Summer 8-31-2023

Computational Methods for Assessment and Prediction of Viral Evolutionary and Epidemiological Dynamics

Fatemeh Mohebbi
Georgia State University

Computational Methods for Assessment and Prediction of Viral Evolutionary and Epidemiological Dynamics

by

Fatemeh Mohebbi

Under the Direction of Pavel Skums, 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

2023

ABSTRACT

The ability to comprehend the dynamics of viruses' transmission and their evolution, even to a limited extent, can significantly enhance our capacity to predict and control the spread of infectious diseases. An example of such significance is COVID-19 caused by the severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2). In this dissertation, I am proposing computational models that present more precise and comprehensive approaches in viral outbreak investigations and epidemiology, providing invaluable insights into the transmission dynamics, and potential interventions of infectious diseases by facilitating the timely detection of viral variants. The first model is a mathematical framework based on population dynamics for the calculation of a numerical measure of the fitness of SARS-CoV-2 subtypes. The second model I propose here is a transmissibility estimation method based on a Bayesian approach to calculate the most likely fitness landscape for SARS-CoV-2 using a generalized logistic sub-epidemic model. Using the proposed model I estimate the epistatic interaction networks of spike protein in SARS-CoV-2. Based on the community structure of these epistatic networks, I propose a computational framework that predicts emerging haplotypes of SARS-CoV-2 with altered transmissibility. The last method proposed in this dissertation is a maximum likelihood framework that integrates phylogenetic and random graph models to accurately infer transmission networks without requiring case-specific data.

INDEX WORDS: SARS-CoV-2, Fitness, Transmissibility, Sub-epidemic model, Bayesian inference, Genomic surveillance, Haplotype forecasting, Epistasis, Network community, Genomic epidemiology, Transmission network, Maximum likelihood inference

Computational Methods for Assessment and Prediction of Viral Evolutionary and Epidemiological Dynamics

by

Fatemeh Mohebbi

Committee Chair:	Pavel Skums
Committee:	Alex Zelikovsky
	Murray Patterson
	Artem Rogovskyy

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2023

DEDICATION

To my family and friends.

ACKNOWLEDGMENTS

I extend my sincere gratitude to my advisor, Dr. Pavel Skums, for creating a supportive research environment. Their unwavering guidance, mentorship, and continuous support have been vital throughout my Ph.D. journey. I am honored to have Dr. Alex Zelikovsky, Dr. Murray Patterson, and Dr. Artem Rogovskyy as my committee members, whose guidance, expertise, and support have been invaluable in completing my Ph.D. research.

I am grateful for the exceptional group of peers and colleagues in the lab with whom I had the privilege of working. I would like to express my thanks to Dr. Pelin Icer Baykal, Dr. Andrew Melnyk, Dr. Viachaslau Tsyvina, Dr. Sergey Knyazev, Dr. Kiril Kuzmin, Hossein Saghaeiannejad, Dr. Filipp Rondel, Akshay Juyal, and Alina Nemira for their contributions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Severe acute respiratory coronavirus syndrome 2 (SARS-CoV-2), which jumped into the human population from an uncharacterized animal reservoir in late 2019, caused Coronavirus disease 2019 (COVID-19). The virus has gradually accumulated mutations leading to new strains of SARS-CoV-2. So far five of these SARS-CoV-2 strains were labeled by the World Health Organization as "variants of concern": the Alpha, Beta, Gamma, Delta, and Omicron variants. COVID-19 was an example of a global pandemic where respiratory virus infections resulted in substantial morbidity and mortality as well as economic losses. The transmission mechanism and how easily respiratory viruses spread (transmissibility) differ between respiratory viruses belonging to different families but they can also be varied within a single family. It has been shown that some of the mutations that occurred in SARS-CoV-2 such as D614G and P681R are associated with increased transmissibility and virulence of the variants[111,203].

Data from surveillance and observational epidemiological studies are usually used to estimate transmissibility by estimating the basic reproduction number ($R0$). The basic reproduction number ($R0$) also known as the basic reproductive rate is defined as the average number of successful transmissions per infected individual in a population at the start of an epidemic. In more precise terms, $R0$ represents the average number of secondary infections caused by a primary infection[49]. A mathematical or statistical model is often used to estimate the transmissibility of a respiratory virus in a population, especially during pandemics. In this dissertation, I propose more efficient and accurate fitness (growth rate) and reproduction number, and therefore transmissibility esti-

mation models for SARS-CoV-2 lineages (mutations) and subtypes. Then I use these models to predict the potential epistasis mutation networks of the SARS-CoV-2 spike protein. The community structure of epistatic networks within the SARS-CoV-2 spike protein offers a promising avenue for efficiently detecting or predicting emerging haplotypes with modified transmissibility. Notably, dense network communities related to these haplotypes become observable significantly earlier than the prevalence of the corresponding viral variants. Here I propose a computational framework that leverages this observation. This model identifies highly connected communities of SAV alleles and merges them into haplotypes that accurately predicted known SARS-CoV-2 variants of concern (VOCs) and variants of interest (VOIs) months before they became noticeably prevalent.

The last method proposed in this dissertation is a modeling and algorithmic framework to infer viral transmission networks from genomic data by integrating phylogenetic and random graph models. In this model, the social component of epidemics is considered by estimating the probability that sampled networks are subgraphs of a random contact network (social networks of contacts between individuals at risk) and summarizing them accordingly into a consensus network.

1.1 Assessment of the fitness landscape and transmissibility of SARS-CoV-2 variants and subtypes

According to population genetics theory, the majority of mutations are neutral[136], but some may be advantageous or deleterious. Mutations that are highly deleterious will be rapidly removed from the population; mutations that are only mildly deleterious may be retained, if only temporarily. While neutral mutations, and especially advantageous mutations, can reach higher frequencies.

The early detection of such mutations could potentially prove useful in controlling the COVID-19 pandemic. However, it can be difficult to distinguish neutral mutations from advantageous mutations that directly increase the virus' transmission.

It has been demonstrated that phylogenetic-tree-based analyses can lead to overinterpretation in studying SARS-CoV-2 genomic variations and transmissibility[115], and therefore a quantitative assessment based on an epidemiological and evolutionary modeling framework is necessary.

Here, I propose a model for the calculation of a numerical measure of the fitness (growth rate) of SARS-CoV-2 subtypes and lineages by adapting a measure of selective fitness which originally was introduced for calculating differential interferon resistance coefficients for quasispecies using HCV sequence data[169]. The fitness coefficient is an assessment of the selective fitness of a subtype, based on the number of sequences in the corresponding lineage, and the rate at which it grows over time.

Then I explore and examine a reproduction number estimation method based on a Bayesian approach using a generalized logistic sub-epidemic model[32] which was previously used for epidemiological forecasting of SARS-CoV-2 successfully[157]. Considering each variant of SARS-CoV-2 as a subpopulation, we calculate the most likely fitness landscape for each subpopulation using the generalized logistic sub-epidemic model. Using the proposed model we evaluate the transmissibility of SARS-CoV-2 variants, the obtained values are consistent with other studies[105,43]. For each pair of mutations in SARS-CoV-2, our Bayesian model is used to assess epistasis interaction and then build the epistatic networks of the spike protein.

1.1.1 Problem formulation

This chapter addresses the following problem:

- Given nucleotide sequences of SARS-CoV-2 and the corresponding metadata including the sequences' collection dates:

  (i) Estimate fitness/transmissibility of SARS-CoV-2 variants/lineages and subtypes.

  (ii) Estimate the epistatic interaction networks of SARS-CoV-2.

1.2 Prediction of emerging variants of SARS-CoV-2 with altered phenotypes

Epistasis occurs when a mutation's phenotypic effect is dependent on the presence of other mutations in the genome. The genomes of RNA viruses display complex patterns of epistatic interactions within and between genes despite their structural simplicity. These pathogens' evolutionary dynamics are profoundly affected by such complex patterns[60].

Moreover, these interactions determine the complexity of the genotype-fitness landscape[54]. Depending on prior substitutions, epistasis can influence the order in which mutations can occur[174]. Due to their combinatorial nature, epistatic interactions constitute a powerful influence at the population level, determining the long-term evolutionary trajectory of evolving populations[54]. The combinatorial nature of epistatic interactions makes an exhaustive laboratory exploration difficult and therefore limits our ability to predict long-term evolution[160].

An essential component of the infectivity of SARS-CoV-2 is the spike protein, a homotrimeric glycoprotein complex encoded by the S-gene. In the exposed regions of the spike protein, a large number of mutations have led to variants that have a higher affinity for the human ACE2 receptor,

are more transmissible, and are less neutralizing to antibodies[3,113,142]. I examine the mutation pairs in spike protein in SARS-CoV-2 and build epistatic/coordinated substitution networks. An epistatic network $\mathcal{G}$ is defined as a graph with nodes representing SAVs (single amino acid variations), and two nodes being adjacent whenever the corresponding non-reference alleles are simultaneously observed more frequently than expected by chance. I propose a novel computational framework that predicts haplotypes of SARS-CoV-2 with altered phenotypes based on analyzing dense communities of the epistatic networks of the spike protein. This approach which is called HELEN (Heralding Emerging Lineages in Epistatic Networks), was validated by accurately identifying known SARS-CoV-2 VOCs and VOIs up to 10-12 months before they reached high prevalences and were designated by the WHO.

1.2.1 Problem formulation

This chapter addresses the following problem:

- Given nucleotide sequences of SARS-CoV-2 and the corresponding metadata including the sequences' collection dates.
- Predict the haplotypes of SARS-CoV-2 with altered phenotypes.

1.3 Viral outbreak investigation and transmission history reconstruction

Genomic epidemiology, which involves analyzing viral genomes to understand how viruses spread and evolve, has become a vital tool for investigating outbreaks and tracking transmission dynamics[6,18]. The development of efficient computational methods has enabled the rapid progress of

genomic epidemiology, leading to the creation of transmission history inference tools such as Outbreaker and Outbreaker 2[91,25], SCOTTI[46], SeqTrack[92], SCOTTI[46], Phybreak[95], and more[95,195,47,196,45,53,81,171,170,74,102,112,24,162,163,51,200,121,52,122,41,26,80]. These tools have successfully been applied to various viruses, including SARS, MERS, and SARS-CoV-2[191,150,205,146,99,23].

The extremely high genomic diversity of viruses resulting from their error-prone replication means that each infected individual typically hosts a heterogeneous population of numerous genomic variants, known as viral quasispecies. The first generation of transmission inference methods largely ignored intra-host viral diversity, only considering a single sequence per host. Later, it was demonstrated that taking intra-host diversity into account greatly enhances the predictive power of transmission inference algorithms, allowing for the detection of viral evolution directionality in situations where reliable phylogenetic rooting is not possible[196,170,5,156,99]. Several tools have been developed specifically to address this issue, including TNeT[51], TiTUS[163], SharpTNI[162], and BadTrIP[47].

Despite the progress made in the development of transmission inference methods, there are still several computational, modeling, and algorithmic challenges that need to be addressed. These challenges include the use of maximum parsimony principles, while maximum likelihood or Bayesian models can be used for more accurate reconstruction of transmission links[125]. Using genomic data alone. However, genomic data alone are not able to reverse transmission network ambiguities in many cases, requiring additional evidence[89,182,91]. And the assumption of independent transmission network edges which means any person can infect any other person with the same probability. Yet, this is not always the case[71]. We propose a maximum likelihood transmission networks infer-

ence framework, SOPHIE (SOcial and PHilogenetic Investigation of Epidemics), that overcomes these challenges by combining phylogenetic and random graph models. SOPHIE samples from the joint distribution of phylogeny ancestral traits defining transmission networks, and estimates the probabilities that sampled networks are subgraphs of a random contact network and summarize them accordingly into the consensus network. This approach is scalable, accounts for intra-host diversity, and accurately infer transmissions without case-specific epidemiological data. We applied SOPHIE to synthetic data simulated under different epidemiological and evolutionary scenarios, as well as to experimental data from epidemiologically curated HCV outbreaks. The experiments confirm the effectiveness of this methodology.

1.3.1 Problem formulation

This chapter addresses the following problem:

- Given a time-labelled phylogeny $T = (V(T), E(T))$ with n_l leafs corresponding to viral haplotypes sampled from n_h infected hosts; each leaf u is assigned the label $\lambda_u \in [n_h]$ corresponding to its host.
- Estimate a transmission network.

1.4 Contributions

This dissertation discusses the following contributions:

- Estimation of a fitness coefficient for SARS-CoV-2 subtypes based on a measure of selective fitness which originally was introduced for calculating drug resistance coefficients for quasispecies in HCV.

- Designing a novel method that estimates the reproduction number of SARS-CoV-2 variants/lineages. This model is based on a Bayesian inference using a so-called generalized logistic sub-epidemic framework, which is a growth model for forecasting epidemic trajectories. We model SARS-CoV-2 variants as overlapping sub-epidemics (dividing the whole population into the variants and the wild type). Using this model we estimate the transmissibility of each sub-epidemic.

- Estimating the fitness coefficient and transmissibility of SARS-CoV-2 variants and lineages.

- Designing a novel computational framework that predicts the haplotypes of SARS-CoV-2 with altered phenotype. It identifies densely connected communities of SAV alleles and merges them into haplotypes using a combination of statistical inference, population genetics, and discrete optimization techniques.

- Designing a maximum likelihood framework based on the integration of phylogenetic and random graph models. It infers transmission networks from viral phylogenies and expected properties of inter-host social networks modeled as random graphs with given expected degree distributions.

- Discussing validation of estimated values and obtained results.

1.5 Refereed Journal Articles

3. Melnyk, A., **Mohebbi, F.**, Knyazev, S., Sahoo, B., Hosseini, R., Skums, P., ... & Patterson, M. (2021). From alpha to zeta: Identifying variants and subtypes of sars-cov-2 via clustering. **Journal of Computational Biology**, 28(11), 1113-1129.

