, we have an access to public health data on the level of individual regions, which makes agent-based model predictions more comprehensive. Similarly, the advances and cost reduction of next-generation sequencing (NGS) methods allowed rapid genomic data acquisition at early stages of the epidemic [4]. These data processed by advanced phylogenetic and phylodynamic models allow to assess the virus importation and intra-country transmission dynamics from a “different angle” [29]. Furthermore, in cases when two methods produced independent estimations of the basic reproduction number $\mathcal{R}_0$, the obtained results are comparable, thus highlighting their consistency. The uncertainty estimates for the stochastic estimates are wider, which may be due to the fact that stochastic model has more parameters and higher variability in the outputs while phylodynamic models has pretty strong priors.

The study has limitations since the available surveillance incidence and genomic molecular data are limited. Ukraine is one of the poorest countries in Europe (based on GDP per capita), and, therefore, the health care infrastructure in Ukraine lacks in some parts the resources of its close and distant European neighbors [18]. As a result, both availability of screening tests and the reporting of incidence data during the initial epidemic likely (substantially) underestimated the burden of disease in terms of incidence counts. Likewise, as the pandemic evolved, the scarcity of genetic sequencing limited the number of sequence comparison in the phylogenetic analysis. As such, the actual number of viral clusters of local transmission remains unknown and should be interpreted as *at least seven* clusters which only form the "tip of the iceberg" of all transmission clusters. Moreover, the local population compliance with the NPI regulations implemented by officials is always a question, which might have reduced the effectiveness of such measures [9][14].

In summary, this study was among the first to explore the characteristics of the initial pandemic as it spread to Ukraine *and* provided additional *genomic analysis* not previously published.

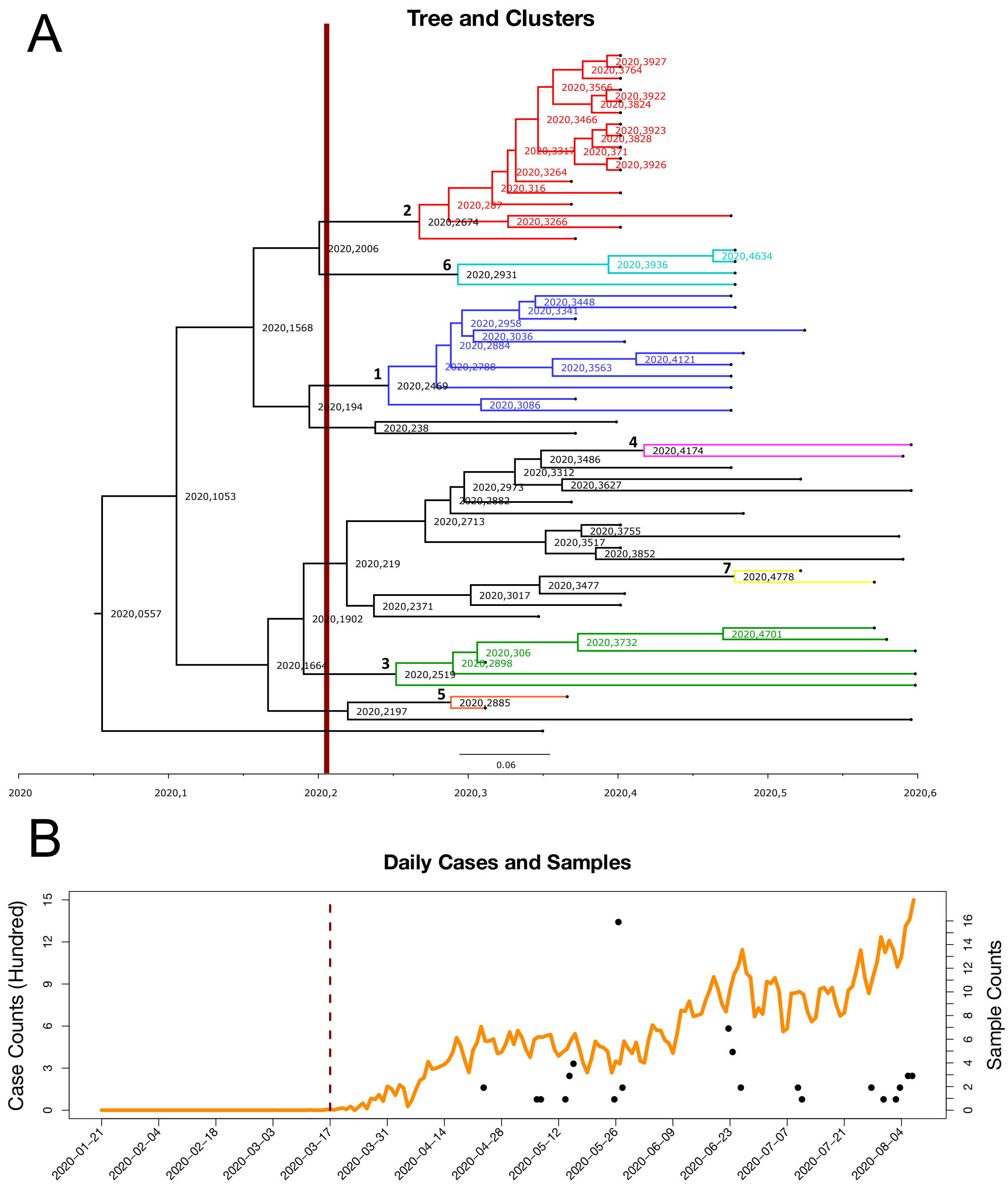

Figure 2.2 Panel A: The SARS-CoV-2 clusters are presented in the Ukrainian phylogenetic tree. Clusters colored by blue, red, green, pink, orange, azure, yellow, and numbered from one to seven, respectively. Panel B: Daily incidence of reported cases for Ukraine (orange) together with the sample counts and collection dates for sequenced samples (black). The travel restriction has happened on March 17, 2020, which is indicated by a vertical dark red bar time separator in both panels.

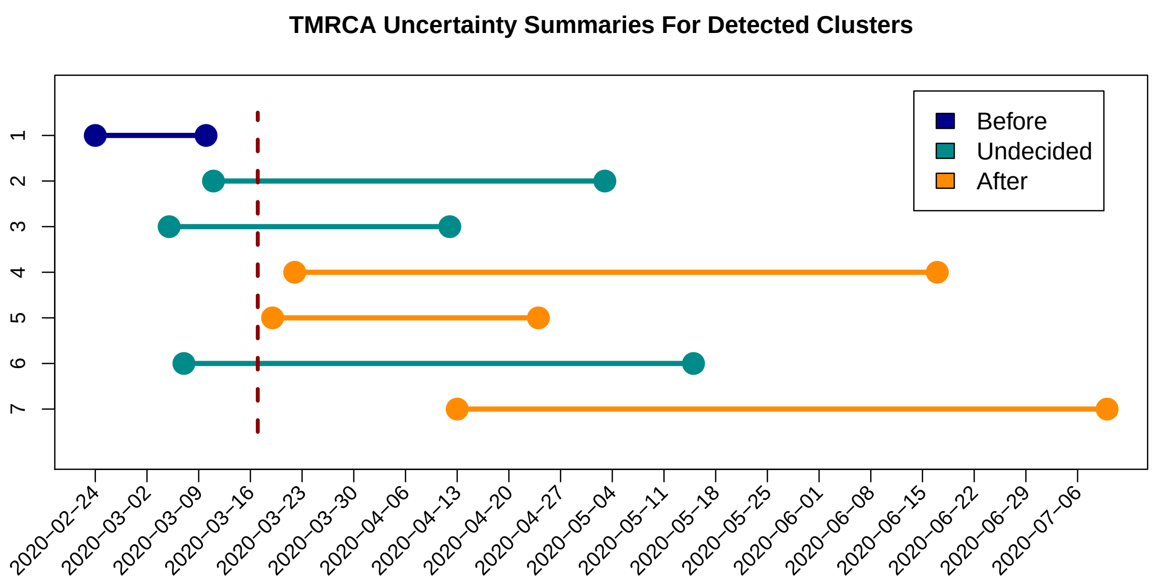

Figure 2.3 Introduction times for seven largest Ukrainian transmission lineages visualized from Table A2 from the supplement. The estimated intervals for introduction times are depicted as horizontal lines, the border closure date (March 17, 2020) is indicated by a vertical line.

CHAPTER 3

SARS-COV-2 TRANSMISSION DYNAMICS IN BELARUS IN 2020 REVEALED BY GENOMIC AND INCIDENCE DATA ANALYSIS

Published in *Communication Medicine* 1, Article number: (31), 2021

3.1 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 [38, 106] and lower outward population mobility [21]. At the same time, Belarus has a relatively modern healthcare system [90], and the country's Human Development Index (HDI) has been categorized as "very high" (vhHDI) [25].