2. Skums, P., **Mohebbi, F.**, Tsyvina, V., Baykal, P. I., Nemira, A., Ramachandran, S., & Khudyakov, Y. (2022). SOPHIE: Viral outbreak investigation and transmission history reconstruction in a joint phylogenetic and network theory framework. **Cell Systems**, 13(10), 844-856.

1. Batool, M., Hillhouse, A. E., Ionov, Y., Kochan, K. J., **Mohebbi, F.**, Stoica, G., ... & Rogovskyy, A. S. (2019). New Zealand White rabbits effectively clear Borrelia burgdorferi B31 despite the bacterium's functional vlsE antigenic variation system. **Infection and immunity**, 87(7), e00164-19.

1.6 Refereed Articles in Conference Proceedings

1. Melnyk, A., **Mohebbi, F.**, Knyazev, S., Sahoo, B., Hosseini, R., Skums, P., ... & Patterson, M. (2020, December). Clustering based identification of SARS-CoV-2 subtypes. In International Conference on Computational Advances in Bio and Medical Sciences (pp. 127-141). Springer, Cham.

1.7 Invited Talks

1. **Mohebbi, F.**, Skums, P., Chowell, G. Bayesian assessment of the fitness landscape and epistatic interaction network of SARS-CoV-2. 9th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), 2021

1.8 Not submitted yet

1. **Mohebbi, F.**, Zelikovsky, A., Mangul, S., Chowell, G., & Skums, P. (2023). Community structure and temporal dynamics of SARS-CoV-2 epistatic network allows for early detection of emerging variants with altered phenotypes. bioRxiv, 2023-04.

CHAPTER 2

Assessment of the fitness landscape and transmissibility of SARS-CoV-2 variants and subtypes

As the COVID-19 pandemic continues its unabated global spread, it is critical to monitor the relative transmissibility, severity, and resistance to pharmaceutical interventions of novel variants of the coronavirus pandemic.

The lineage B.1.1.7 is defined by a specific family of 17 non-synonymous SNVs, that includes 8 SNVs in the spike protein[147]. Likewise, lineage B.1.617.2 includes 8 SNVs in the spike protein[94]. The majority of these genomic alterations are known to have phenotypic effects[175,82]. The spike mutation(s) such as P681R are responsible for the enhanced replication fitness of lineage B.1.617.2 over B.1.1.7.[109] As well, the rapid lineage growth prompted indicates that the variants of B.1.1.7 and B.1.617.2 have altered transmissibility and a higher fitness with respect to other SARS-CoV-2 subpopulations[147,105].

Currently measuring the transmissibility and fitness of a variant can be done only when a lineage is large enough in the population which can be too late for controlling the spread of a variant. Therefore, a quantitative assessment based on a sound epidemiological and evolutionary modeling framework is required to evaluate the viruses' fitness and transmissibility[105]. In this chapter, I am proposing two models for measuring the fitness of SARS-CoV-2 haplotypes.

The first model is based on the quasispecies model for measuring differential interferon (INF) resistance of HCV quasispecies presented in[169]. The INF-resistance method is built on an analysis of HCV population dynamics considering the relative frequencies of variants during the first hours of interferon therapy, at a set of observed time points. Where the abundance of different viral vari-

ants is formulated as a system of differential equations based on the death rate and the replication rate of each variant. Then the evolution of frequencies of quasispecies is modeled using this system of differential equations. The frequencies are approximated using cubic spline approximation at each time point. They define fitness functions $g_i(t)$ based on the approximated frequencies and approximated titer which reflect changes of quasispecies fitness under the selection pressure.

Since this framework agrees with the standard population genetic models, we applied this mathematical model to calculate a measure of *fitness* coefficient of SARS-CoV-2 variants and lineages. In Section 2.1 this adaptation is explained in detail. We estimated the fitness coefficients of SARS-CoV-2 variants and subtypes. The subtypes were obtained by clustering sequences into groups using a CliqueSNV-based clustering model (Section 2.1) in order to identify novel variants and subtypes of SARS-CoV-2. The fitness coefficients agree with R_0 value of corresponding variants and other validation metrics used in this study[119,120], which indicates the accuracy of our estimation model.

We also propose a Bayesian assessment for the fitness of SARS-CoV-2 variants and lineages based on a generalized logistic sub-epidemic model[32]. This model which supports various epidemic wave trajectories, has been successfully used for SARS-CoV-2 epidemiological forecasting[157,158] and proved to be accurate and reliable. This framework tries to capture the heterogeneity of population and temporal changes that shape the incidence curves of the larger-scale epidemic wave patterns. This is achieved by dividing the population into overlapping sub-epidemics. Each sub-epidemic is modeled by a generalized logistic growth model (GLM) (2.1) which has been used for short-term forecasting of the trajectory of infectious disease outbreaks effectively[31,140]. GLM

equation is as follows:

$$\frac{dC(t)}{dt} = rC^p(1 - \frac{C(t)}{K_0}) \tag{2.1}$$

Where $\frac{dC(t)}{dt}$, $C(t)$, r, $p \in [0, 1]$, K_0 respectively denote the incidence curve of a sub-epidemic over time t, the cumulative number of cases at time t, the growth rate, the scaling of growth parameter, the final epidemic size. In Section 2.2 the Bayesian estimation of SARS-CoV-2's fitness and transmissibility using GLM differential equations is discussed in detail.

We also, use this method to find epistasis mutation pairs in spike protein and estimate the epistasis networks for different countries. Epistasis is defined by an interaction of genetic variation in which the effect of a gene mutation depends on mutations in one or more other genes. Epistatic analysis is paramount for drug and vaccine development, and prediction of many evolutionary hypotheses[21]. Some studies already tried to predict epistasis interactions of SARS-CoV-2. In[152], a global phylogenetic tree was constructed over a global alignment of the whole genomes. They found out more than 100 nonsynonymous mutations appeared multiple times on around 200 or more terminal branches and leaves in the phylogenetic tree. Which claimed to be evidence of positive selections. An epistasis network was inferred based on the co-occurrence network of these positively selected residues. The inferred epistasis interactions are mostly located in the receptor-binding domain (RBD) of the spike protein and the region of the nucleocapsid protein. No quantitative methods are suggested for epistasis interaction estimation, and the inferred network is only based on the co-occurrence of mutations in the tree.

Another study[201] used in vitro binding measurements to predict epistatic mutations in SARS-CoV-2. In vitro evolution which is an experimental method for screening of large random-sequence

libraries, was applied to select for higher affinity binding of the SARS-CoV-2 spike RBD to the host cell receptor angiotensin-converting enzyme 2 (ACE2). They suggest that the spike mutations such as S477N, E484K, and N501Y are positively correlated with increased binding affinity to ACE2. Consequently, they are responsible for more transmissibility of viruses. As well, they identified mutations N501Y and Q498R as epistatic pair.

A Genome-wide epistasis analysis over a global alignment is proposed in[202] for epistasis prediction. Pseudo-likelihood maximization for direct coupling analysis (DCA) is utilized to infer epistatic interactions. DCA is a method to extract approximate information about coevolving residues in a protein family from data. Since DCA is time-consuming for such a large dataset, only the top 200 mutation pairs regarding pseudo-likelihood maximization score are considered for epistatic analysis. Out of 200 pairs, 8 pairs are reported as potential epistatic which are located in genes ORF3a, ORF8, nsp2, nsp6, nsp12, nsp13, and nsp14.

In this study, we use our reliable fitness assessment to evaluate the fitness of mutation pairs that occurred both independently and concurrently in the population (Section 2.2.3). If a pair of mutations occurring together enhance or decline the fitness of the subpopulation over independent occurrences of individual mutations, it is reported as a positive or a negative epistasis pair respectively. If it did not affect fitness it is considered additive epistasis.

Variant	Region	Lineage	S/Gen.	Source
Gamma	Brazil	P.1(B.1.1.28.1)	10 / 21	127
Zeta	Brazil	P.2(B.1.1.28.2)	1 / 5	
Epsilon	California	B.1.427/B.1.429	3 / 5	204
Iota	New York	B.1.526	6 / 16	192
Beta	S. Africa	B.1.351	9 / 21	70
Alpha	UK	B.1.1.7	8 / 17	
Kappa	India	B.1.167.1	8 / 17	197
Delta	India	B.1.167.2	8 / 17	

Table 2.1 Some known variants of SARS-CoV-2. The five columns, starting from the left, are: Variant (Greek name); Region where it was first identified; PANGOLIN Lineage identifier; Number of mutations on the S gene / entire genome; and Source.

2.1 Fitness coefficient estimation of SARS-CoV-2 subtypes

2.1.1 Datasets

In this section, we outline the datasets that we used in the fitness coefficient estimation of SARS-CoV-2 subtypes. We first give a brief overview of subtypes, or *variants* we study here, and then describe the datasets we use, which are known to contain different proportions of these variants.

2.1.1.1 Known variants

Since its emergence in November 2019[48], SARS-CoV-2 has evolved into different variants. Divergences in mutation at the genomic level have been observed in different regions of the world as new infectious variants are emerging. The following is a description of some of the well-known variants to date. A more complete list can be found in Table 2.1.

Alpha variant (UK) The Alpha variant, also known as the B.1.1.7 variant of SARS-CoV-2 was first identified in Kent, UK, in late summer to early autumn 2020. The lineage was the most transmissible of all those that appeared before, with a 50% to 100% reproduction rate[185]. The first

case was reported on December 14, 2020, and this variant is now detected in over 30 countries, with more than 15 thousand people affected worldwide[70]. Of the many genomic mutations that characterize this variant, it has a 69/70 deletion and a mutation at position 501, which affects the conformation of the receptor binding domain (RBD) of the spike protein of SARS-CoV-2. It has 17 mutations which include 14 amino acids and 3 in-frame deletions at open-reading frame (ORF) 1 a/b, ORF 8, spike (S), and N gene regions. These mutations have biological implications and have resulted in diagnostic failures[148].

Beta variant (South Africa) The first case of the Beta variant, also known as B.1.351, was identified in Nelson Mandela Bay, South Africa, in October 2020. This lineage was predominant by the end of November 2020 in the Eastern and Western Cape Provinces of South Africa. By January 2021, there were 415 known cases of infection with this variant, were found in 13 different countries. This variant has eight mutations in the S gene region, including three mutations SK417N, E484K, and N501Y that affect the RBD of the spike protein. These three mutations can be the reason for increased transmissibility, and can also lead to alterations in conformation that could pose a challenge for the effectiveness of vaccines[70,207,178].

Gamma and Zeta variants (Brazil) The Gamma variant, also known as P.1(B.1.1.28.1), was initially identified in February 2020, in Japanese travelers coming from Amazonas State, Brazil. It was first reported in a 29-year-old female resident of Amazonas State. The P.1 lineage has mutations K417T, E484K, and N501Y in the S gene region, which affects the RBD of the spike protein. The Zeta variant, also known as P.2(B.1.1.28.2) was first identified in Rio de Janeiro, Brazil. It shares the mutation E484K with the Gamma variant[126].

Dataset	Database	Start	End	No. Sequences
UK	EMBL-EBI	2020-01-29	2020-12-29	88 008
GISAID 2	GISAID	2019-12-24	2021-04-04	1 000 982

Table 2.2 The datasets that are used in the experiments of Section 2.1.3. The five columns, starting from the left, are: Name we use here; Database it is from (GISAID[59] or EMBL-EBI[61]); Earliest collection date of any sequence; Latest collection date; and number of sequences.

Epsilon variants (California, USA) In July 2020, the first case of the Epsilon variants, also known as the CAL.20C or B.1.427/B.1.429 variants of SARS-CoV-2, was identified in Los Angeles County, California, USA. The Cedars-Sinai Medical Center (CSMC) reported that the second B.1.429 Epsilon variant contains five mutations at ORF 1 an (I4205V), ORF 1 b (D1183Y), and S gene mutations S13I, W152C and L452R. Mutation L452R is correlated with higher infectivity[204]. The Epsilon variants are spreading in the US and in 29 other countries[118].

2.1.1.2 Datasets

We use two different datasets, which are summarized in Table 2.2, and then each one is explained in more detail in its corresponding subsection below.

UK The Uk data set consists of sequences submitted to the EMBL-EBI[61] database from the end of January 2020 to the end of December 2020. Since this database is in the UK, and given the collection period, this dataset contains a sizeable portion of the Alpha variant.

GISAID 2 The second data set consists of all sequences submitted to GISAID up until April 2021. Since many of the known variants mentioned above have been well-documented by April 2021, this dataset contains a sizable portion of sequences annotated as being from the Alpha, Beta, Gamma, Epsilon and Zeta variants. Such labels correspond to "ground truth clusters" for which we can compute the precision, specificity, F_1 score, etc., of clustering obtained with a given method.

2.1.2 Methods

2.1.2.1 Fitness

An Interferon (IFN) resistance coefficient model is adapted here to calculate the fitness of SARS-CoV-2 subtypes, based on how the rate of change in size (number of sequences it contains) varies over time. For a set $C_1, \ldots, C_k$ of clusters, $X_i(t)$ denotes the size of subtype or cluster C_i at a particular time t. The fitness coefficient is calculated using h_i, which is the cumulative sum of the X_i. It follows that $h(t) = \sum_{i=1}^{k} h_i(t)$ is the total infected population size at time t. Each $h_i(t)$ is normalized over $h(t)$, which is denoted by $u_i(t)$, that is,

$$u_i(t) = \frac{h_i(t)}{\sum_{i=1}^{k} h_i(t)} \ . \tag{2.2}$$

Using cubic splines, $u_i(t)$ and $h(t)$ are interpolated over the time period and the derivatives $\dot{u}_i(t)$ and $\dot{h}(t)$ are calculated. The *fitness function* g_i, for each cluster C_i is then defined as

$$g_i(t) = \frac{\dot{u}_i(t)}{u_i(t)} + \frac{\dot{h}(t)}{h(t)} \ . \tag{2.3}$$

The *fitness coefficient* r_i, which is the average fitness over the time period T (composed of the times t) for cluster C_i is then

$$r_i = \frac{1}{T} \int_1^T g_i(t) dt \ . \tag{2.4}$$

In order to reduce sampling error, we use the Poisson distribution to draw random samples. For each cluster at time t, a sufficiently large number of random samples are drawn from the Poisson distribution on $X_i(t)$ as the expectation of the interval. Then $X_i(t)$ is replaced by the mean value of these random samples. This is repeated a sufficiently large number of times (*e.g.*, 100) to cal-

culate a set of Poisson-distributed sizes. The fitness coefficient calculation is then applied on each repetition separately and a confidence interval of the fitness coefficient is obtained.

This framework is tested on SARS-CoV-2 variants (alpha and delta) where it correctly estimated the growth rates and the R_0 values calculated based on the growth rates. Then we applied it on clusters obtained by clustering nucleotide sequences of the SARS-CoV-2 virus with a CliqueSNV-based clustering method, which is explained in the following.

2.1.2.2 CliqueSNV

We are clustering viral sequences in order to identify subtypes. The idea is that we use, CliqueSNV to find haplotypes in the massively interhost viral population, using them as cluster centers in categorical clustering algorithms such as k-modes (Huang, 1997) to find subtypes. We propose to use currently existing tools that were developed to identify subtypes in intra-host viral populations from next-generation sequencing (NGS) data reviewed in[98], such as Savage[11], PredictHaplo[144], aBayesQR[2], *etc.* However, our setting is slightly different, where the data consists of large collections of *inter-host* consensus sequences gathered from different regions and countries around the world[59,61]. We expect, however, that such tools are appropriate at this scale: now the "host" is an entire region or country, and we reconstruct the subtypes, or variants, and their dynamics within these regions or countries. The SARS-CoV-2 sequences in GISAID are consensus sequences of approximate length 30K. Such sequences by quality and length have similar properties as PacBio reads. We choose CliqueSNV since it performed very well on PacBio reads[96].

2.1.2.3 k-modes Clustering

We also considered known general techniques for clustering from the literature as a baseline for comparison. Since we are clustering sequences, which are on the *categories* A, C, G, T (and –, a gap), we chose k-modes[84,85] for this purpose. This approach is almost identical to k-means[4,116], but it is based on the notion of *mode* (rather than Euclidean mean), making it appropriate for clustering categorical data. Indeed, the Euclidean mean of three nucleotides has little meaning in this context, and may not even be well-defined. An example of the latter is when the "distance" from A to G is different than from G to A. A similar observation was made in the context cancer mutation profiles[37] in the form of absence/presence information. Treating these as categories in using k-modes (rather than as 0's and 1's in using k-means) resulted in a clustering approach[35] that, when used as a preprocessing step, allowed cancer phylogeny building methods to attain a higher accuracy[36], and in some cases with much lower runtimes[88].

The *mode* q of a cluster C of sequences is another "sequence" (on A, C, G, T, –) which minimizes

$$D(C, q) = \sum_{s \in C} d(s, q) \ , \tag{2.5}$$

where d is some dissimilarity measure (such as Hamming distance) between the sequences we are considering. Note that q is not necessarily an element of C. Aside from finding the mode instead of the Euclidean mean, the k-modes algorithm operates similarly to k-means, following the same iteration:

1. Initialize cluster centers (or centroids);

2. Assign each sequence to the closest center based on dissimilarity d;

3. For each cluster resulting from this assignment, find its (new) center (2.5); and

4. Return to step 2. until convergence (clusters do not change between 2. and 3.).

In this work, we use k-modes. We first initialize cluster centers (1.) by using the centers (the subtypes) that were found by CliqueSNV.

Then, the dissimilarity d that we use is either the (i) Hamming distance, or (ii) TN-93 distance[177].

2.1.2.4 Clustering entropy

Because of the lack of ground truth, we need to consider an *internal* evaluation criteria. The clustering entropy[106] (2.8 and 2.9) is an internal evaluation criterion that was shown to generalize any distance-based criterion, and does not even require any notion of distance or dissimilarity. Since sequences are objects on categorical attributes which take values A, C, G, T (and –, a gap), the clustering entropy criterion is appropriate in our case. Moreover, clustering entropy naturally reflects the fact that the population of viral sequences comes from a number of subtypes. Clustering entropy can be formally derived using a likelihood principle based on Bernoulli mixture models. In these mixture models, the observed data are thought of as coming from a number of different latent classes. In[106], the authors prove that minimizing clustering entropy is equivalent to maximizing the likelihood that set of objects are generated from a set of (k) classes. This reflects the underlying processes which generate a set of viral sequences: that they evolved from a set of (k) subtypes.

This relates closely to the widely-used notion of *sequence logo*[164]: a graphical representation of a set of aligned sequences which conveys at each position both the relative frequency of each base (or residue), and the amount of information (the entropy) in bits. A clustering of viral sequences of low entropy then relates to a reliable set of sequence logos (in terms of information), and can hence shed light on the possible biological function of the viral subtype that each such logo (or related motif) represents.

2.1.2.5 Monte-Carlo Based Entropy Minimization

We use clustering entropy[106] to assess the clustering method. For this reason, we also employ a technique aimed directly at minimizing clustering entropy as the objective. We first define clustering entropy in the following.

Formally, we have a set S of aligned nucleotide sequences on the set X of genomic sites. Since they are aligned, sequences can be viewed as rows of a matrix and, when restricted to a site $x \in X$, can be viewed as columns of this matrix. Let $\mathcal{N} = \{\text{A, C, G, T}\}$ be the four nucleotides, not counting the gap (–) character. Using the notation of[106], the entropy $\hat{H}_x(C)$ of a subset C (a cluster) of sequences from S at site $x \in X$ is then

$$\hat{H}_x(C) = -\sum_{s \in C} \sum_{a \in \mathcal{N}} p_x(s = a) \log p_x(s = a) \ . \tag{2.6}$$

Note that $p_x(s = a)$ — the probability that a sequence $s \in C$ has nucleotide a at site x — essentially amounts to the relative *frequency* of nucleotide $a \in \mathcal{N}$ in C at site x. The entropy

$\hat{H}_X(C)$ of subset C of sequences on a subset X of sites is then

$$\hat{H}_X(C) = \sum_{x \in X} \hat{H}(x) \ , \tag{2.7}$$

that is, we simply sum up the entropies at the individual sites. Since the set of sites will always correspond to the SNV sites of our sequences, we will use simply $\hat{H}(C)$ for the entropy of a subset (a cluster) of sequences from hereon in. The *expected* entropy[106] of a clustering $\mathbb{C} = C_1, \ldots, C_k$ of sequences is then

$$H(\mathbb{C}) = \frac{1}{n} \sum_{i=1}^{k} n_i \hat{H}(C_i) \ , \tag{2.8}$$

where $n_i = |C_i|$, the number of elements in cluster C_i, and n is the total number of sequences. For completeness, the *total* entropy of a clustering is simply the sum

$$T(\mathbb{C}) = \sum_{i=1}^{k} \hat{H}(C_i) \tag{2.9}$$

of the individual entropies of each cluster (not weighted by n_i).

In[106], the authors prove that the entropy 2.8 is a convex function, allowing any optimization procedure to reach a global minimum. It is because of this property that we can use techniques aimed directly at minimizing clustering entropy as the objective. The Monte-Carlo method is a broad class of computational algorithms that rely on repeated random sampling to optimize some criterion. In this context, we are randomly sampling clusterings of sequences in order to minimize 2.8. The basic idea is that we start with some clustering — note that the clustering corresponding to placing all sequences in the same cluster has maximum entropy, by definition. The Monte-Carlo process then operates according to the iteration:

- from the current clustering, randomly pick a sequence from some cluster and place it into another cluster, resulting in a new clustering;
- compute the entropy (2.8) of the new clustering; and
- accept this new clustering, if the entropy has decreased, otherwise keep the current clustering;

until convergence, *i.e.*, the clustering does not change after some number θ of iterations.

In [106], the authors prove the concept of applying the Monte-Carlo method to entropy minimization by implementing a very basic procedure similar to the above, and then demonstrate it on a small dataset. Since our datasets are on a much larger scale (millions of sequences on 30K genomic sites), the basic iteration which randomly samples a single sequence in each iteration would need many iterations for a very small improvement. For this reason, we apply the following preprocessing step, to improve the convergence. Rather than using all (30K) columns, we first sort the columns according to their (unclustered) entropy value. We then select the n columns, or *tags*, with highest entropy. Next, we then run the above Monte-Carlo process on the reduced dataset with the n tags. This results in a clustering (of the rows), to which we then apply to the original set of all columns.

2.1.2.6 Filling gaps

Finally, the set of SARS-CoV-2 sequences that we deal with contains missing nucleotides, due to gaps or deletions. This is particularly true with GISAID sequences collected from December 2019 to the end of March 2020, when sequencing, alignment, *etc.*, were less refined. This is further complicated by the presence of deletions, which could be confused with gaps.

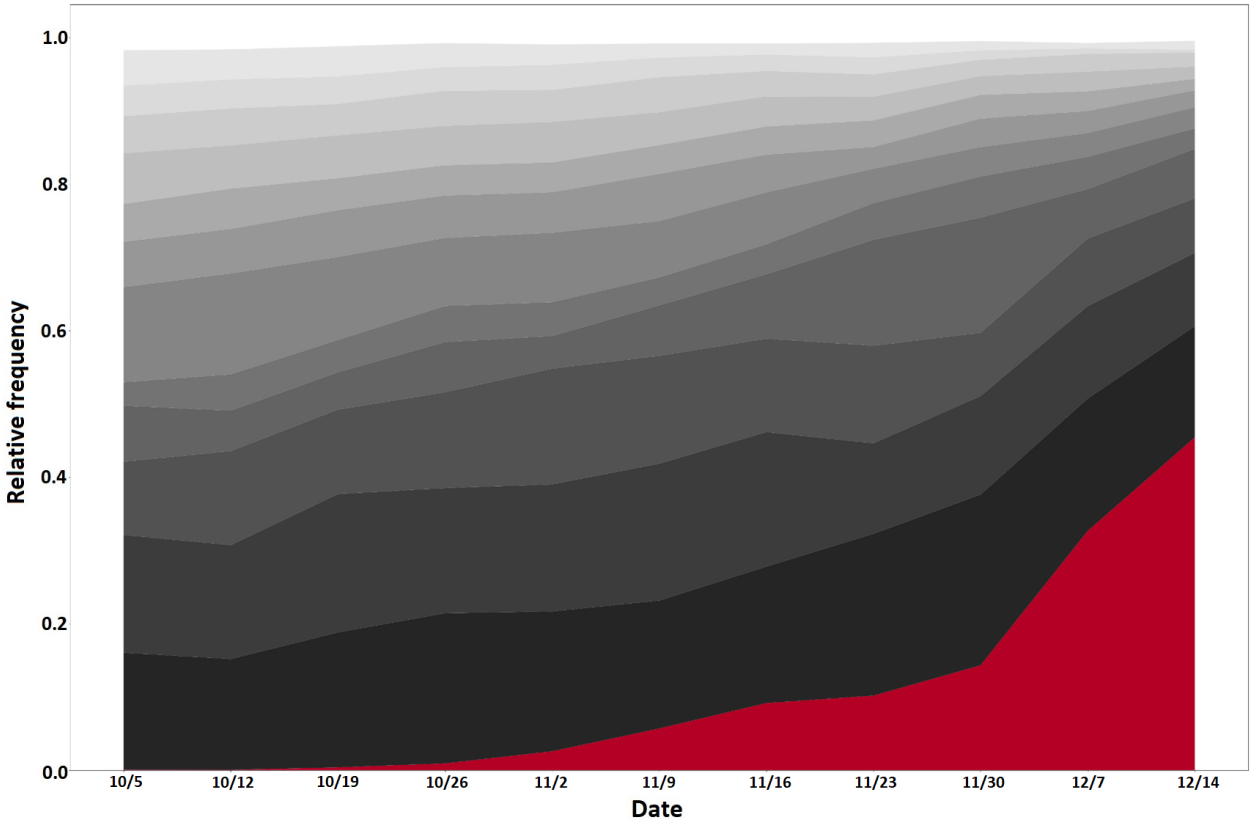

Figure 2.1 Subtype distribution (the UK dataset, weekly window, relative count), produced our CliqueSNV-based clustering method. The subtype in red contributes to sequences that correspond to the Alpha variant.

Here, we attempt to use the clustering obtained by some clustering methods in order to fill the gaps. That is, rather than uniformly filling all sequences with, *e.g.*, the reference genome, we fill each sequence with the center of its cluster. The idea is that if a clustering performs well, then the sequences of a cluster should correspond to a subtype. In this case, the center — a consensus sequence of this subtype — should be much closer to any sequence of its cluster than the reference genome, resulting in a more accurate filling of the gaps.

2.1.3 Results

Our CliqueSNV-based clustering method was able to detect one subtype which tends to dominate the population in the UK dataset, in attaining good entropy and F_1 scores[120]. However, we wanted to further validate if this is consistent with other independent measures of quality, such as the cluster-based fitness coefficient. To compute this, we chose our time points t to be intervals of one week over the period from the beginning of October to the middle of December, exactly as in Figure 2.1. The size $X_i(t)$ of each cluster C_i (of $k = 15$ clusters) for every week t was obtained, and each fitness coefficient r_i was computed according to 2.4. In order to reduce sampling error, we drew 2000 random samples from the Poisson distribution on $X_i(t)$. We repeated this 100 times, and we report in Table 2.3 the 95% confidence interval of the top five clusters, sorted by interval lower bound. We note that similar results are obtained with either Hamming or TN-93 distance, with TN-93 distance corresponding to slightly higher fitness coefficients. We confirm that in either case, the mostly highly ranked cluster in terms of fitness (with cluster ID 6) corresponds to the cluster containing all of the sequences pertaining to the Alpha variant from above (specificity $> 99\%$).