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 [8, 12]. 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 notably narrower scope of non-pharmaceutical interventions (NPIs) in comparison to other vhHDI countries [30]. 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 [13]. 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 [59, 109]. 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 [13].

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 [27, 40, 43, 46, 52, 54, 68, 77, 94]. 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 large 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. Specifically, the analysis revealed a large number of unobserved infections, multiple virus introductions from different regions and a considerable amount of virus importations between Belarus and its geographic neighbors. Furthermore, the epidemiological dynamics during the first epidemic wave in Belarus was comparable to that in the neighboring Ukraine, where NPIs were broader. We also were able to observe a moderate reduction of the basic reproduction number after the introduction of limited NPIs, which, however, was lower in magnitude in comparison to the countries with large-scale NPIs.

3.2 Methods

3.2.1 Data

The SARS-CoV-2 genomic data for analysis were downloaded from GISAID [95] on March 15, 2021. The Belarusian dataset consists of 41 full-length genomes sampled between March 2020 and February 2021. Specimen and their metadata were originally collected by the Republican Research and Practical Center for Epidemiology and Microbiology (RRPCEM, 40 specimen) and Gomel State Medical University (1 specimen) in surveillance settings.

The sequenced samples were selected to represent all 7 major regions of Belarus; however, 60% of sequences represent the capital city of Minsk, where the testing facilities were better developed and whose population constitutes 22% of the country's total population. Females constitute 51% of sampled individuals, the age distribution was: ≤ 20 years – 15%, $21 - 40$ years – 26%, $41 - 60$ years – 49% and ≥ 61 years – 10%. 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 [22]

3.2.2 Global phylogenetic analysis

For the phylogeny reconstruction, we utilized the SARS-CoV-2-specific phylogenetic inference pipeline implemented in Nextstrain [56]. 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 [56]. The sequences were aligned using MAFFT [67], and a maximum likelihood (ML) phylogenetic tree was constructed using IQ-TREE [83] under Hasegawa-Kishino-Yano (HKY)+Γ nucleotide substitution model with a gamma-distributed site rate variation [58].

In the resulting time-labelled tree, ancestral geolocation traits have been inferred using so-called "mugration model" [92]. 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) [89] via KCMD Dynamic Data Hub [21]. 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 [85]. 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 [23].

3.2.3 Intra-country phylodynamic analysis

In this work we largely followed a general analytic pipeline adopted in other similar country-level studies (see e.g. [52, 68, 70, 94]), 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) [88]. Next, BEAST (v.2.6.3) [32] 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} [28, 52], 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) [86] 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) [61]. 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 using Birth Death Skyline Serial (BDSKY) model [98] 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 [52, 60, 68, 80]. 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.

3.2.4 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) [102]. In the second approach, we utilized the method of [109]. 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 [109] 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.

3.3 Results

SARS-CoV-2 genomic diversity. In total, the observed Belarusian SARS-CoV-2 41 sequences constitute 0.02% of officially reported cases. Despite such sparse sampling, the observed Belarusian SARS-CoV-2 sequences belong to 11 genomic lineages (by the nomenclature of [87], Fig.3.1a). 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 [39]. The root-to-tip regression analysis demonstrated moderately strong temporal signal ($R^2 = 0.62$, $p = 1.03 \cdot 10^{-9}$, Fig.3.1c).

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

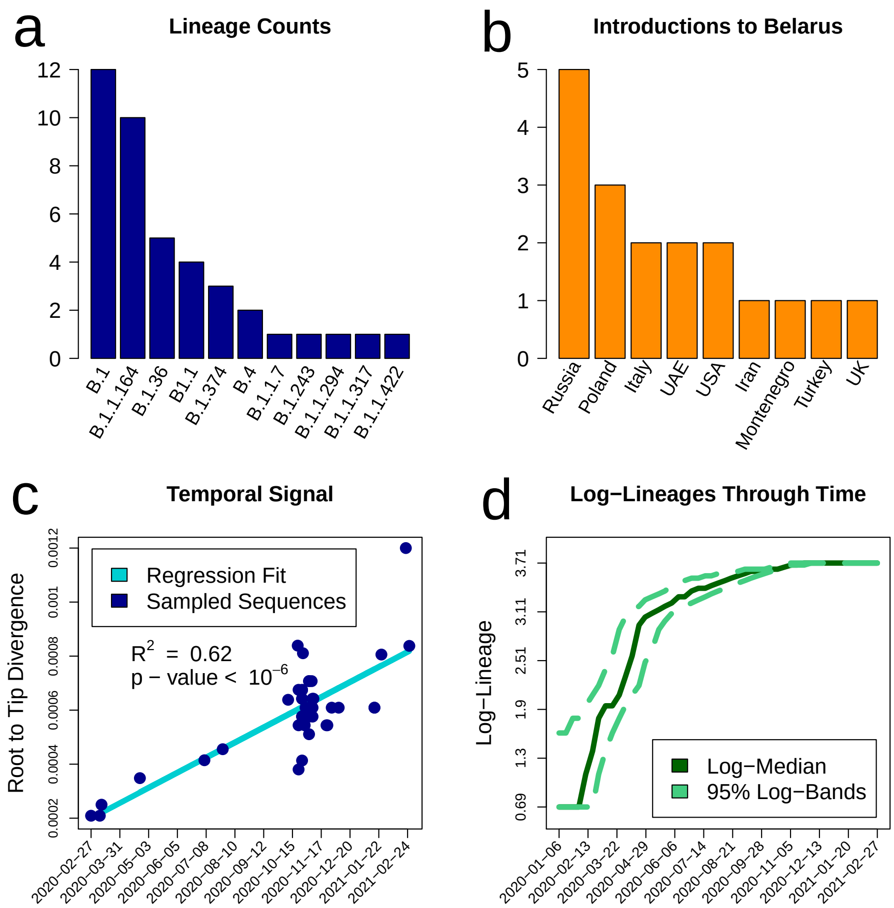

Figure 3.1 Lineages summaries. A) Abundances of genomic lineages. B) Estimated introduction sources to Belarus. C) Temporal signal. D) Lineages-through-time (on logarithmic scale).

when they were available. In particular, the first confirmed SARS-CoV-2 case was the individual who arrived from Iran [8], and the phylogenetics reaffirmed that. The agreement also held for the second detected case brought by the travelled from Italy [6]. The first introduction produced at least one secondary case as indicated by the tree; however, both lineages were not sampled after March, 2020 (Fig.3.2b). This can be attributed to the timely isolation of those individuals and their first order contacts [7]. In general, SARS-CoV-2 importations into the country could be attributed to a mixture of regional and global transmissions. As illustrated by Fig.3.1b and Fig.3.2c, the most frequent alleged virus introduction sources were the neighboring countries of Russia (5 introductions) and Poland (3 introductions).