The GISAID 2 dataset Since our CliqueSNV-based clustering approach was able to clearly pinpoint the Alpha variant within the UK dataset, we tested it also on the GISAID 2 dataset, which contains many of the variants listed in Table 2.1. CliqueSNV-based clustering identified 36 subtypes in this dataset. We first computed fitness coefficients r_i (2.4) for these 36 clusters using one-week time intervals t. Table 2.4 reports the 95% confidence interval due to subsampling of the top and bottom five clusters, sorted by interval lower bound. One will notice immediately that the

Distance	Rank	Cluster ID	Int. Lower B.	Int. Upper B.
Hamming	1	6	1.343	1.504
	2	3	0.354	0.369
	3	14	0.284	0.324
	4	8	0.08691	0.0881
	5	2	0.08690	0.0878
TN-93	1	6	1.390	1.510
	2	2	0.789	0.795
	3	3	0.351	0.364
	4	14	0.353	0.390
	5	8	0.086	0.0869

Table 2.3 The 95% confidence interval of the top five fitness coefficients, according to the interval lower bound, of the 15 clusters of the UK dataset obtained using our CliqueSNV-based clustering method with Hamming distance and TN-93 distance, respectively.

Rank	Cluster ID	Int. Lower B.	Int. Upper B.
1	1	0.0601	0.0602
2	17	0.0486	0.0489
3	21	0.0463	0.0463
4	20	0.0456	0.0457
5	35	0.0440	0.0440
32	4	0.0143	0.0143
33	29	0.0138	0.0138
34	28	0.0120	0.0120
35	32	0.0118	0.0118
36	34	0.0110	0.0110

Table 2.4 The 95% confidence interval of the top and bottom five fitness coefficients, according to the interval lower bound, of the 36 clusters of the GISAID 2 dataset were obtained using our CliqueSNV-based clustering method. The mean (μ) $\pm$ standard deviation (σ) of the interval lower and upper bounds are 0.0281 ± 0.0122 and 0.0281 ± 0.0122, respectively.

fitness coefficient is much more evenly distributed across the clusters of this dataset, compared to the UK dataset (Table 2.3).

Table 2.5 reports some of the variants found by our CliqueSNV-based approach in terms of specificity, F_1 score, and fitness rank (Table 2.4). Notice that specificity/F_1 score generally decreases with rank and cluster size, as would be expected. Exceptions to this trend are the Gamma/Zeta variant in F_1 score vs. Rank (having a high F_1 score for its rank) and the Epsilon variant (having a large cluster size for its F_1 score and rank). 50% and F_1 scores ≥ 0.5.

Variant	ID	Specificity	F_1	Rank	Size
Alpha (UK)	1	93.16%	0.96	1	265 255
Gamma & Zeta (Brazil)	25	51.21%	0.68	7	1892
Beta (S. Africa)	21	45.85%	0.62	3	2754
Epsilon (California)	13	41.08%	0.58	13	9251

Table 2.5 Specificity, F_1 score and fitness rank (Table 2.4) of the cluster containing the largest number of sequences of the corresponding variant.

2.2 Bayesian assessment of the fitness landscape of SARS-CoV-2

2.2.1 Datasets

We obtained the genomic data and associated metadata analyzed in this section from GISAID[167]. The earliest date when genomes that belong to the B.1.1.7 lineage were sampled is September, 20[147]. Thus, we analyzed the sequences from the UK generated after September, 13 (one week before the detection of B.1.1.7) and before December 17. 2021. All genomes were separated into two groups based on the presence of the set of 8 SNVs in the spike protein identified in[147], which includes two codon deletions and 6 point mutations. Similarly, for B.1.617.2 lineage, the sequences from the UK were clustered into two groups based on the presence of the 7 SNVs in the spike protein from April first, 2021 to May 31, 2021 (the first two months that B.1.617.2 lineage spread throughout the UK). The total SARS-CoV-2 case counts for the analyzed time period were obtained from COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University[55].

For epistasis analysis, a global multisequence alignment up to July 25th, 2021 is downloaded from GISAID. The spike protein is extracted from the whole genome. Ambiguous characters were replaced by blanks. Then Gaps with a 90% threshold over the whole alignment were removed. The

consensus of this reduced alignment is 99.8% identical to the spike protein reference. Eventually, the sequences belonging to each country were separated using the metadata, which is also downloaded from GISAID.

2.2.2 Methods

2.2.2.1 Growth rate estimation

In this section, we describe our method for estimation of growth rates and basic reproduction numbers for subpopulations of SARS-CoV-2 genomes spreading in the population of susceptible individuals. In what follows, vectors are highlighted in bold. Formally, we are given a viral population $\mathcal{P} = \mathcal{P}_1 \cup ... \cup \mathcal{P}_n$ consisting of n subpopulations with different phenotypic features, and the goal is to estimate the most likely *fitness landscape* $\boldsymbol{f} : \mathcal{P}_i \mapsto f_i$. In this study, $n = 2$, but the framework described below is applicable for any number of subpopulations. We assume that the viral population has been sampled over the discrete time interval $\boldsymbol{\tau} = (\tau_1, ..., \tau_S)$. For each time-point τ_j, we consider the observed subpopulation counts $\boldsymbol{k}^j = (k_1(\tau_j), ..., k_n(\tau_j))$, the total number of sequenced cases $l(\tau_j) = \sum_{i=1}^{n} k_i(\tau_j)$ and the observed incidence $c(\tau_j)$. Each subpopulation $\mathcal{P}_i$ also has its first sampling time $t_i \leq \min\{\tau : k_i(\tau) > 0\}$.

We describe the spread of viral subpopulations using a generalized logistic sub-epidemic model[32]. We consider the overall epidemic consisting of n overlapping sub-epidemic waves corresponding to the spread of particular haplotypes. The waves are described by the system of differential equations

$$\frac{d}{dt}X_i(t) = f_i\alpha_i(t)X_i^d(t)\left(1 - \frac{X_i(t)}{Q_i}\right), \quad i = 1, ..., n. \tag{2.10}$$

where $\overline{\boldsymbol{X}}(t) = (X_1(t), ..., X_n(t))$ are the cumulative numbers of infections for the ith sub-epidemic and $\alpha_i(t)$ is the indicator function for the onset timing of the ith sub-epidemic:

$$\alpha_i(t) = \begin{cases} 0, \text{ if } t < t_i; \\ 1, \text{ if } t \geq t_i. \end{cases} \tag{2.11}$$

Besides the growth rates f_i, another model parameters are "scaling of growth" parameter d and the maximum sub-epidemic size $Q_i = Q_0 e^{-qt_i}$, where Q_0 is the initial sub-epidemic size and q is the rate of sub-epidemic decline caused by public health interventions or population behavior changes that mitigate transmission[32]. In addition, in what follows we consider the vector of sub-epidemic *daily incidences* $\overline{\boldsymbol{Y}}(\tau_j) = (Y_1(\tau_j), ..., Y_n(\tau_j))$, where $Y_i(\tau_j) = X_i(\tau_j) - X_i(\tau_{j-1})$, $i = 1, ..., n$. We estimate the model parameters $\boldsymbol{f}$, Q_0, d, q by maximizing $p(\boldsymbol{f}, d, Q_0, q, \boldsymbol{t}|\boldsymbol{C}, \boldsymbol{k})$, the posterior probability of model parameters given the observed incidence data $\boldsymbol{C} = (c(\tau_1), ..., c(\tau_S))$ and sampled subpopulation data $\boldsymbol{k} = (\boldsymbol{k}^1, ..., \boldsymbol{k}^S)$ in a Bayesian way:

$$p(\boldsymbol{f}, d, Q_0, q, \boldsymbol{t}|\boldsymbol{C}, \boldsymbol{k}) \propto p(\boldsymbol{C}|\boldsymbol{f}, d, Q_0, q, \boldsymbol{t})p(\boldsymbol{k}|\boldsymbol{f}, d, Q_0, q, \boldsymbol{t})\left(\prod_{i=1}^{n} p(f_i)p(t_i)\right)p(d)p(Q_0)p(q) \tag{2.12}$$

The likelihoods $p(\boldsymbol{C}|\boldsymbol{f}, d, Q_0, q)$ and $p(\boldsymbol{k}|\boldsymbol{f}, d, Q_0, q)$ are defined by assuming that for each time-point τ_j

(a) the observed incidence $c(\tau_j)$ is drawn from the Poisson distribution with the mean equal to

the model-based estimated total incidence $\boldsymbol{Y}(\tau_j) = \sum_{i=1}^{n} Y_i(\tau_j)$.

(b) the observed subpopulation counts $\boldsymbol{k}^j = (k_1(\tau_j), ..., k_n(\tau_j))$ follow the multinomial distribution with sampling probabilities $\boldsymbol{p^j} = (p_1^j, ..., p_n^j)$, where $p_i^j = \frac{Y_i(\tau_j)}{\boldsymbol{Y}(\tau_j)}$ (similar to[183]).

In this study, the priors $p(f_i)$, $p(d)$, $p(Q_0)$,$p(q)$ and $p(t_i)$ were defined by assuming that the corresponding parameters are distributed uniformly on the intervals $[0, f^{\max}]$, $[0, p^{\max}]$, $[0, Q_0^{\max}]$,$[0, q^{\max}]$ and $[\tau_1, \min\{\tau : k_i(\tau) > 0\}]$, respectively. Thus, after transition to log-likelihoods and the dropping of constant terms, the parameters were estimated by solving the following constrained optimization problem:

$$(\boldsymbol{f}^*, d^*, Q_0^*, q^*, \boldsymbol{t}) = \underset{\boldsymbol{f}, p, Q_0, q, \boldsymbol{t}}{\arg\max} \quad \boldsymbol{C} \cdot \log(\boldsymbol{Y}) + \sum_{j=1}^{S} \boldsymbol{k^j} \cdot \log(\overline{\boldsymbol{Y}}(\tau_j)) - \boldsymbol{1} \cdot \boldsymbol{Y} - \boldsymbol{l} \cdot \log(\boldsymbol{Y}) \quad (2.13)$$

subject to constraints $0 \leq f_i \leq f^{\max}$, $0 \leq p \leq p^{\max}$, $0 < Q_0 \leq Q_0^{\max}$, $0 \leq q \leq q^{\max}$ and $\tau_1 \leq t_i \leq \min\{\tau : k_i(\tau) > 0\}$. Here $\overline{\boldsymbol{Y}}$ and $\boldsymbol{Y}$ are functions of $\boldsymbol{f}, p, Q_0, q, \boldsymbol{t}$; the symbol "$\cdot$" denotes a scalar product of vectors and $\log$ is a coordinate-vise natural logarithm. To solve the problem (2.13), we used the gradient-free pattern search approach[10], as implemented in Matlab 2019b (MathWorks, Natick, MA).

The inferred model parameters were used to estimate the reproduction numbers associated with each subpopulation. The generation interval of SARS-CoV-2 was modeled assuming gamma distribution with a mean of $\mu = 5.2$ days and a standard deviation of $\sigma = 1.72$ days. Then, if $\rho(t)$ is a probability distribution of the generation interval t and $Y_i^*(\tau_j)$ is the model-based incidence of the i-th subepidemic calculated using optimal parameters, then the i-th effective reproduction number was calculated using the renewal equation[133,68] as follows:

$$R^i_{\tau_j} = \frac{Y^*_i(\tau_j)}{\sum_{l=1}^{j} Y^*_i(\tau_j - \tau_l)\rho(\tau_l)} \tag{2.14}$$

Here the numerator represents the total number of new cases at a given time τ_j, and the denominator represents the total number of cases that contribute (as primary cases) to generate the new cases at τ_j. Note that we consider only the first wave of each subpopulation because later waves usually arise when the genomic variants under consideration are ”passenger mutations” of other variants. To do so,

2.2.3 Evaluating epistasis interactions

Epistasis networks of the spike protein for each country were built using our growth rate estimation method. A pair of mutations can be evaluated as an epistasis pair if the mutations happened both independently and concurrently in the population. Therefore we consider only the mutation pairs for which all four haplotypes 00, 01, 10, and 11 occurred significantly in the population. In order to find pairs that meet this requirement, a zero-one mutation matrix is calculated. A $m.n$ matrix, where m is the number of sequences in the population and n is the number of positions in the spike gene (1274 amino acids or 3822 nucleotides). Each sequence is compared to the spike gene extracted from the reference. A zero is placed at (i, j) entry of the matrix, if the character at a position j in sequence i agrees with the character at the position j in the reference, and a one if they don’t agree. Blanks and ambiguous characters were ignored. In other words, non-reference alleles are one, and reference alleles are zero in the matrix. To make sure low-quality sequences are not considered, rows with more than the average number of mutations plus 10 are removed.

Also, the identical columns (almost all zeros or all ones) with a threshold of 200 got removed. Considering the permutation of remaining alleles in the matrix, the count of haplotypes 00, 01, 10, and 11 are calculated for each pair (i, i') in the spike gene. Then pairs with a frequency of more than 100 for each haplotype are passed to the next step, which is to estimate the fitness values.

For each haplotype $ab \in \{00, 01, 10, 11\}$ of pair (i, i'), all sequences are divided into two clusters, one including all the sequences having haplotype ab, Y_{ab} and the other including the rest of sequences, N_{ab}. The daily frequencies (subpopulation counts) for each cluster are obtained over the time period (up to July 25th, 2021) using the collection date of each sequence extracted from the metadata. Growth rate values $F(Y_{ab})$ and $F(N_{ab})$ are computed and $F_{ab} = F(Y_{ab})/F(N_{ab})$ is considered as a fitness value of the pair ab at (i, i'). To generate parameter distributions, we used parametric bootstrapping with $k = 500$ bootstraps and a Poisson noise added to the observed total numbers of cases and the observed subpopulation counts.

The final step is to evaluate the epistasis interactions using $\Delta = F_{11} - F_{00} - F_{01} - F_{10}$ for each position pairs (i, i')[67,166,16,42]. Δ indicates positive epistasis if bigger than zero and negative epistasis if less than zero. $\Delta = 0$ indicates no epistasis (additive epistasis). Figure 2.2 demonstrates the pipeline of epistasis mutation analysis.

An epistasis network $G = (V, E)$ is defined for each country. Where, V is a set of nodes including the positions in the spike protein, associated with epistasis pairs and, E is a set of edges showing their epistasis interactions where $\Delta \neq 0$.

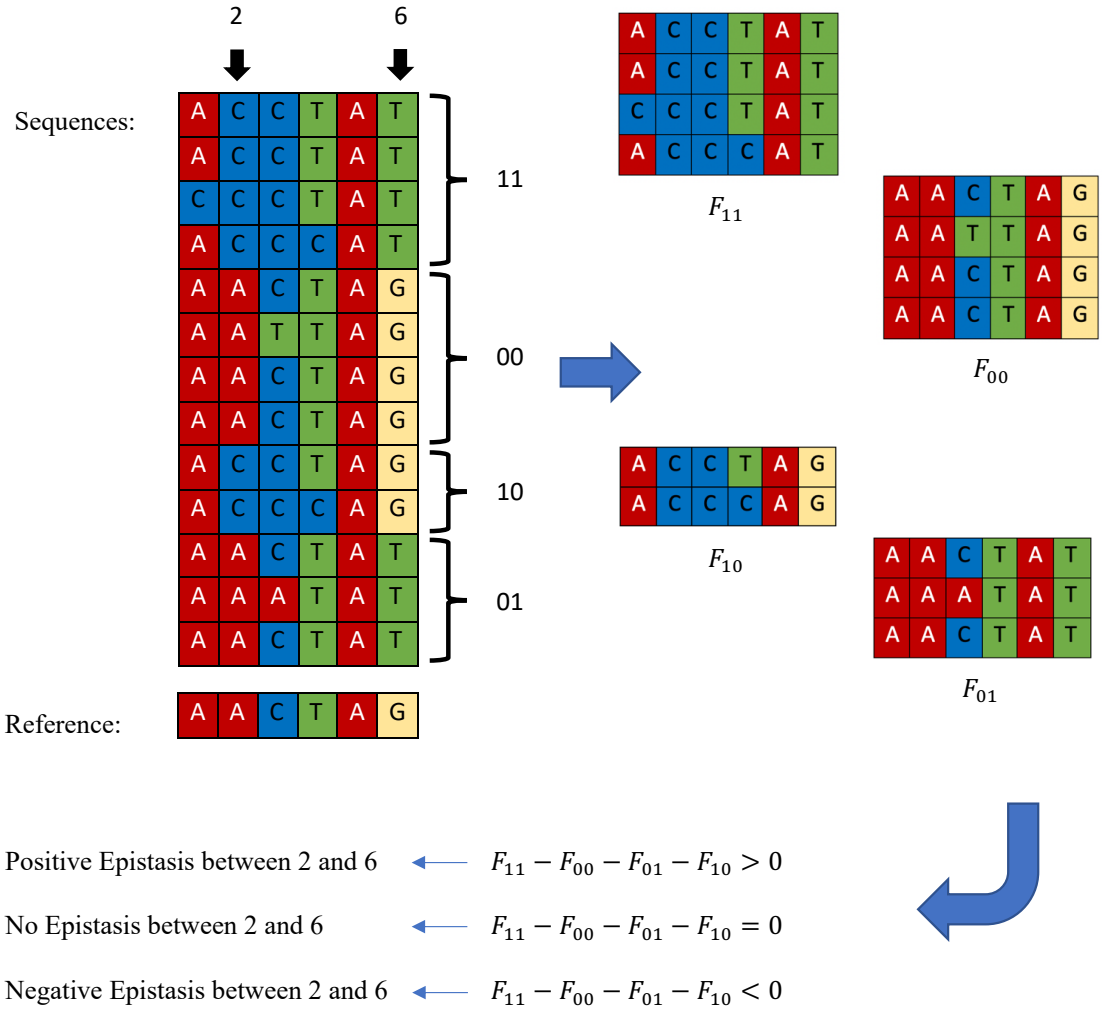

Figure 2.2 Pipeline of Epistasis interaction analysis of SARS-CoV-2 sequences. The sequences including each pair of haplotypes are separated and the fitness values of the four groups are estimated using our Bayesian model. Then the value of $F_{11} - F_{00} - F_{01} - F_{10}$ determines if there are any epistasis interactions between the considered positions.

2.2.4 Results

2.2.4.1 The structure of S-gene epistatic network

The obtained networks for the USA and the UK are shown in figures 2.4 and 2.3. The structure of both networks is pretty similar. Interestingly, each network contains a single ”giant” component that includes 126 (11.1%) vertices. In addition, there are components of sizes 3 and 2. The s-gene epistatic network appears to be scale-free, with the right-skewed degree distribution. Degree distributions of such networks follow power-law (i.e. the probability of having a particular degree is proportional to the power of that degree), and they are often the result of a preferential attachment process, where a vertex joining a network gets connected to an existing vertex with the probability proportional to the degree of that vertex - the model is often described by the metaphor ”the rich get richer”. We fitted negative binomial, beta negative binomial, Poisson, Yule-Simon, Generalized Pareto, and Pareto distributions to the observed degree distribution of the transmission network. To compare the goodness of fit yielded by different models, we used the Akaike (AIC) and Bayesian (BIC) Information Criteria (Table 2.6). The Yule-Simon distribution, which represents the classical power-law, demonstrated the best fit. The exponent of the Pareto distribution was estimated to be 1.20 ($95\% CI = [1.12, 1.34]$), which is lower than for most complex scale-free networks studied in the literature, thus indicating the higher tendency of vertices to be connected to hubs (high-degree vertices).

Distribution	UK		USA	
	AIC	BIC	AIC	BIC
Negative binomial	658.7643	654.7643	1340.8	1336.8
Beta negative binomial	484.8638	478.8638	889.6	883.6
Poisson	627.5770	625.5770	1467.7	1465.7
Yule-Simon	396.6424	394.6424	765.1	763.1
Pareto	425.7476	423.7476	871.6	869.6
Generalized Pareto	476.4723	472.4723	891.8	887.8

Table 2.6 The Akaike (AIC) and Bayesian (BIC) Information Criteria for the largest connected component in positive epistasis networks of the UK and the USA

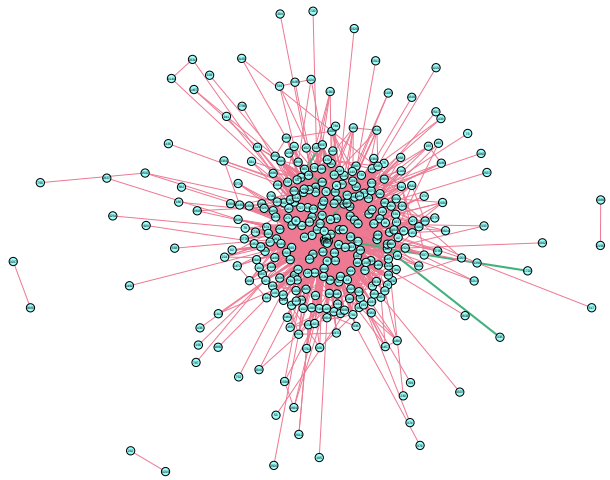

Figure 2.3 Estimated epistasis network for USA. The red edges indicate the negative epistasis interactions and the green edges indicate positive epistasis interactions.

2.2.4.2 Validation of fitness inference model on known Variants of Concern

Maximum a posteriori basic reproduction numbers $R_0(1)$ and $R_0(2)$ were estimated for non-B.1.1.7 and B.1.1.7 lineage, and similarly for non-B.1.617.2 and B.1.617.2 lineage subpopulations using the methods described above. A parametric bootstrapping with $k = 500$ bootstraps was generated and a Poisson noise was added to the observed total numbers of cases and the observed subpopulation counts. The upper bounds for the Bayesian inference were: $f^{\max} = 2$, $p^{\max} = q^{\max} = 1$, $Q_0^{\max} = 10^8$. Basic reproduction numbers were calculated using equation 2.14

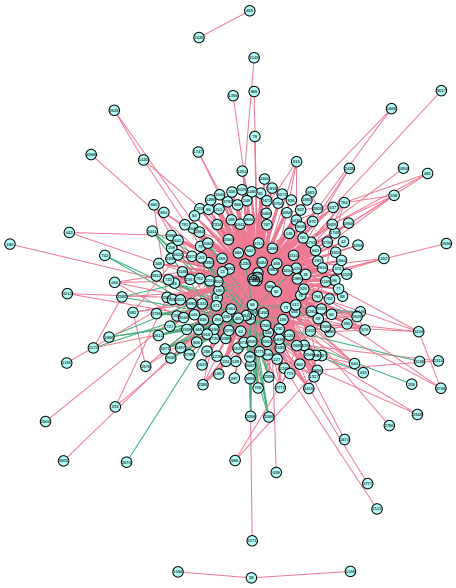

Figure 2.4 Estimated epistasis network for Uk. The red edges indicate the negative epistasis interactions and the green edges indicate positive epistasis interactions.

and the gamma distribution parameters $\mu = 5.2$ days and $\sigma = 1.72$ days taken from[72].

The estimated mean maximum a posteriori ratio of basic reproduction numbers of B.1.1.7 and non-B.1.1.7 subpopulations was $R_0(2)/R_0(1) = 1.641$ (95% percentile bootstrap CI $= [1.615, 1.754]$), see Figure 2.5. Thus, the estimated transmissibility of SARS-CoV-2 variants of B.1.1.7 lineage is approximately $\sim 64\%$ higher than for non-B.1.1.7 variants ($p < 0.001$, Kruskal-Wallis test). This estimation agrees with the estimations of relative transmissibility of emerging UK-based SARS-CoV-2 variants presented in other early studies[43,105]. It can be forecasted, that within 21 day period from the day of the last observation, B.1.1.7 variants may constitute $\sim 58.7\%$ of all new cases in the UK (Figure 2.5). As well, the estimated transmissibility of B.1.617.2 lineage is approximately $\sim 36\%$ more than B.1.1.7 lineage (R_0 95% percentile bootstrap CI $= [2.069, 2.070]$).

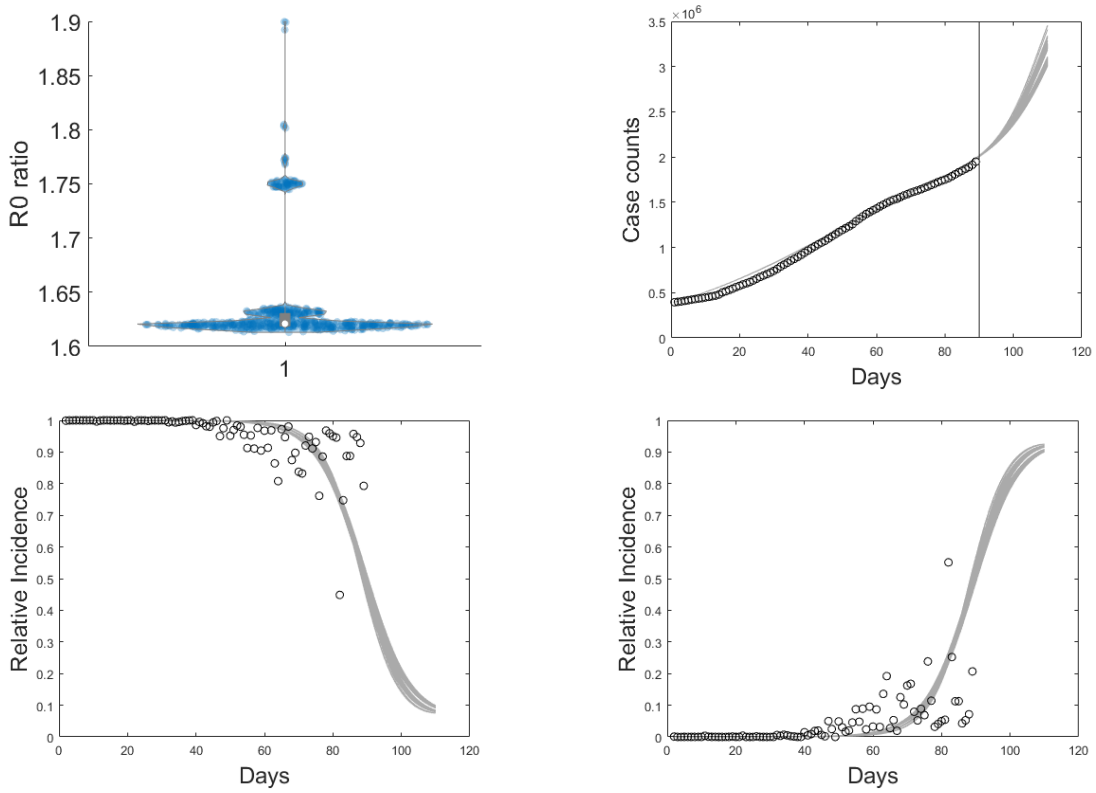

Figure 2.5 Upper left: violin plot of basic reproduction number ratios for B.1.1.7 and non-B.1.1.7 subpopulations. Upper right: total case counts. Lower left and right: relative incidence of non-B.1.1.7 and B.1.1.7 subepidemics (between September 20, 2020, and December 17, 2020, as well as forecasted to 21 days after the latter date). Circles depict frequencies of B.1.1.7 variants observation among sequenced genomes. Different predicted relative incidence trajectories are depicted by grey curves

CHAPTER 3

Prediction of emerging variants of SARS-CoV-2 with altered phenotypes

Understanding the predictability of evolution and the relative impact of random and deterministic factors in evolutionary processes is a fundamental problem in life sciences. This problem gains an applied significance in the context of viruses and other pathogens, as even a modest degree of predictability of pathogen evolution can enhance our ability to forecast and, therein, control the spread of infectious diseases[104,87,117,154].

The most evident example of the importance of this problem is the case of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The successive waves of COVID-19 are driven by the emerging variants of interest (VOIs) or variants of concern (VOCs) that have been associated with altered phenotypic features, including transmissibility[44,175,82,203], antibody resistance and immune escape[186,83,73,142]. Each variant is defined as a phylogenetic lineage characterized by a specific combination of single amino acid variants (SAVs) and/or indels acquired over the course of SARS-CoV-2 evolution. For instance, lineages B.1.1.7 (alpha variant by WHO classification) and B.1.617.2 (delta variant) are defined by distinct families of 7 SAVs in the spike protein[147,94], many of which have been linked to enhanced fitness compared to preceding SARS-CoV-2 lineages[175,82,203,184,109,147,105].