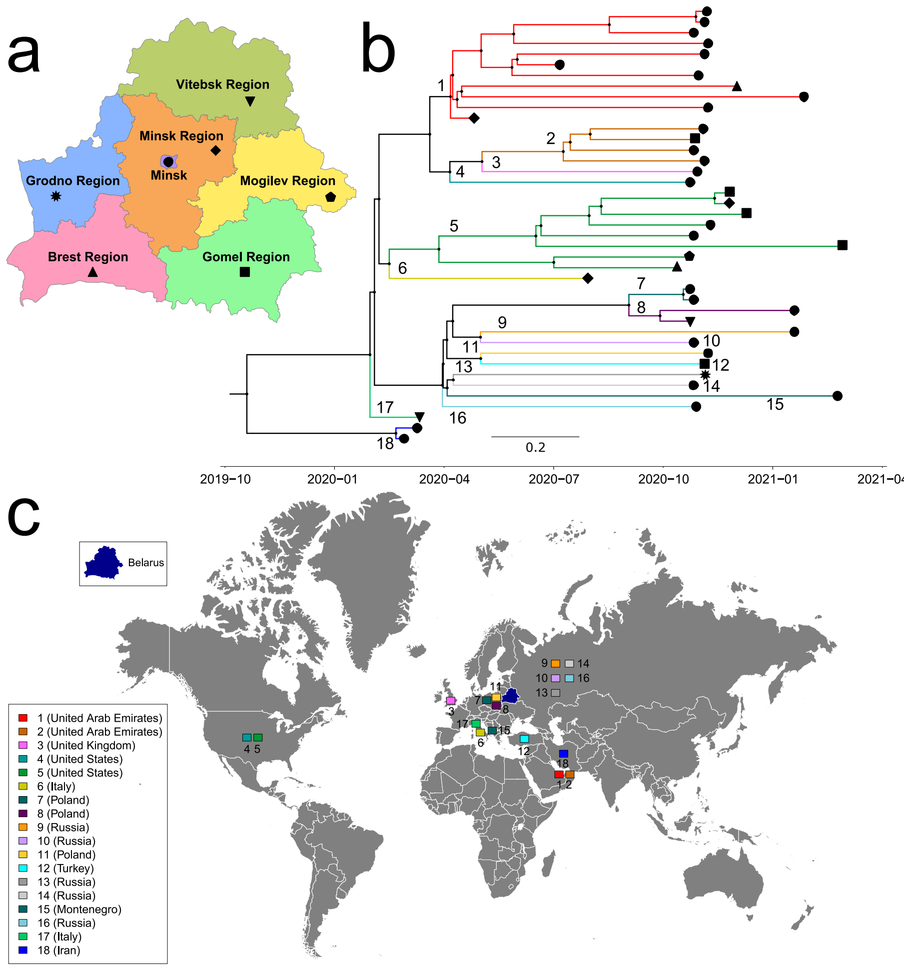

Figure 3.2 The annotated maximum clade credibility tree. A) Administrative regions of Belarus. B) The tree with the leaves color-coded by sampling regions (using the colors from panel A) and with the branches coded by the corresponding clusters. C) Cluster sources marked on the world map by their ids that correspond to panel B. Maps for figures were downloaded from Vemaps.com

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.

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 into April (Fig. 3.1d, Fig.3.2b). 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 [26]. 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.3.4d), 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.3.2b). 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 Birth-Death Skyline (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. 3.3b). The obtained HPD intervals, however, are rather wide due to the relatively small genome sample size. Thus, we also 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.

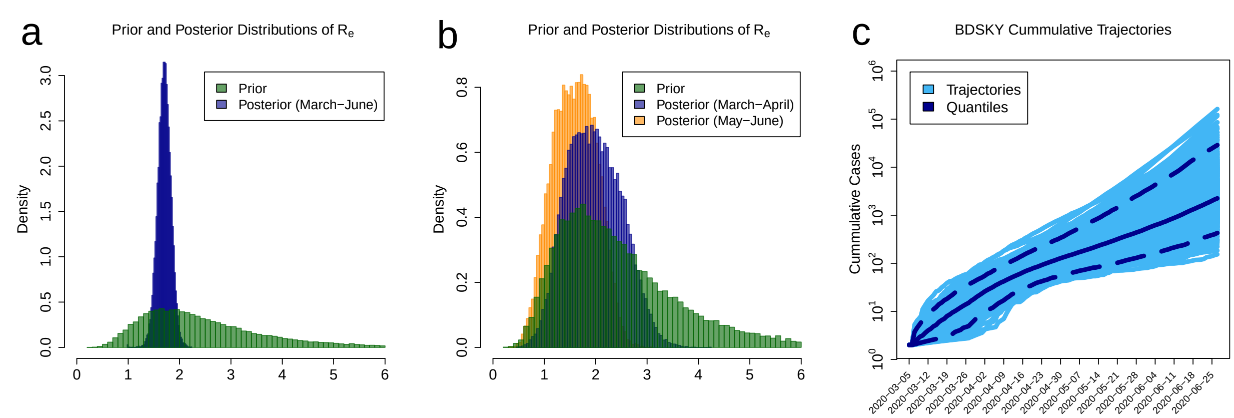

Figure 3.3 BDSKY model estimations. A), B) prior (green) and posteriors (blue and orange) distributions of the effective reproduction number estimate $\hat{\mathcal{R}}_e$ for Belarus during the first COVID-19 wave. C) The cumulative case count trajectories on $\log_{10}$ scale. Solid blue and dashed lines represents a median and 95% confidence intervals, respectively.

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 [30]. The same Birth-Death Skyline Serial model was applied to two best-sampled Ukrainian clusters with the total of 28 sequences defined as in [49]. 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 [49] and appeared to be comparable to $\hat{\mathcal{R}}_e$ estimates for Belarus.

Cumulative incidence and case counts. Cumulative case count trajectories for Belarus implied by the BDSKY model are reported on Fig.3.3c. 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.3.4. The model [109] 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.3.4b). 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.3.4c).

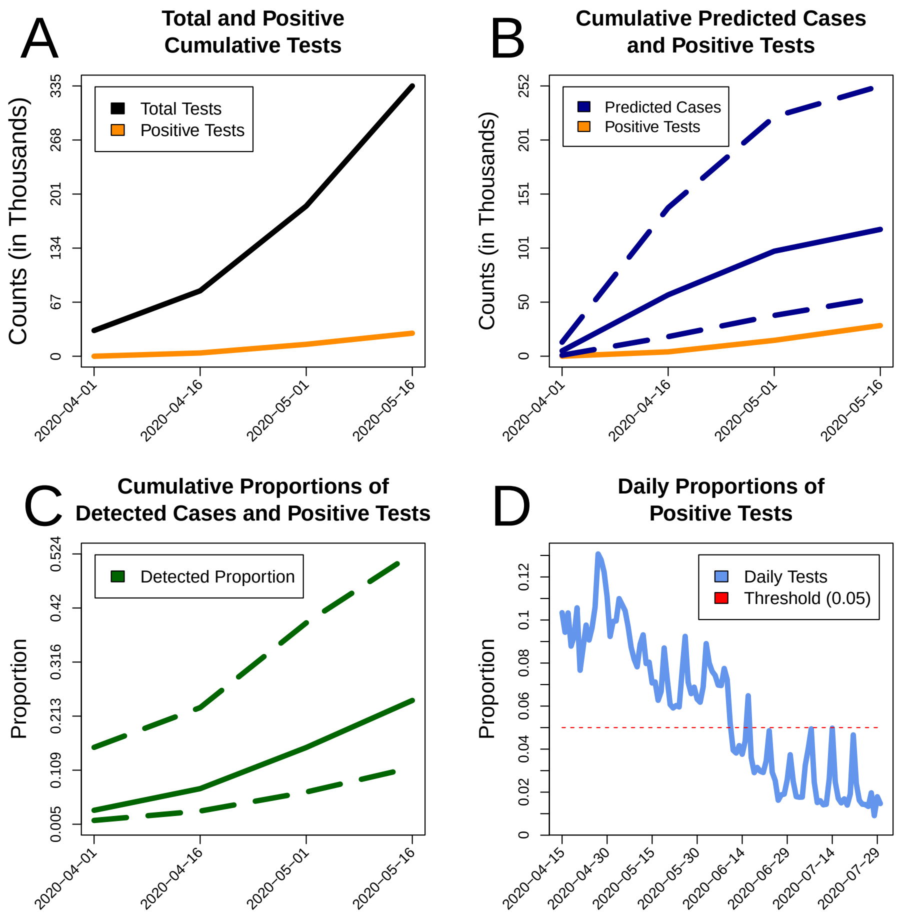

Figure 3.4 The summary of counts data analysis. A) Input data: officially reported cumulative numbers of cases $D(t)$ (orange) and conducted tests $T(t)$ (black). B) The cumulative numbers of officially reported cases ($D(t)$, orange) and the counts that were inferred by the model ($C(t)$, blue). C) 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.

3.4 Discussion and Conclusion

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 substantially 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 11 genomic lineages, which is comparable with similar early estimates for other regions [27, 40, 54, 68, 94]. However, for countries with denser sequence sampling these numbers are considerably larger [43, 52, 77]; thus the estimate for Belarus is most likely a lower bound on the real number of introductions. Still, it allows for several important observations. In contrast to most Western European and North American countries [27, 40, 43, 46, 52, 54, 77, 80, 94], the larger portion of estimated transmission links was with geographic neighbors. It 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. 3.3b). 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 [37, 77, 80, 94]. For example, for Victoria, Australia this value is 1.63 (CI: $(1.45; 1.8)$) [94]. 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 much 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 [9, 14]. Similar estimate $\mathcal{R}_e = 1.76$ $(0.91; 2.71)$ has been also reported for Russia [68] 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 [99, 109]. 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 [59].

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. BDSKY 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 [51, 80, 98], 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. Furthermore, new data may allow to compare first and second COVID-19 waves, which is especially interesting for understanding the epidemiological effects of social and public health developments during the second half of 2020. In our opinion, however, 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 [54, 68, 80]. 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 [92] 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.

The third limitation is the sparsity of the SARS-CoV-2 incidence and testing data. In contrast to other countries [17, 24], 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 considerable 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. Molecular surveillance of SARS-CoV-2 in Belarus is coordinated by Republican Research and Practical Center for Epidemiology and Microbiology (RRPCEM) represented by several co-authors of this paper. Prior to this study, such surveillance allowed to confirm the virus importation routes for the first cases and to detect newly emerging viral lineages (including alpha and delta variants) relatively soon after their appearance in neighboring countries. This information has been promptly reported to the government decision makers. However, in post-COVID era, advanced molecular surveillance necessary for informed and timely public health interventions should be based on deeper analysis, that resembles or extends the analysis reported here. Thus, 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.

CHAPTER 4

EPISTATIC DENSITY OF VIRAL VARIANTS IN ACUTE AND CHRONIC HCV PATIENTS

Written as *Conference Paper* and submitted for publication to Springer LNBI

4.1 Introduction

In evolutionary biology, fitness of any organism is often quantified as the reproductive success of a genotype or phenotype compared to others in the population. Fitness of viral variant determines the ability of a virus to survive, replicate, and transmit within a host or a population [42]. Viruses demonstrate remarkable evolutionary capabilities through their ability to adapt and maintain fitness. For instance, SARS-CoV-2 variants have shown repeated epidemic surges through enhanced spreading potential. This evolutionary process involves continuous genetic modifications that can either increase or decrease fitness of viral variants [65].

The fitness distribution of a virus is shaped by various genetic factors, including mutations at the level of single nucleotide variants (SNVs). However, fitness is not solely dictated by individual mutations but also by the interactions between them - a phenomenon known as epistasis. Recently, epistasis-based approaches have shown promise in predicting emerging COVID-19 variants [78]. Epistasis refers to the phenomenon where the effect of one genetic mutation is modified or influenced by the presence of another mutation. These epistatic interactions influence viral adaptation and evolution through driven adaptive trajectories by enabling mutations to collectively enhance or decrease fitness.

Traditional fitness estimations such as the log fitness ratio (LFR) [84], log relative fitness (LRF) [107], and production rate ratio (PRR) [108] have been employed to describe viral and cellular dynamic interactions and define competitive fitness indices of viral variants, they overlook the combinatorial effects of epistatic interactions.

To address this limitation, we propose to estimate epistatic density using an epistasis-based network approach that can enhance our understanding of fitness of viral variants in rapidly evolving viruses, such as the Hepatitis C Virus (HCV).

The core idea behind a novel machine learning method is that certain pairs of SNVs appear together more frequently than expected under the assumption that they are independent. This co-occurrence is indicative of potential epistatic interactions, where the presence of one mutation influences the effect or occurrence of another. These linkages are crucial for understanding patterns of correlated mutations that may influence a pathogen's evolution, particularly traits such as fitness, pathogenicity, or drug resistance. The method represents linked SNVs in the form of a coordinated substitution network, a graphical framework where each SNV is a node (or vertex), and significant pairwise linkages between SNVs are represented as links (or edges) connecting the nodes. By constructing this network, we can visualize and analyze the relationships between mutations. A key strength of the coordinated substitution network approach is its ability to identify dense communities within the network. These dense communities often correspond to groups of mutations that collectively influence the viral phenotype and its characteristics such as transmissibility, immune escape, or antiviral drug resistance [78].

Epistatic density definition and model. We define epistatic density as the number of edges representing advantageous interactions or connections and calculate as the number of edges representing positively linked SNV pairs within the coordinated substitution network. Our model assumes that emerging strains have higher epistatic density and more positive edges as dense community, in comparison, to other viral variants.

Data. We computed the epistatic density of intra-host HCV populations sampled from 85 chronic and 28 acute patients with the HCV 1a genotype, using data provided by the Centers for Disease Control and Prevention (CDC). We provide an optional step for classifying acute and chronic patients within the HCV dataset using epistatic density values and a statistical approach.

Contributions. First, we observe that chronic HCV populations exhibit higher epistatic density compared to acute cases, suggesting a more stable adaptation over time. Second, epistatic density distributions in acute cases appear more irregular and choppier than in chronic cases, indicating dynamic evolutionary pressures during early infection. Finally, we demonstrate that epistatic density metrics can serve as a discriminating factor between acute and chronic infections, offering potential applications in viral surveillance.

4.2 Methods

We constructed an epistatic network to analyze the epistatic density of viral variants in HCV patients by identifying positively linked SNV pairs using statistical tests. The epistatic network was built for all HCV patients, by representing SNVs as nodes and significant epistatic interactions as edges, enabling e-fitness calculation based on the number of edges representing positively linked SNP pairs within the epistatic network. E-fitness scores were computed at both the group level using aggregated patient data and the individual level by comparing each patient's mutation profile against epistatic network. To characterize viral evolution, we performed e-fitness statistical analysis, computing measures such as median fitness, normalized twins fitness, and normalized distinct fitness. Finally, we evaluated the model's classification accuracy using ROC analysis, demonstrating that e-fitness parameters effectively distinguish between acute and chronic HCV infections.

4.2.1 Research Roadmap Description

The research roadmap for constructing and analyzing epistatic networks of the Hepatitis C Virus (HCV) encompasses several critical steps, from data processing to network modeling, fitness calculation, and performance evaluation (Figure 4.1). These methodological steps are described in detail below.

The first step involved data processing, where genomic sequences were analyzed to identify single nucleotide polymorphism (SNP) candidates. For each viral haplotype, nucleotide positions were compared to the reference sequence: matching positions were assigned a value of 0, and mismatched positions a value of 1. As a result, binary matrices were generated for each patient sample, where rows represented the viral haplotypes and columns represented positions in the sequence. These matrices served as input for subsequent analyses of mutation patterns.

Following data processing, epistatic networks were constructed to represent interactions,

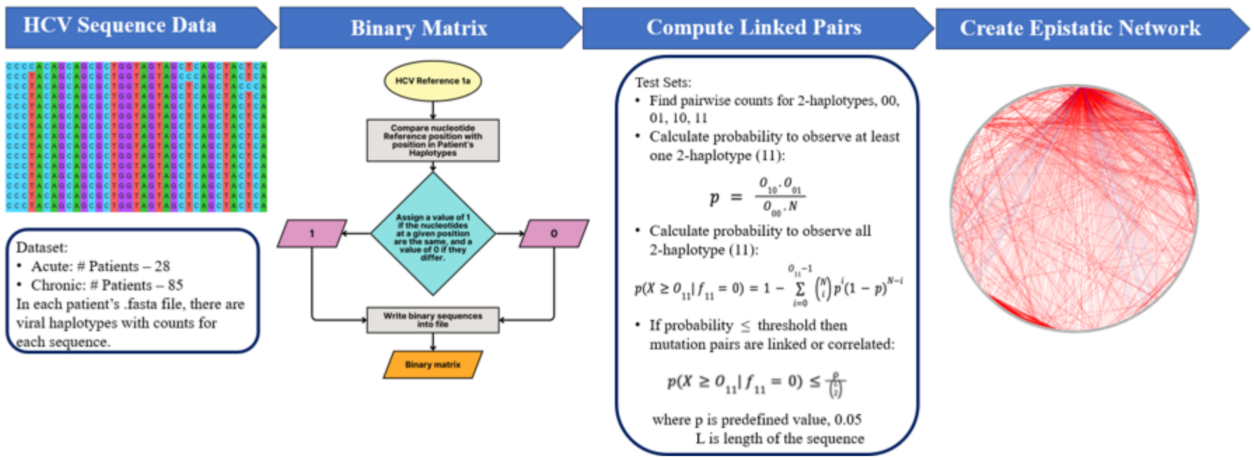

Figure 4.1 Epistatic Network Construction Pipeline

.

or dependencies, between mutations. SNPs were treated as nodes, and edges represented significant interactions between these loci. To detect interactions, a statistical test was used to calculate the likelihood of observing co-occurring mutations based on counts of four pairwise SNP combinations:

((0,0), (0,1), (1,0), (1,1)). Significant interactions were identified using a p-value threshold with a Bonferroni correction to control false positives. An adjacency matrix was generated to encode these interactions, which was later converted into an edge matrix, a key component for fitness calculations.

The next step involved fitness evaluation to quantify how well the mutations in a viral population adhered to the network of interactions. A fixed epistatic network (EN) was derived as a global model based on aggregated data from all patients. Using this EN, fitness scores were calculated for both patient groups and individual patients. Group-level fitness scores were computed by combining the edge matrix (capturing SNP interactions) and the population matrix (representing mutation presence across patients). Similarly, individual-level fitness scores were calculated to evaluate how each patient's mutation profile aligned with the global epistatic constraints.