Genomic epidemiology has been crucial for monitoring the emergence and spread of SARS-CoV-2 variants since the start of the COVID-19 pandemic. SARS-CoV-2 genomes sampled around the globe and produced using high-throughput sequencing technologies have been analyzed by a plethora of phylogenetic, phylodynamic, and epidemiological models[101] to detect spreading lin-

eages and measure their reproductive numbers and other epidemiological characteristics. However, these methods, powerful and valuable as they are, are primarily applied retrospectively. In other words, they allow to *detect* growing lineages and measure their fitness only when these lineages are already sufficiently prevalent. Moreover, existing phylogenetic and phylodynamic approaches are computationally expensive. They must use subsampling, simplifying assumptions, and heuristic algorithms without performance guarantees to handle the vast amounts of available genomic data (e.g., more than 14 million sequences in the GISAID database[167] at the time of submission of this paper). These considerations can impact their power, accuracy, and reliability.

In contrast to retroactive detection, the task of *early detection* or *forecasting* involves the proactive identification of SARS-CoV-2 genomic variants that have the potential to become prevalent in the future. This problem is more challenging as it is intertwined with the fundamental question of whether viral evolution can be predicted or whether one can ”replay the tape of life” for the global SARS-CoV-2 evolution, using the metaphor of S.J. Gould[76]. For viruses, the possibility of evolutionary predictions remains a topic of debate[14]. Nevertheless, studies attempting to address the SARS-CoV-2 evolutionary forecasting problem have emerged[117,154,12,1,199,135]. Most of these studies have focused on the emergence of *individual mutations*, with some methods assuming that mutations accumulate independently or that the effects of their interactions can be averaged out over their genomic backgrounds[117,135].

Meanwhile, a number of studies have highlighted the significance of *epistasis*, i.e., the non-additive phenotypic effects of combinations of mutations, for SARS-CoV-2[202,152,201,154,153,128,124]. Using various methodologies, including phylogenetic analysis[152,128], direct coupling analysis[154],

and in vitro binding measurements[201,124], these studies suggest the existence of an epistatic network that includes many genomic sites in the receptor-binding domain of the spike protein that is associated with increased binding affinity to angiotensin-converting enzyme 2 (ACE2) receptor[3,113,142]. Epistasis is closely linked to the complex structures of viral fitness landscapes[16,154,124,202], which determine the evolutionary trajectories of SARS-CoV-2 lineages and contribute to the high non-linearity of its evolution, making forecasting challenging. The emergence of new Variants of Concern, such as the lineage B.1.1.529 (Omicron variant), is an example of such non-linear phenomena[124]. Its rapid emergence does not align with the gradual mutation accumulation hypothesis and is still a topic of debate, with hypothesized origins including immune-suppressed hosts and reverse zoonosis[137,39,103,187,124].

Given the role of epistasis, it can be argued that selection often acts on combinations of mutations, or *haplotypes*, rather than on individual mutations. Therefore, effective forecasting should focus on viral haplotypes instead of solely on SAVs. However, predicting haplotypes is a significantly more challenging problem than predicting individual SAVs – in particular, simply due to the exponential increase in the number of possible haplotypes with genome length. This complexity precludes the use of traditional approaches utilized in most mutation-based studies, where a feature vector of epidemiological, evolutionary, and/or physicochemical parameters is calculated for each SAV, and a statistical or machine learning model is trained to predict SAV phenotypic effects. As a result, even studies that account for epistatic effects usually focus on assessing the phenotypic effects of individual mutations[154].

This paper focuses on predicting haplotypes of SARS-CoV-2 using a novel approach based

on analyzing dense communities of the *coordinated substitution networks* of the spike protein, which reflects potential positive epistatic interactions[153,128,27]. We demonstrate that emerging haplotypes with altered phenotypes can be accurately predicted by leveraging these communities and introduce HELEN (Heralding Emerging Lineages in Epistatic Networks) - a variant reconstruction framework that integrates graph theory, statistical inference, and population genetics methods. HELEN was validated by accurately identifying known SARS-CoV-2 VOCs and VOIs up to 10-12 months before they reached high prevalences and were designated by the WHO. Importantly, the majority of predictions were derived from data collected independently from different countries, further supporting their credibility. These results demonstrate that network density is a more precise, sensitive, and scalable measure than lineage frequency, allowing for reliable early detection or prediction of potential variants of concern before they become prevalent. For instance, our approach suggests that the spread of the Omicron haplotype or a closely related genomic variant could have been predicted as early as the beginning of 2021, almost a year before its designation as a VOC. Furthermore, the computational complexity of our method depends on genome length rather than the number of sequences, making it significantly faster than traditional phylogenetic methods for VOC detection and enabling it to handle millions of currently available SARS-CoV-2 genomes.

Our approach to the early detection of viral haplotypes utilizes a certain methodological similarity with the problem of inference of rare viral haplotypes from noisy sequencing data, particularly when produced by long-read sequencing technologies like Oxford Nanopore and PacBio. This problem has gained significant attention in recent years, with several new tools appearing

each year[97,100,110,22,90]. Some of these tools accurately infer rare haplotypes with frequencies comparable to the sequencing noise level. In particular, several tools developed by the authors of this paper achieve such results by identifying and clustering statistically linked groups of SNV alleles[100,7,8,120]. Although this approach is not directly transferable to haplotype prediction, it provided a foundation for this study.

3.1 Methods

3.1.1 Rationale

3.1.2 Construction of coordinated substitution networks

Following other studies (e.g.[152,128,27]), we consider networks of *potential* positive epistatic interactions or, in other terms, *coordinated substitution networks*. Specifically, given a multiple sequence alignment consisting of N genomes of length L, we define a coordinated substitution network $\mathcal{G}$ as a graph with nodes representing SAVs, and two nodes being adjacent whenever the corresponding non-reference alleles are simultaneously observed more frequently than expected by chance.

To formalize this definition, we extend the idea proposed in our previous studies[8,100]. Specifically, let U_0, U_1 and V_0, V_1 be the reference and SAV alleles at two particular genomic positions $U, V \in \{1, ..., L\}$, respectively. Let further E^t_{ij} and O^t_{ij} be the expected and observed counts of allele pairs (or 2-haplotypes) (U_i, V_j) at a time t.

We assume that viral evolution is driven by mutation and selection, where (a) 2-haplotypes (U_i, V_j) have replicative fitnesses f_{ij}; (b) allele transitions at positions U and V are random, and transitions between alleles i and j happen at rates q^U_{ij}, q^V_{ij}. Thus, expected 2-haplotype counts can

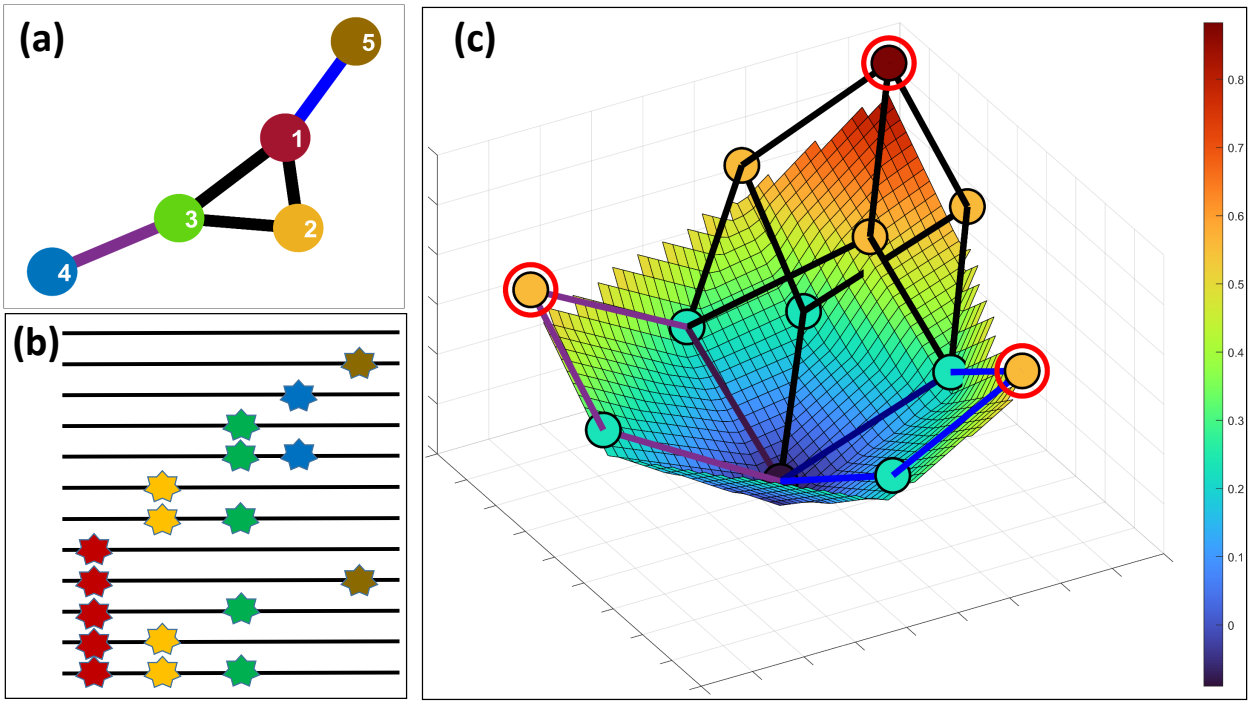

Figure 3.1 The model of an epistatically-constrained sequence space and fitness landscape. (a) The epistatic network $\mathcal{G}$. Edges of maximal cliques are displayed in blue, black, and purple. (b) Genotypes that are viable under the constraints imposed by the epistatic networks. Stars represent 1-alleles, and colors denote loci. (c) The viable space is depicted alongside the corresponding fitness landscape. Surface and vertex colors represent fitness values on a scale from blue (low fitness) to red (high fitness). Sub-hypercubes corresponding to three maximal cliques of the epistatic network $\mathcal{G}$ are highlighted in blue, black in red, respectively, with edges belonging to two sub-hypercubes colored in intermediate shades. The circled vertices represent local maximums within each sub-hypercube.

be described by the quasispecies model[134,58] (or mutation-selection balance model in the classical population genetics terms[194]) in the following form:

$$E_{ij}^{t} = \sum_{k,l=0,1} f_{kl} q_{ki}^{U} q_{lj}^{V} E_{kl}^{t-1} \tag{3.1}$$

We do not make any assumptions about the rate values, except that the rate of allelic change is

smaller than the rate of no-change, i.e.

$$q^U_{ij} < q^U_{ii}, q^V_{ij} < q^V_{ii}, i, j = 0, 1. \tag{3.2}$$

We use the model (3.1) to devise a statistical test that decides whether the 2-haplotype (U_2, V_2) is viable or its observed appearances can be plausibly explained by random mutations. The proposed test is based on the following fact:

Theorem 1. *Suppose that the 2-haplotype (U_2, V_2) is not viable, i.e. $f_{11} = 0$. Then*

$$E^t_{11} \leq \frac{E^t_{01} \cdot E^t_{10}}{E^t_{00}} \tag{3.3}$$

Proof. The proof follows the same lines as the proof in[8]. Given that $f_{11} = 0$, we have

$$\begin{aligned} E^t_{00} \cdot E^t_{11} &= \left(\sum_{k,l=0,1} f_{kl} q^U_{k0} q^V_{l0} E^{t-1}_{kl} \right) \left(\sum_{k,l=0,1} f_{kl} q^U_{k1} q^V_{l1} E^{t-1}_{kl} \right) \\ = q^U_{00} q^V_{00} q^U_{01} q^V_{01} (f_{00} E^{t-1}_{00})^2 &+ q^U_{10} q^V_{00} q^U_{11} q^V_{01} (f_{10} E^{t-1}_{10})^2 + q^U_{00} q^V_{10} q^U_{01} q^V_{11} (f_{01} E^{t-1}_{01})^2 + \\ &+ (q^U_{00} q^V_{00} q^U_{01} q^V_{11} + q^U_{00} q^V_{10} q^U_{01} q^V_{01}) f_{00} f_{01} E^{t-1}_{00} E^{t-1}_{01} + \\ &+ (q^U_{00} q^V_{00} q^U_{11} q^V_{01} + q^U_{10} q^V_{00} q^U_{01} q^V_{01}) f_{00} f_{10} E^{t-1}_{00} E^{t-1}_{10} + \\ &+ (q^U_{00} q^V_{10} q^U_{11} q^V_{01} + q^U_{10} q^V_{00} q^U_{01} q^V_{11}) f_{01} f_{10} E^{t-1}_{01} E^{t-1}_{10} \end{aligned} \tag{3.4}$$

and

$$E_{01}^t \cdot E_{10}^t = \left(\sum_{k,l=0,1} f_{kl} q_{k0}^U q_{l1}^V E_{kl}^{t-1}\right)\left(\sum_{k,l=0,1} f_{kl} q_{k1}^U q_{l0}^V E_{kl}^{t-1}\right)$$

$$\begin{aligned} = q_{00}^U q_{01}^V q_{01}^U q_{00}^V (f_{00}E_{00}^{t-1})^2 + q_{10}^U q_{01}^V q_{11}^U q_{00}^V (f_{10}E_{10}^{t-1})^2 + q_{00}^U q_{11}^V q_{01}^U q_{10}^V (f_{01}E_{01}^{t-1})^2 + \\ +(q_{00}^U q_{01}^V q_{01}^U q_{10}^V + q_{00}^U q_{11}^V q_{01}^U q_{00}^V) f_{00} f_{01} E_{00}^{t-1} E_{01}^{t-1} + \\ +(q_{00}^U q_{01}^V q_{11}^U q_{00}^V + q_{10}^U q_{01}^V q_{01}^U q_{00}^V) f_{00} f_{10} E_{00}^{t-1} E_{10}^{t-1} + \\ +(q_{00}^U q_{11}^V q_{11}^U q_{00}^V + q_{10}^U q_{01}^V q_{01}^U q_{10}^V) f_{01} f_{10} E_{01}^{t-1} E_{10}^{t-1} \end{aligned} \tag{3.5}$$