To better understand these fitness distributions, statistical analysis was performed on the fitness data for each patient. Various metrics such as median, range, minimum, maximum, twin mutations (pairs differing by exactly one), and diversity measures (e.g., unique values) were calculated. Missing values were identified by comparing observed fitness values with their expected ranges. Data processing and statistical calculations were performed using Python libraries such as NumPy, SciPy, and Pandas. Finally, the model's performance was evaluated using receiver operating characteristic (ROC) analysis. Patients were classified as recently or persistently infected based on fitness metrics like median, twin mutation ratios, and unique value ratios. Sensitivity and specificity were computed iteratively, normalized, and used to generate ROC curves. These curves plotted the true positive rate (sensitivity) against the false positive rate (1-specificity), and the area under the curve (AUC) was calculated to quantify classification accuracy. This process enabled an objective evaluation of the model's capability to distinguish between patient groups.

4.2.2 Finding pairs of positively linked SNPs

To identify genomic positions that could be potential candidates for single nucleotide polymorphism (SNP) variants, we performed a comparison for each patient's sample by aligning nucleotide reference positions with all corresponding positions in multiple viral haplotypes, one at a time. Each position was scored as follows: a match between the reference and haplotype positions was assigned a value of 0, while a mismatch was assigned a value of 1. Binary matrices were then constructed with dimensions $n \times m$, where nn represents the number of multiple viral haplotypes and mm denotes the number of positions in the patient's sequence.

We analyzed a viral population composed of N haploid genotypes, each with a length of L. For any two distinct loci u and v within the set $\{1, \ldots, L\}$, we examine the four possible 2-haplotypes, which are defined by the pairs $(i, j) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Here, 0 and 1 represent the reference and alternative alleles at loci u and v, respectively. Additionally, let O_{ij} denote the observed count of each 2-haplotype (i, j) for $i, j = 0, 1$.

We approximately estimate the probability of the event that there exist a large number of genomes with the 2-haplotype $(U_1; V_1)$ given that this 2-haplotype is not viable. Specifically, by formula approximates the largest probability of observing a genome containing 2-haplotypes $(U_1; V_1)$ among N sequenced genomes given that $f_{11} = 0$:

$$p = \frac{O_{10} \cdot O_{01}}{O_{00} \cdot N} \tag{4.1}$$

Then we can assume that the number of such genomes X follows the binomial distribution $B(N; p)$, and the probability that $X \geq O_{11}$ can be calculated as

$$p(X \geq O_{11} \mid f_{11} = 0) = 1 - F_X(O_{11} - 1) = 1 - \sum_{i=0}^{O_{11}-1} \binom{N}{i} p^i (1-p)^{N-i} \tag{4.2}$$

where F_X is the cumulative distribution function of the binomial distribution. We assume that SNPs U_1 and V_1 are linked (i.e., adjacent in the epistatic interaction network G), when the probability in equation (2) is low enough, i.e.,

$$p(X \geq O_{11} \mid f_{11} = 0) \leq \frac{p}{\binom{L}{2}} \tag{4.3}$$

where p is a predefined p-value (in this study we used $p = 0.05$) and the denominator $\binom{L}{2}$ is a Bonferroni correction.

In what follows, we use the statistical test (4.3) to construct epistatic interaction networks G_i using HCV sequences.

An epistatic network is an SNP linkage graph where vertices are SNPs and edges are linked SNP pairs.

4.2.3 Epistatic Network (EN) Construction and Fitness Calculation

In the context of viral genomes, an epistatic network is used to represent the complex interactions which are non-linear between different mutations within the viral genome, this representation lets us understand where the effect of one mutation is dependent on the presence of others. These interactions play a critical role in understanding viral evolution, adaptation, and drug resistance. Nodes or vertices in an epistatic network represent the nucleotide or amino acid positions in the viral genome. Edges represent the dependencies or correlation/dependence between mutations, often derived from covariance analysis of sequencing data or like in our study fitness.

4.2.4 Epistatic Network for All HCV Patients Together

A Epistatic Network (EN) is a network where mutation interactions (epistasis) are considered constant across all patients. In this context, for HCV patients:

- The adjacency matrix (computed using `create_adj_matrix_binom`) captures significant interactions between mutations.
- This EN is derived from aggregated viral sequence data across all patients.
- It serves as a global model of mutational dependencies, assuming that the same set of interacting mutations influences viral evolution in all patients.

The function `create_adj_matrix_binom` constructs an adjacency matrix based on binomial significance testing for epistatic interactions between mutations.

1. Initialize Adjacency Matrix

$$A \in R^{columns\_count \times columns\_count}$$

where $A_{i,j}$ represents the interaction strength between mutation sites i and j.

2. For Each Pair (i, j):

Extract counts:

$$z_z, z_o, o_z, o_o$$

where:

- z_z = both positions have wild type (0,0)

- z_o = first is wild type, second is mutant (0,1)
- o_z = first is mutant, second is wild type (1,0)
- o_o = both positions have mutant (1,1)

3. Compute Binomial Probability

- If $(z_z \cdot o_o) > (z_o \cdot o_z)$, compute:

$$p = \frac{(o_z \cdot z_o) + 1}{(z_z \cdot columns\_count) + 1}$$

- If $(z_z \cdot o_o) < (z_o \cdot o_z)$, compute:

$$p = \frac{(o_o \cdot z_z) + 1}{(z_o \cdot columns\_count) + 1}$$

4. Check Statistical Significance

- Compute binomial cumulative distribution function:

$$\text{prob} = 1 - F(o_o; \text{columns\_count}, p)$$

- If prob $\leq \frac{0.05}{\text{combinations}(columns\_count,2)}$, assign:

$$A_{i,j} = 1 \text{ (positive interaction)} \quad \text{or} \quad A_{i,j} = -1 \text{ (negative interaction)}$$

4.2.5 Fitness Calculation for Groups Using Fixed EN

To calculate fitness for patient groups:

- The edge matrix is calculated using `create_edge_matrix`, encoding mutation interactions.
- The population matrix is constructed by transposing the group's sequence data.
- The fitness matrix is computed via matrix multiplication:

$$F = E \times P \tag{4.4}$$

where:

- E is the edge matrix from the fixed EN.
- P is the mutation presence matrix for a group of patients.

The fitness vector is computed as the sum of fitness contributions per sequence and normalized.

The function `create_edge_matrix` converts an adjacency matrix into an edge matrix that encodes epistatic interactions between mutations.

1. Preprocessing the Graph

   - Converts `adj_matrix` into a graph G.
   - Removes isolated nodes to retain only interacting mutations.

2. Initialize Edge Matrix

$$E \in R^{|edges| \times columns\_count}$$

where $|edges|$ is the number of significant interactions.

3. Extracting Edge Information

For each edge (i, j), assign:

$$w_{i,j} = \begin{cases} +1, & \text{if color} = \text{red} \\ \\ -1, & \text{if color} = \text{blue} \end{cases}$$

4. Populating Edge Matrix

For each edge (i, j), update:

$$E[k, i] += w_{i,j}, \quad E[k, j] += w_{i,j}$$

This encodes the interaction between mutation sites.

4.2.6 Fitness Calculation for Each Patient Using Fixed EN

- The process is similar to group-level fitness but applied at the individual patient level.
- The individual patient's mutation profile is used instead of a group-level population matrix.
- The same fixed EN adjacency and edge matrix are used to ensure consistency.

- The resulting fitness score quantifies how well a given patient's viral sequence aligns with the global epistatic constraints.

This approach ensures that each patient's viral evolution is evaluated using a common reference network, allowing for direct fitness comparisons across patients or groups.

The repository to the pipeline can be found at: `https://github.com/Akshayjuyal/fitness_viral_gnome`

4.2.7 E-fitness statistics calculation and data processing

For each patient, we computed a set of statistical measures based on fitness values obtained from multiple viral haplotypes, including: minimum, maximum, median, range, number of twins, twins divided by range, number of distinct values, distinct values divided by range, total number of measurements, missing values, skewness, number of segments of lengths 1, 2, and 3 or more, and the total number of segments. Data processing and statistical calculations were performed using Python (version 3.10.9) [48] with the NumPy [57], SciPy [103], Pandas [76], and statistics libraries.

Fitness values were sorted in ascending order. The minimum and maximum fitness values were extracted directly from the sorted dataset. The median was computed as the middle value of the sorted fitness scores. The range was defined as the difference between the maximum and minimum fitness values:

$$\text{Range} = \max(\text{fitness values}) - \min(\text{fitness values}) \tag{4.5}$$

The total number of measurements was counted as the total number of recorded fitness values.