It is easy to see that the terms in (3.4) and (3.5) except for the last ones are equal. Thus we have

$$\begin{aligned} E_{01}^t \cdot E_{10}^t - E_{00}^t \cdot E_{11}^t = \\ = (q_{00}^U q_{11}^V q_{11}^U q_{00}^V + q_{10}^U q_{01}^V q_{01}^U q_{10}^V - q_{00}^U q_{10}^V q_{11}^U q_{01}^V - q_{10}^U q_{00}^V q_{01}^U q_{11}^V) f_{01} f_{10} E_{01}^{t-1} E_{10}^{t-1} = \\ = \left(1 - \frac{q_{01}^U q_{10}^U}{q_{00}^U q_{11}^U}\right)\left(1 - \frac{q_{01}^V q_{10}^V}{q_{00}^V q_{11}^V}\right) q_{00}^U q_{11}^V q_{11}^U q_{00}^V f_{01} f_{10} E_{01}^{t-1} E_{10}^{t-1} \geq 0, \end{aligned} \tag{3.6}$$

where the last inequality follows from (3.2). Thus, the inequality (3.3) holds. □

We use Theorem 1 to approximately evaluate 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. Considering the density of sampling and the number of available genomes, we assume that observed and expected numbers of 2-haplotypes are close to each other. Then, by (3.3), the value $p = \frac{O_{10} \cdot O_{01}}{O_{00} \cdot N}$ approximates the largest probability of observing a genome containing 2-haplotypes (U_1, V_1) among N sequenced genomes given that $f_{11} = 0$. 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

$$p(X \geq O_{11}|f_{11} = 0) = 1 - F_X(O_{11} - 1) = 1 - \sum_{i=0}^{O_{11}-1} \binom{N}{i} p^i (1-p)^{N-i}, \tag{3.7}$$

where F_X is the cumulative distribution function of the binomial distribution. We assume that SAVs U_1 and V_1 are *linked* (i.e. adjacent in the coordinated substitution network $\mathcal{G}$), when the probability (3.7) is low enough, i.e.

$$p(X \geq O_{11}|f_{11} = 0) \leq \frac{\rho}{\binom{L}{2}}, \tag{3.8}$$

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

3.1.3 Sampling of connected k-subgraphs and estimation of density-based p-values of viral haplotypes

In what follows, we will use the standard graph-theoretical notation: $V(\mathcal{G})$ and $E(\mathcal{G})$ are the sets of vertices and edges of the graph $\mathcal{G}$, respectively; $N_{\mathcal{G}}(v)$ is the set of neighbors of a vertex v in $\mathcal{G}$; the subgraph of G induced by a subset S is denoted by $\mathcal{G}[S]$.

We use the statistical test (3.8) to construct temporal coordinated substitution networks $\mathcal{G}_t$ for different time points t using SARS-CoV-2 sequences sampled before or at the time t. These networks have the same set of vertices but different sets of edges. A viral haplotype thus can be associated with a subset of vertices $H \subseteq V(\mathcal{G}_t)$ of a network $\mathcal{G}_t$. The density of a haplotype H is

thus defined as the density of the subgraph of $\mathcal{G}_t$ induced by H, i.e.

$$d_{\mathcal{G}_t}(H) = \frac{|E(\mathcal{G}_t[H])|}{|H|} \tag{3.9}$$

We hypothesize that viral haplotypes corresponding to potential VOCs and VOIs form dense subgraphs of $\mathcal{G}_t$. Below we describe how we verify and exploit this hypothesis.

The first step is to demonstrate the statistical significance of our hypothesis by producing density-based p-values of known VOC and VOI haplotypes H. The simplest way to assess these p-values is to randomly sample subgraphs of $\mathcal{G}_t$ of the size $|H|$ and calculate the proportion of sampled subgraphs with densities higher than H. However, SARS-CoV-2 temporal coordinated substitution interaction networks are relatively sparse, and thus many sampled subgraphs will be a priori disconnected and, consequently, also sparse. As a result, such a sampling scheme is inherently biased towards assigning low p-values to haplotypes corresponding to connected subgraphs and subgraphs with few connected components. Known VOCs and VOIs at most time points have these properties, and thus their statistical significance could be overestimated.

To overcome this problem, we utilize a more sophisticated randomized enumeration subgraph sampling scheme based on the network motif sampling algorithm introduced in[189]. This scheme uniformly samples only connected subgraphs and can be described as follows. Let us assume that all vertices of $\mathcal{G}_t$ are labeled by the unique integers $1, ..., L$. The sampling is performed using a recursive backtracking algorithm that, starting from each vertex $v \in V(\mathcal{G}_t)$, iteratively extends previously constructed connected subgraph S by adding a random new vertex w from the set of allowed extensions W. After that, the set of allowed extensions is updated by adding the neighbors

of w that do not belong to the set of avoided extensions X. The set of avoided extensions at each iteration contains the vertices that are neighbors of vertices previously added to S and the vertices with labels larger than v. These steps allow the algorithm to avoid double-sampling[189]. Extension stops, when a subgraph of the given size k is produced. Generation of k-subgraphs containing a given vertex v continues until the pre-defined sample size is achieved.

Given the subgraph sample $\mathcal{S}^* = \{S_1, ..., S_{|\mathcal{S}^*|}\}$, p-value of a haplotype H in the network $\mathcal{G}_t$ is defined as

$$p_{\mathcal{G}_t}(H) = \frac{|\{S_j \in \mathcal{S}^* : d_{\mathcal{G}_t}(S_j) \geq d_{\mathcal{G}_t}(H)|}{|\mathcal{S}^*|} \tag{3.10}$$

If, at some point, the subgraph induced by H is disconnected, we replace H with its largest connected component. For each analyzed spike coordinated substitution network $\mathcal{G}_t$, the sampling was performed until $k = \min\{3000, \eta_{\mathcal{G}_t}(v)\}$ subgraphs for each vertex v are generated, where $\eta_{\mathcal{G}_t}(v)$ is the total number of connected subgraphs containing v.

3.1.4 Inference of viral haplotypes as dense communities in coordinated substitution networks

We propose to infer viral haplotypes as dense communities of coordinated substitution networks. Community detection is a well-established field of network science, with numerous algorithmic solutions proposed over the last two decades[30,151,130]. Typically (though not always), the collection of communities in a network is defined as a partition[181]. However, in the case of viral genomic variants, there can be overlaps, as observed in known VOCs and VOIs. Additionally, most existing algorithms are heuristics designed to scale to the sizes of extremely large networks rather than to

produce optimal solutions. S-gene coordinated substitution networks, although containing hundreds of vertices, are typically smaller than most networks studied in applied network theory. Thus we use our own community detection approach, which extends our previously developed methodology[100]. This approach uses exact algorithms rather than heuristics and is tailored to account for the characteristics of viral data.

Firstly, we use a Linear Programming (LP) formulation[29] to find the densest subgraphs of networks $\mathcal{G}_t$ at each time point t. This formulation contains variables x_i for each vertex $i \in V(\mathcal{G}_t)$, variables y_{ij} for each edge $ij \in E(\mathcal{G}_t)$, and the following objective function and constraints:

$$\sum_{ij \in E(\mathcal{G}_t)} y_{ij} \to \max \tag{3.11}$$

$$y_{ij} \leq x_i, \quad y_{ij} \leq x_j, \quad ij \in E(\mathcal{G}_t) \tag{3.12}$$

$$\sum_{i \in V(\mathcal{G}_t)} x_i \leq 1 \tag{3.13}$$

$$x_i, y_{ij} \geq 0, \quad i \in V(\mathcal{G}_t), ij \in E(\mathcal{G}_t) \tag{3.14}$$

Note that the variables x_i, y_{ij} are continuous rather than integer since it can be shown that the value of the optimal solution of the LP (3.11)-(3.14) and the maximum subgraph density of $\mathcal{G}_t$ coincide[29]; furthermore, if $U \subseteq V(\mathcal{G}_t)$ is the vertex set of the densest subgraph, then $(x_i = \frac{1}{|U|}, i \in$

$U; x_i = 0, i \notin U; y_{ij} = \frac{1}{|U|}, i, j \subseteq U; y_{ij} = 0, i, j \not\subseteq U)$ is the optimal solution of (3.11)-(3.14). Thus, densest subgraphs of the networks $\mathcal{G}_t$ can be found in a polynomial time.

The single densest subgraph can, however, provide only a single haplotype per time point. We need to generate multiple dense communities to infer multiple haplotypes that could correspond to VOCs and VOIs. Our method produces these communities is as follows. We iterate through a given range of fixed subgraph sizes k ($k = k_{\max}, k_{\max} - 1, ..., k_{\min}$); at each iteration, we generate a set $\mathcal{S}_k$ of up to $n_{\max}$ densest subgraphs of size k that are not contained in subgraphs generated in the previous iterations. Here $k_{\max}$,$k_{\min}$ and $n_{\max}$ are parameters of the algorithm. However, finding the densest subgraph of a given size is an NP-hard problem[64,9]. Therefore, for each value of k, we use the following Integer Linear Programming formulation:

$$\frac{1}{k} \sum_{ij \in E(\mathcal{G}_t)} y_{ij} \to \max \tag{3.15}$$

$$y_{ij} \leq x_i, \quad y_{ij} \leq x_j, \quad ij \in E(\mathcal{G}_t) \tag{3.16}$$

$$\sum_{i \in V(\mathcal{G}_t)} x_i = k \tag{3.17}$$

$$\sum_{i \in V(\mathcal{G}_t) \setminus S} x_i \geq 1, \quad S \in \bigcup_{k'=k+1}^{k_{\max}} \mathcal{S}_{k'} \tag{3.18}$$

$$x_i, y_{ij} \in \{0,1\}, \quad i \in V(\mathcal{G}_t), ij \in E(\mathcal{G}_t) \tag{3.19}$$

The problems (3.11)-(3.14) and (3.15)-(3.19) are solved using Gurobi[78]; for the latter, we used an option to continue the search until the pool of up to $n_{\max}$ optimal solutions is produced.

Now, let $\hat{\mathcal{S}}t = S_{t,1}, ..., S_{t,|\hat{\mathcal{S}}t|}$ be the set of generated densest subgraphs with sizes ranging from $k_{\min}$ to $k_{\max}$. This set does not necessarily have a one-to-one correspondence with the true haplotypes due to two reasons. First, some haplotypes may consist of more than $k_{\max}$ SAVs, so the generated subgraphs only cover parts of these haplotypes. Second, many generated subgraphs overlap significantly, and thus most likely correspond to the same haplotypes. To obtain full-length haplotypes, we employ an algorithmic pipeline described below. Initially, we split the generated dense subgraphs into clusters such that each cluster ideally corresponds to a single true haplotype. Then, we locate the corresponding haplotype for each cluster by finding the densest core community in a subgraph induced by the union of elements of that cluster. Figure 3.2 illustrates the pipeline, which we describe in detail in the following Algorithm.

Algorithm 2: inference of viral haplotypes.

Input: the set of dense subgraphs $\hat{\mathcal{S}}_t = \{S_{t,1}, ..., S_{t,|\hat{\mathcal{S}}_t|}\}$

Output: the set of haplotypes $\mathcal{H}_t = \{H_{t,1}, ..., H_{t,|\mathcal{H}_t|}\}$.

1) Construct *an intersection graph* $\mathcal{L}(\hat{\mathcal{S}}_t)$, whose vertex set is $\hat{\mathcal{S}}_t$, and two vertices $S_{t,i}$ and $S_{t,j}$ are adjacent, whenever $|S_{t,i} \cap S_{t,j}| \geq \min\{|S_{t,i}|, |S_{t,j}|\} - 1$.

2) Partition $\mathcal{L}(\hat{\mathcal{S}}_t)$ into clusters $L_{t,1}, ..., L_{t,r}$:

2.1) Split $\mathcal{L}(\hat{\mathcal{S}}t)$ into connected components and then subdivide each component into $(\kappa + 1)$-connected components, where κ denotes the vertex connectivity. To achieve this, we use a modified version of the algorithm proposed by[62], which computes the vertex connectivity and corresponding vertex cut as the smallest of (s, t)-cuts between specifically chosen vertices of the graph. The algorithm computes these (s, t)-cuts using network flow techniques[63]. We further augment this algorithm by adding an extra step. Consider a pair of vertices (s, t) for which the minimal vertex cut of size $\kappa_{s,t}$ has been found, and $P^1_{s,t}, ..., P^{\kappa_{s,t}}_{s,t}$ are the corresponding internal vertex-disjoint (s, t)-paths (which can be found using network flows[63] and whose existence is guaranteed by Menger's theorem[193]). If a vertex s' is adjacent to the internal vertices of all of these paths, then we can exclude the pair (s', t) from further consideration because $\kappa_{s',t} \geq \kappa_{s,t}$. This step significantly accelerates the connectivity calculation for graphs with many high-degree vertices, and the connected components of $\mathcal{L}(\hat{\mathcal{S}}_t)$ typically exhibit this property.

2.2) Suppose that $L_{t,1}, ..., L_{t,r'}$ are the components produced at the previous step. Further subdivide each component $L_{t,i}$ as follows: first, find an embedding of the subgraph $\mathcal{L}(\hat{\mathcal{S}}_t)[L_{t,i}]$ into $\mathbb{R}^3$ using a force-directed graph drawing algorithm[69]; second, cluster the obtained embedded graph by a spectral clustering algorithm[131] using the largest Laplacian eigenvalue gap to estimate the number of clusters.

Each cluster produced at steps 2.1)-2.2) is supposed to correspond to a single haplotype.

3) For every cluster $L_{t,i}$, we examine the induced subgraph $\mathcal{G}_{t,i} = \mathcal{G}_t[\bigcup_{S_{t,j} \in L_{t,i}} S_{t,j}]$, which consists of the SAVs covered by the subgraphs that correspond to the vertices of $L_{t,i}$.