The number of distinct values was defined as the count of unique fitness scores, irrespective of repetition. To quantify the diversity of fitness scores relative to their range, the number of distinct values was divided by the range:

$$\text{Distinct values ratio} = \frac{\text{Number of distinct values}}{\text{Range}} \tag{4.6}$$

Twin values were defined as consecutive fitness values differing by exactly one. This was calculated using the `np.diff()` function in Python, counting occurrences where the difference equals one. The twin values were then normalized by the range:

$$\text{Twin values ratio} = \frac{\text{Number of twin values}}{\text{Range}} \tag{4.7}$$

The number of missing values was calculated by comparing the expected range of values with the actual observed values. First, we determined the total number of possible fitness values (expected count) within the observed range:

$$\text{Expected count} = \max(\text{fitness values}) - \min(\text{fitness values}) + 1 \tag{4.8}$$

Then, we counted the number of distinct fitness values present in the dataset. Finally, the missing values were computed as:

$$\text{Missing values} = \text{Expected count} - \text{Number of distinct values} \tag{4.9}$$

4.2.8 Performance Evaluation Using ROC Analysis

The statistical analysis employs sensitivity, specificity, receiver operating characteristic (ROC) curves, and area under the curve (AUC) metrics to evaluate the classification performance between acute and chronic HCV patients. We analyze the complete set of HCV patients with calculated e-fitness parameters, including median values, twin values ratios, and distinct values ratios, arranged in ascending order. The iterative process begins with zero counts for both sensitivity and specificity measures. For acute group patients, we increment the sensitivity count while maintaining the specificity count. Conversely, for chronic group patients, we increment the specificity count while maintaining the sensitivity count. The normalization procedure standardizes both sensitivity and specificity measurements. We calculate normalized sensitivity by dividing the cumulative count by the total number of acute group patients. Similarly, we compute normalized specificity by dividing the respective count by the total number of chronic group patients. The ROC curves are generated by plotting the false positive rate (1 - specificity) on the x-axis against the true positive rate (sensitivity) on the y-axis. Each plot includes a diagonal reference line representing random classifier performance. The plots are generated for three distinct scenarios: twin values ratios, distinct values ratios, and median values. The ROC analysis is implemented using scikit-learn (version 1.2.1)[**?**], a robust Python machine learning library. The AUC calculations utilize scikit-learn metrics.auc function, providing a quantitative measure of the classifier's discriminative capability. This comprehensive methodology enables objective evaluation of the classification method effectiveness in distinguishing between acute and chronic

HCV patients, providing a robust framework for assessing the discrimination potential of the proposed approach.

4.3 Results

We analyzed intra-host HCV populations from acute (N=98) and chronic HCV patients (N=257), as described in [71]. From this dataset, we selected intra-host populations of the HCV 1a genotype, including samples from acute (N=28) and chronic (N=85) patients, following the criteria outlined in [71]. HCV 1a genotype haplotypes were extracted from a mixed population of various HCV genotypes using a matching threshold of 83 % or higher to the HCV 1a reference sequence from NCBI [81]. The E1/E2 junction of the HCV 1a reference sequence was selected through pairwise alignment and trimmed from both ends using an HCV 1a sequence from a patient. The E1/E2 junction of the HCV genome (L=263 nt each), which contains the hypervariable region 1 (HVR1), was sequenced using the GS FLX System along with the GS FLX Titanium Sequencing Kit from 454 Life Sciences, Roche, Branford, CT. Obtained sequences were processed using the Kmer Error Correction (KEC) algorithm [97] for error correction and haplotyping. All HCV sequences were aligned using MUSCLE [44]. The average number of sequences obtained per patient for acute and chronic HCV intra-host populations was n=11,899 and n=6,054, respectively.

To examine the evolutionary dynamics of intra-host HCV populations, we analyzed the epistatic density distributions of acute and chronic infections. Epistatic density distribution histograms were generated to visualize the distribution of sequences across different density values, allowing for a comparison of selective pressures and adaptive patterns.

Distinct patterns were noted in the sequence distribution across epistatic density values within these populations, emphasizing the choppy nature of their distributions. Analyzed density histograms illustrate the distribution of sequences across different density values for intra-host HCV populations, with a focus on the choppy nature of the density distribution (Figure 4.2).

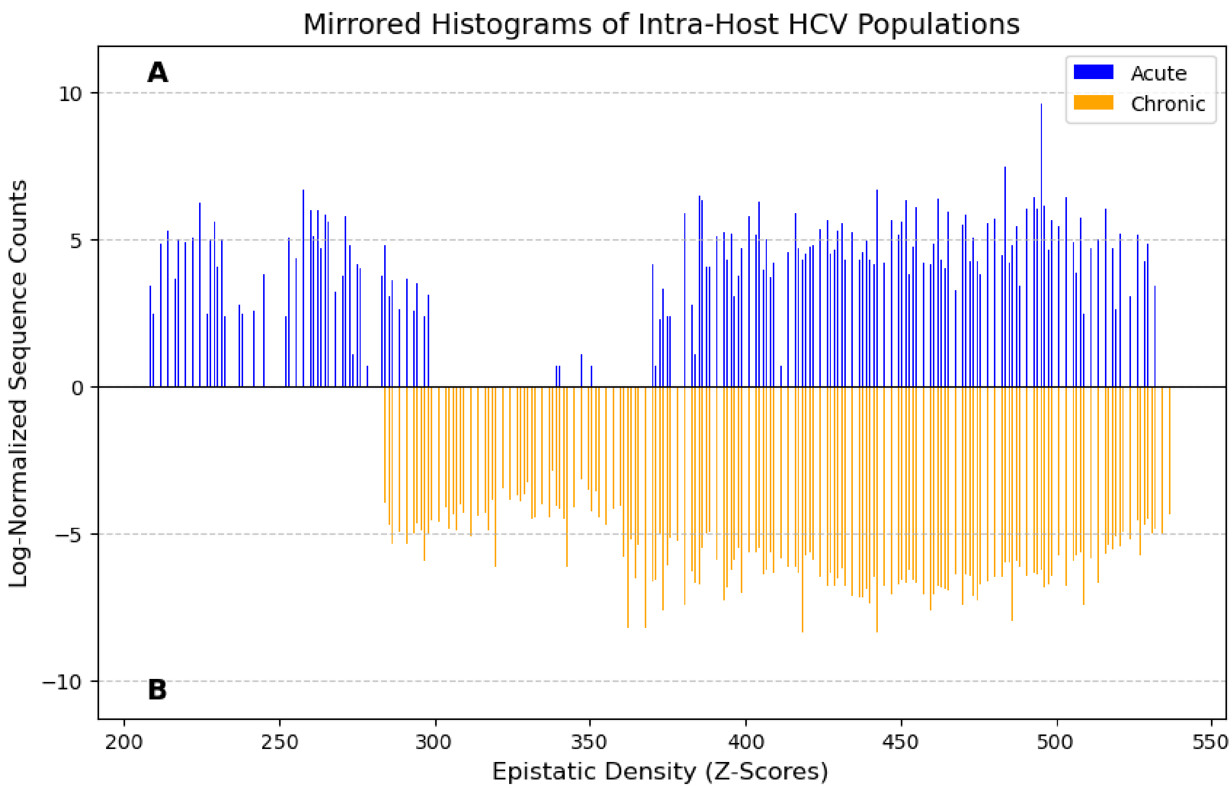

Figure 4.2 Comparison of epistatic density distributions of acute (Panel A) and chronic (Panel B) intra-host HCV populations. The x-axis represents epistatic density values, which were normalized using z-score transformation, while the y-axis denotes sequence counts, which were log-normalized to improve visualization.

Epistatic density distribution for acute intra-host HCV population showed a broader range of density values from 500 to 8000 but the sequence counts are relatively sparse, indicating that only a few density values are represented. Density value peaks are less pronounced, suggesting that there may be fewer advantageous mutations or adaptations in this broader range, leading to a less diverse population in terms of fitness (Panel A, Figure 4.2). In contrast, density distribution for chronic intra-host HCV population has a narrower range of values from 500 to 2250, revealing a more intricate and choppy pattern. Here, the peaks are more pronounced, indicating that certain density values are significantly more favorable, resulting in a higher concentration of sequences. This choppiness suggests that the population is experiencing strong selective pressures, favoring specific traits that enhance fitness (Panel B, Figure 4.2).

Overall, the comparison of these two histograms highlights the complexity of the density distribution in HCV populations. The choppy nature of the second histogram indicates a dynamic evolutionary process, where specific epistatic density peaks may represent adaptive strategies that are being favored in chronic versus acute infections. Understanding these patterns can provide insights into viral evolution.

We conducted a comprehensive analysis using receiver operating characteristic (ROC) curves and area under the curve (AUC) values to assess the classification performance of our model (Figure 4.3).

Figure 4.3 presents three ROC curves, each evaluating the classification performance of different metrics distinguishing acute and chronic HCV patients.

Panel A, Figure 4.3 the ROC curve (in orange) illustrates the classification performance

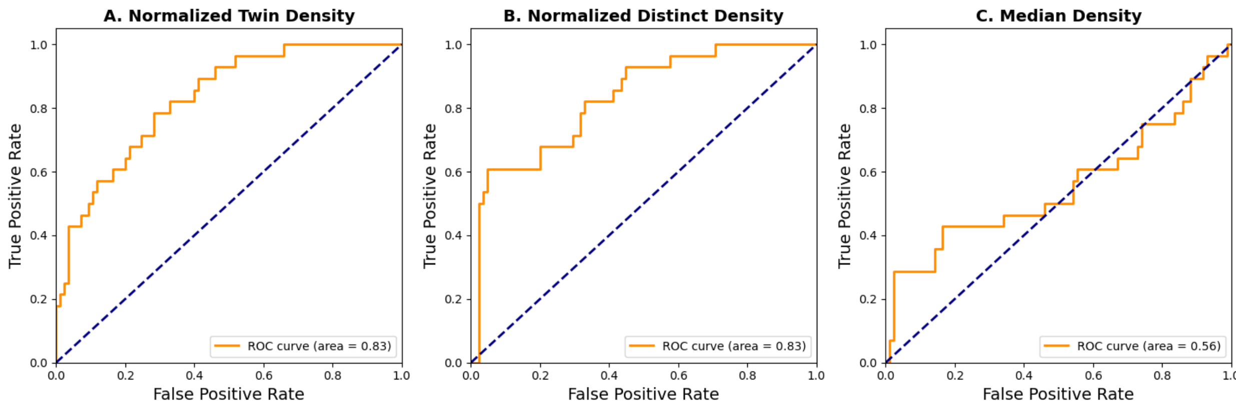

Figure 4.3 Receiver Operating Characteristic (ROC) curves for classifying HCV acute and chronic patients based on normalized twin density values (Panel A), normalized distinct density values (Panel B), and median density values (Panel C).

based on normalized twin density values. The AUC is 0.83, indicating strong model performance in distinguishing between acute and chronic HCV cases. Higher AUC values (closer to 1.0) suggest superior discriminative ability.

Panel B, Figure 4.3 displays the ROC curve for classification using normalized distinct density values. The curve follows a similar trajectory to Panel A, with an AUC of 0.83, confirming robust classification efficacy comparable to the normalized twin density values.

In contrast to the previous panels, Panel C, Figure 4.3 the ROC curve for median density values exhibits significantly lower discriminative performance, with an AUC of 0.56. This value suggests poor classification capability, as the curve is closer to the diagonal reference line (dashed blue), which represents a random classifier.

Overall, the results highlight that both normalized twin and distinct density values are effective for classifying HCV infection types, whereas median density values lack sufficient discriminatory power.

4.4 Discussion and Conclusion

This study indicates that the epistatic density of viral variants increases with evolution within the viral population. Our proposed epistatic density model assumes that emerging strains have higher density in the population and more positively linked edges as a dense community within the coordinated substitution network compared to other viral variants. The epistatic density is defined as the number of positively linked edges representing advantageous interactions in a genome, calculated as the number of edges representing positively linked SNV pairs within the coordinated substitution network. Additionally, our research findings suggest that analyzing coordinated substitution networks and calculating statistical parameters of epistatic density can effectively distinguish acute from chronic HCV patients. Specifically, the normalized twin and distinct density values, derived from the epistatic density statistical analysis, demonstrate strong discriminatory power in classifying HCV infection types. Thus, we highlight the potential of this approach for understanding viral evolution and disease progression. These findings support the hypothesis that chronic HCV infections, due to their longer evolutionary development time, exhibit distinct epistatic density patterns compared to acute infections.

While the study presents a novel approach for classifying HCV infections using coordinated substitution networks, there are potential limitations to consider. The sample size, particularly the smaller number of acute individuals (N=28) compared to chronic individuals (N=85), may introduce bias and affect the generalizability of the findings. Furthermore, we restricted viral variants to the HCV 1a genotype to exclude the effects of genetic heterogeneity on classification results. We acknowledge that the reliance on specific statistical thresholds and parameters for constructing the coordinated substitution networks could influence the network structure and, consequently, the epistatic density calculations, potentially affecting the robustness of the classification.

This study introduces a novel machine learning model that leverages coordinated substitution networks for classifying HCV infection. By incorporating epistasis as a feature in the classification model, the approach captures epistatic interactions between mutations, offering valuable insights into coordinated mutation patterns that influence viral epistatic density. The model demonstrates strong discriminatory power, effectively distinguishing between acute and chronic HCV patients. This classification method has the potential to be adapted for other infectious diseases, offering a robust approach to the identification of mutation patterns and epistatic interactions that influence viral evolution. The results highlight that normalized twin and distinct density values are highly effective in classifying HCV infection types, while median density values do not provide sufficient discriminatory power, as discussed in the Results section.

REFERENCES

[1] 2020, BEAST Software, https://beast.community/estimating_R0.html, accessed: 2021-02-20

[2] 2020, BEAST Software, https://beast.community/phylodynamics_of_epidemic_influenza, accessed: 2021-02-20

[3] 2020, CDC, https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html, accessed: 2021-02-20

[4] 2020, GISAID, A global initiative on sharing avian flu data., https://www.gisaid.org/, accessed: 2021-02-20

[5] 2020, Kyiv Post, https://www.kyivpost.com/ukraine-politics/developing-first-coronavirus-case-identified-in-ukraine.html, accessed: 2021-02-20

[6] 2020, 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/, accessed: 2021-03-31

[7] 2020, 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/, accessed: 2021-03-31

[8] 2020, 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/, accessed: 2021-03-31

[9] 2020, 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, accessed: 2021-06-29

[10] 2020, WHO, https://www.who.int/dg/speeches/detail/%who-director-general-s-opening-remarks-at-the-media-briefing-on-%COVID-19---11-march-2020, accessed: 2021-03-31

[11] 2020, WHO, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public, accessed: 2021-02-20

[12] 2020, WHO Coronavirus disease 2019 (COVID-19). Situation report 40, http://hdr.undp.org/en, accessed: 2021-03-31

[13] 2020, 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, accessed: 2021-03-31

[14] 2021, 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/, accessed: 2021-06-29

[15] 2021, baltsi.mfa.gov.ua - No Entry for Foreigners -

May 26, 2020, https://baltsi.mfa.gov.ua/ru/news/%timchasova-zaborona-do-24-kvitnya-2020-r-na-peretinannya-%derzhavnogo-kordonu-na-vyizd-v-ukrayinu-dlya-inozemciv-ta-%osib-bez-gromadyanstva, accessed: 2021-02-20

[16] 2021, Closed borders, https://www.kmu.gov.ua/news/%uryad-timchasovo-zaboroniv-vyizd-v-ukrayinu-inozemciv-%ta-zakriv-mizhnarodne-pasazhirske-aviaspoluchennya, accessed: 2021-02-20

[17] 2021, COVID-19 Map - Johns Hopkins Coronavirus Resource Center, https://coronavirus.jhu.edu/map.html, accessed: 2021-03-31

[18] 2021, emerging-europe.com, https://emerging-europe.com/news/imf-ukraine-is-emerging-europes-poorest-country/, accessed: 2021-03-01

[19] 2021, github.com, https://github.com/nextstrain/ncov, accessed: 2020-11-01

[20] 2021, Johns Hopkins University - Coronavirus Resource Center, https://coronavirus.jhu.edu/map.html, accessed: 2021-03-01

[21] 2021, KCMD Dynamic Data Hub, https://bluehub.jrc.ec.europa.eu/migration/app/, accessed: 2021-03-31

[22] 2021, Ministry of Public Health of the Republic of Belarus - Official Telegram Channel (in Russian), https://t.me/s/minzdravbelarus, accessed: 2021-03-31

[23] 2021, pangolin.cog-uk.io, https://pangolin.cog-uk.io/, accessed: 2021-03-31

[24] 2021, The National Health Service of Ukraine (NHSU) - COVID-19 Dashboard, https://nszu.gov.ua/e-data/dashboard/covid19, accessed: 2021-03-31

[25] 2021, United Nations Human Development Reports, http://hdr.undp.org/en, accessed: 2021-03-31

[26] 2021, 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, accessed: 2021-03-31

[27] Alteri, C. et al. 2021, Nature communications, 12, 1

[28] Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. 2020, Nature medicine, 26, 450

[29] Armstrong, G. L., MacCannell, D. R., Taylor, J., Carleton, H. A., Neuhaus, E. B., Bradbury, R. S., Posey, J. E., & Gwinn, M. 2019, New England Journal of Medicine, 381, 2569

[30] Åslund, A. 2020, Eurasian Geography and Economics, 1

[31] Benvenuto, D., Giovanetti, M., Ciccozzi, A., Spoto, S., Angeletti, S., & Ciccozzi, M. 2020, Journal of medical virology, 92, 455

[32] Bouckaert, R. et al. 2014, PLoS Comput Biol, 10, e1003537

[33] Caddy, S. 2020, BMJ: British Medical Journal (Online), 370

[34] Chen, N. et al. 2020, The Lancet, 395, 507

[35] Chen, W.-H., Strych, U., Hotez, P. J., & Bottazzi, M. E. 2020, Current tropical medicine reports, 1

[36] Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W.-C., Wang, C.-B., & Bernardini, S. 2020, Critical Reviews in Clinical Laboratory Sciences, 57, 365

[37] Danesh, G. et al. 2020, medRxiv

[38] Dastanka, A. A. 2015, Regional Formation and Development Studies, 16, 16
[39] Davies, N. G. et al. 2021, Science
[40] Deng, X. et al. 2020, Science, 369, 582
[41] Di Gennaro, F., Pizzol, D., Marotta, C., Antunes, M., Racalbuto, V., Veronese, N., & Smith, L. 2020, International journal of environmental research and public health, 17, 2690
[42] Domingo, E., & Holland, J. 1997, Annual review of microbiology, 51, 151
[43] du Plessis, L. et al. 2021, Science, 371, 708
[44] Edgar, R. C. 2004, Nucleic acids research, 32, 1792
[45] Ellington, S. et al. 2020, Morbidity and Mortality Weekly Report, 69, 769
[46] Fauver, J. R. et al. 2020, Cell
[47] Fokas, A., Dikaios, N., & Kastis, G. 2020, Journal of the Royal Society Interface, 17, 20200494
[48] Foundation, P. S. 2022, Python Language Reference, version 3.10.9, accessed: 2025-01-24
[49] Gankin, Y., Koniukhovskii, V., Nemira, A., Chowell, G., Weppelmann, T. A., Skums, P., & Kirpich, A. 2021, medRxiv
[50] Ganyani, T., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J., & Hens, N. 2020, Eurosurveillance, 25, 2000257
[51] Geidelberg, L. et al. 2021, Virus evolution, 7, veaa102
[52] Geoghegan, J. L. et al. 2020, Nature communications, 11, 1
[53] Ghinai, I. et al. 2020, The Lancet
[54] Gonzalez-Reiche, A. S. et al. 2020, Science, 369, 297
[55] Grassly, N. C., & Fraser, C. 2008, Nature Reviews Microbiology, 6, 477
[56] Hadfield, J. et al. 2018, Bioinformatics, 34, 4121
[57] Harris, C. R. et al. 2020, Nature, 585, 357
[58] Hasegawa, M., Kishino, H., & Yano, T.-a. 1985, Journal of molecular evolution, 22, 160
[59] Havers, F. P. et al. 2020, JAMA internal medicine, 180, 1576
[60] He, X. et al. 2020, Nature medicine, 26, 672
[61] Heled, J., & Bouckaert, R. R. 2013, BMC evolutionary biology, 13, 1
[62] Huang, C. et al. 2020, The lancet, 395, 497
[63] Islam, N. et al. 2020, bmj, 370
[64] Issanov, A., Amanbek, Y., Abbay, A., Adambekov, S., Aljofan, M., Kashkynbayev, A., & Gaipov, A. 2020
[65] Ito, J., Strange, A. P., Liu, W., Joas, G., Lytras, S., & Sato, K. 2024, A Protein Language Model for Exploring Viral Fitness Landscapes
[66] Iyengar, K., Mabrouk, A., Jain, V. K., Venkatesan, A., & Vaishya, R. 2020, Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14, 943
[67] Katoh, K., & Standley, D. M. 2013, Molecular biology and evolution, 30, 772
[68] Komissarov, A. B. et al. 2021, Nature communications, 12, 1
[69] Kyrychko, Y. N., Blyuss, K. B., & Brovchenko, I. 2020, Scientific reports, 10, 1
[70] Lai, A., Bergna, A., Acciarri, C., Galli, M., & Zehender, G. 2020, Journal of medical virology, 92, 675
[71] Lara, J., Teka, M., & Khudyakov, Y. 2017, BMC genomics, 18, 33

[72] Li, C., Yang, Y., & Ren, L. 2020, Infection, Genetics and Evolution, 104285
[73] Li, Q. et al. 2020, New England Journal of Medicine
[74] Lu, H., Stratton, C. W., & Tang, Y.-W. 2020, Journal of Medical Virology
[75] Mahase, E. 2020, BMJ
[76] McKinney, W. 2010, Proceedings of the 9th Python in Science Conference, 56
[77] Miller, D. et al. 2020, Nature communications, 11, 1
[78] Mohebbi, F., Zelikovsky, A., Mangul, S., Chowell, G., & Skums, P. 2024, Nature Communications, 15, 2838
[79] Mullard, A. 2020, Nature
[80] Nadeau, S. A., Vaughan, T. G., Sciré, J., Huisman, J. S., & Stadler, T. 2021, Proceedings of the National Academy of Sciences, 118
[81] National Center for Biotechnology Information (NCBI). 2025, Hepatitis C Virus Genome, Reference Sequence NC_004102.1, accessed: 2025-01-24
[82] Nesteruk, I. 2020, Europe PMC
[83] Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. 2015, Molecular biology and evolution, 32, 268
[84] Pantho, M. J., Bauder, L. A., Huang, S., & Qin, H. 2024, A Data-Driven Sliding-window Pairwise Comparative Approach for the Estimation of Transmission Fitness of SARS-CoV-2 Variants and Construction of the Evolution Fitness Landscape
[85] Pupko, T., Pe, I., Shamir, R., & Graur, D. 2000, Molecular biology and evolution, 17, 890
[86] Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. 2018, Systematic biology, 67, 901
[87] Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., & Pybus, O. G. 2020, Nature microbiology, 5, 1403
[88] Rambaut, A., Lam, T. T., Max Carvalho, L., & Pybus, O. G. 2016, Virus evolution, 2, vew007
[89] Recchi, E., Deutschmann, E., & Vespe, M. 2019, Robert Schuman Centre for Advanced Studies Research Paper No. RSCAS, 30
[90] Richardson, E., Malakhova, I., Novik, I., Famenka, A., & WHO. 2013, Belarus: health system review (World Health Organization. Regional Office for Europe)
[91] Roy, D., Tripathy, S., Kar, S. K., Sharma, N., Verma, S. K., & Kaushal, V. 2020, Asian Journal of Psychiatry, 102083
[92] Sagulenko, P., Puller, V., & Neher, R. A. 2018, Virus evolution, 4, vex042
[93] Salje, H. et al. 2020, Science
[94] Seemann, T. et al. 2020, Nature communications, 11, 1
[95] Shu, Y., & McCauley, J. 2017, Eurosurveillance, 22, 30494
[96] Silverman, J. D., Hupert, N., & Washburne, A. D. 2020, Science translational medicine, 12
[97] Skums, P. et al. 2012, in BMC bioinformatics, Vol. 13, Springer, 1–13
[98] Stadler, T., Kühnert, D., Bonhoeffer, S., & Drummond, A. J. 2013, Proceedings of the National Academy of Sciences, 110, 228
[99] Stringhini, S. et al. 2020, The Lancet, 396, 313
[100] Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. 2018, Virus evolution, 4, vey016

[101] Tizaoui, K. et al. 2020, International Journal of Biological Sciences, 16, 2906

[102] Vaughan, T. G., Leventhal, G. E., Rasmussen, D. A., Drummond, A. J., Welch, D., & Stadler, T. 2019, Molecular biology and evolution, 36, 1804

[103] Virtanen, P. et al. 2020, Nature Methods, 17, 261

[104] Wallinga, J., & Lipsitch, M. 2007, Proceedings of the Royal Society B: Biological Sciences, 274, 599

[105] Wei, W. E., Li, Z., Chiew, C. J., Yong, S. E., Toh, M. P., & Lee, V. J. 2020, Morbidity and Mortality Weekly Report, 69, 411

[106] White, S., McAllister, I., & Feklyunina, V. 2010, The British journal of politics and international relations, 12, 344

[107] Wu, H., Huang, Y., Dykes, C., Liu, D., Ma, J., Perelson, A., & Demeter, L. 2006, Modeling and Estimation of Replication Fitness of Human Immunodeficiency Virus Type 1 In Vitro Experiments by Using a Growth Competition Assay

[108] ——. 2006, Modeling and Estimation of Replication Fitness of Human Immunodeficiency Virus Type 1 In Vitro Experiments by Using a Growth Competition Assay

[109] Wu, S. L. et al. 2020, Nature communications, 11, 1

[110] Yamamoto, N., & Bauer, G. 2020, Medical Hypotheses, 110160

[111] Yi, H. 2020, Clinical Infectious Diseases

[112] Zhu, N. et al. 2020, New England Journal of Medicine