3.1) Suppose that $D_{t,i}$ is the degree sequence of $\mathcal{G}_{t,i}$. We cluster the elements of $D_{t,i}$ using the k-means algorithm and select the subset of vertices $C_{t,i}$ that corresponds to the cluster with the largest mean value. The goal of this procedure is to identify the "core" of $\mathcal{G}_{t,i}$ consisting of high-degree vertices. To choose the number of clusters k, we use the gap statistics[180].

3.2) Find the densest subgraph $H_{t,i}$ of $\mathcal{G}_{t,i}[C_i]$ using the LP formulation (3.11)-(3.14). If the subgraph is large enough (by default $|H_{t,i}| \geq 5$), then output $H_{t,i}$ as an inferred haplotype.

In addition to the set of haplotypes $\mathcal{H}_t$, Algorithm 2 returns a *support* $s(H_{t,i})$ for each inferred haplotype, that is defined as a relative number of elements (i.e. candidate dense subgraphs) in the cluster $L_{t,i}$: $s(H_{t,i}) = \frac{|L_{t,i}|}{\sum_j |L_{t,j}|}$.

The entire computational framework based on methods described in Subsections 3.1.2-3.1.4 is called HELEN (Heralding Emerging Lineages in Epistatic Networks).

3.2 Results

3.2.1 Data

Genomic data and associated metadata analyzed in this study were obtained from GISAID[167]. Our focus was on analyzing amino acid genomic variants of the SARS-CoV-2 spike protein, which is used for identifying Variants of Concern (VOC) and Variants of Interest (VOI) by standard genomic surveillance tools adopted by WHO[138]. We extracted the spike protein alignment from the whole genome multiple sequence alignment, replacing ambiguous characters with gaps, and

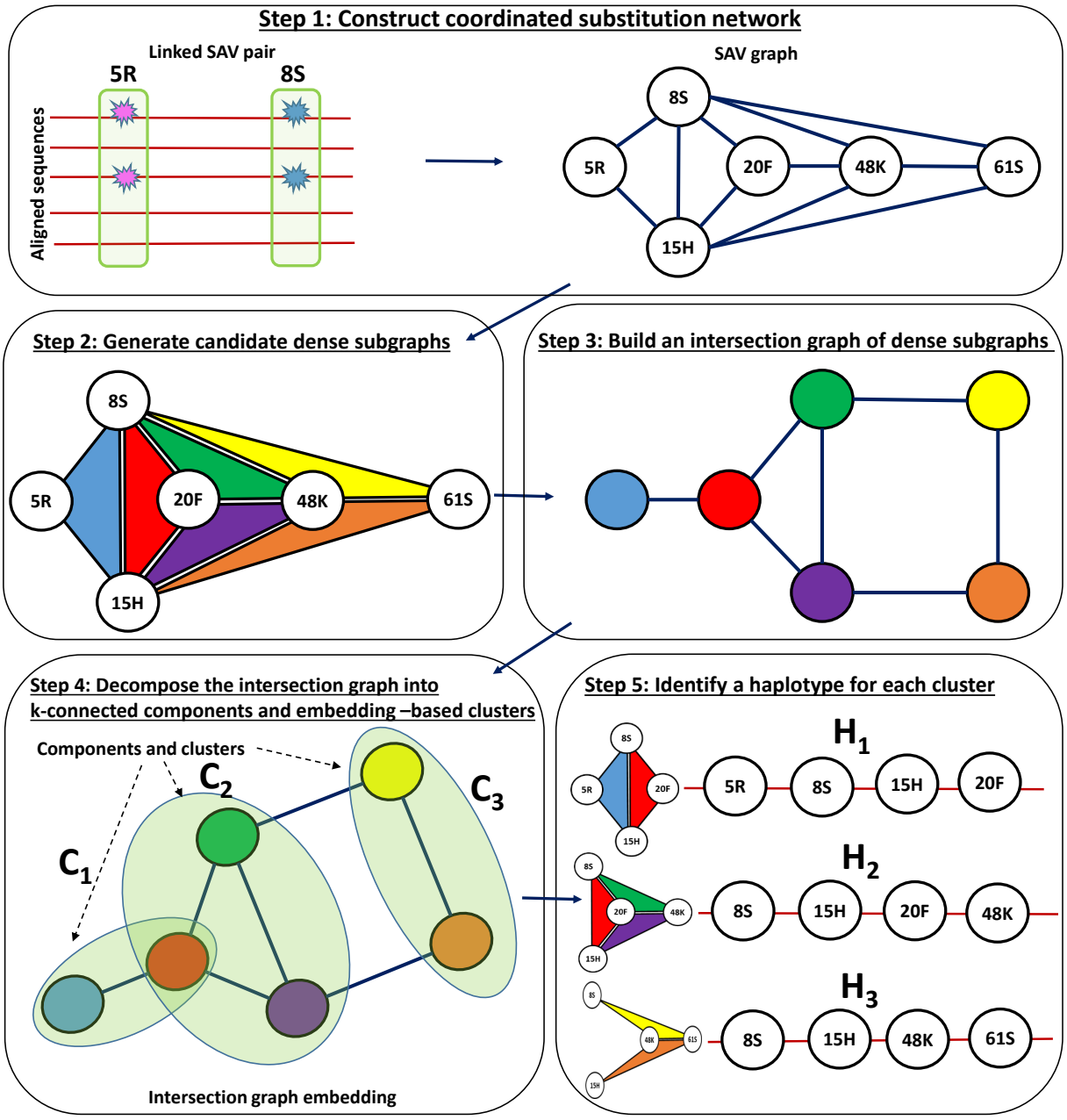

Figure 3.2 General scheme of HELEN. **Step 1**: construction of a coordinated substitution network (CSN) from aligned sequences. **Step 2**: generation of candidate dense subgraphs of CSN (highlighted in different colors). **Step 3**: construction of an intersection graph of subgraphs. Each colored vertex represents a subgraph of the same color; two vertices are adjacent whenever the corresponding subgraphs have sufficiently many common vertices (in this example - two). **Step 4**: decomposition of the intersection graph into clusters (depicted as ovals). Each cluster reflects a single haplotype. **Step 5**: construction of the haplotype for each cluster. The haplotype is found as a densest community in the union of the CSN subgraphs forming that cluster (e.g. the haplotype H_1 is found as the union of the blue and the red subgraphs that form the cluster C_1).

focused solely on SAVs while ignoring long indels. In order to better validate the predictive power of our approach, especially with respect to the Omicron lineage, we analyzed only sequences sampled before November 1, 2021, approximately 1 month before the designation of Omicron as the Variant of Concern by WHO). For defining VOCs and VOIs, we used the notations and lists of SAVs established by WHO[179]: a variant defined by SAVs at k fixed genomic positions was associated with a k-haplotype with minor alleles (with respect to the standard Wuhan-Hu-1 (NC_045512.2) reference) at that positions. Variants epsilon (B.1.427), iota (B.1.526) and zeta (P.2), defined by $3-4$ SAV, were excluded due to their short lengths.

The detection of linked pairs of SAVs and dense communities in coordinated substitution networks is affected by the number of sequences. Thus we focused on data from countries with the largest sample sizes, while maintaining geographic diversity. To do this, we selected two countries per continent (excluding Oceania) with the largest numbers of spike amino acid sequences sampled over the considered time period: the United Kingdom and Germany for Europe, USA and Canada for North America, Brazil and Peru for South America, South Africa and Kenya for Africa, and Japan and India for Asia. Additionally, we included Australia to represent Oceania and 5 extra countries with the largest samples, namely France, Denmark, Sweden, Spain, and Italy. Sequences from the selected countries were identified using GISAID metadata and analyzed separately. Thus, a total of 160 test cases (16 countries $\times$ 10 VOCs/VOIs) have been considered. Figure 3.3a shows the analyzed sample sizes, which were not distributed uniformly, with the USA and United Kingdom accounting for approximately 64% of all sequences.

3.2.2 The structure of S-gene coordinated substitution networks

We utilized the method outlined in Subsection 3.1.2 to construct coordinated substitution networks for 16 countries at 37 uniformly distributed time points between May 1, 2020, and November 1, 2021 (with a 14-day difference between consecutive points). Initially, we evaluated the basic properties of these networks. We found that the majority of networks contained a single ”giant” connected component that could include up to 70% of the vertices. Other connected components were significantly smaller ($p < 10^{-100}$, Kolmogorov-Smirnov test) and made up an average of 0.3% of the network size (Figure 3.3b). Most of these smaller components consisted of isolated vertices.

Coordinated substitution networks of the S-gene tend to gradually evolve towards becoming scale-free, with a right-skewed power-law degree distribution. This type of network structure is often a result of a preferential attachment process, where a new vertex joining the network has a higher probability of connecting to an existing vertex with a higher degree. Indeed, to determine the best distribution fit for the observed degree distribution of the networks, we fitted negative binomial, beta negative binomial, Poisson, Yule-Simon, Generalized Pareto, and Pareto distributions, and compared their goodness of fit using the Bayesian Information Criteria. We found that the Yule-Simon, Pareto, and generalized Pareto distributions, all describing a power-law, provided the best fit for 50.3%, 21.1%, and 16.4% of networks, respectively. Additionally, in all countries, the Yule-Simon distribution eventually became the best fit for the latest networks, i.e., for all networks sampled after a specific date t^* (with the median date being January 11, 2021).

The aforementioned observations indicate that the temporal networks inferred in this study

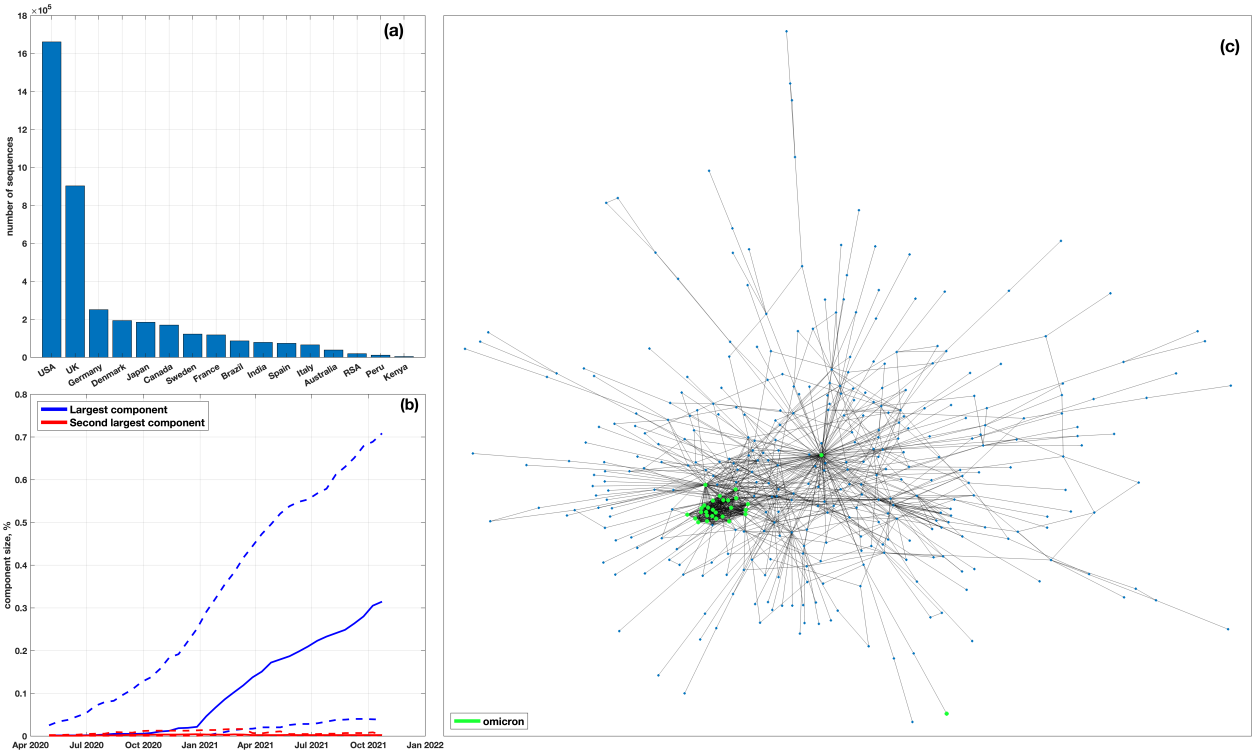

Figure 3.3 (a) Numbers of analyzed spike amino acid sequences per country. (b) Relative sizes of the largest and second largest connected components of coordinated substitution networks over time. Solid and dashed lines depict median and maximum/minimum values over 16 countries at each time point, respectively. (c) An example of a giant component of a coordinated substitution network for the USA on January 11, 2021. The vertices highlighted in green correspond to SAVs of the Omicron variant (lineage B.1.1.529.1). Most of these SAVs form a dense community, which was observed 320 days before the WHO designated the variant, emphasizing the key discovery and an algorithmic concept in this study.

have a sufficiently rich community structure[139,168] that can be analyzed and utilized to evaluate and forecast the SARS-CoV-2 evolutionary dynamics.

3.2.3 Dense communities in S-gene coordinated substitution networks as indicators of variant emergence

We analyzed communities within temporal coordinated substitution networks in search for evidence in support of the following hypotheses:

(H1) known VOCs/VOIs emerge as dense communities in temporal coordinated substitution networks;

(H2) conversely, dense communities within temporal coordinated substitution networks correspond to haplotypes with altered phenotypes;

(H3) such communities can be detected before the corresponding lineages achieve significant frequencies.

To validate the hypotheses (H1)-(H3), we used a two-pronged approach. First, we performed a retrospective statistical analysis of densities of known VOCs and VOIs in temporal coordinated substitution networks. Second, we evaluated the ability to accurately infer haplotypes with altered transmissibilities, both known and unknown, from collections of candidate dense communities. We specifically assessed the promptness of identifying emerging viral haplotypes as dense communities, by measuring so-called *forecasting depth*. This quantitative measure is defined as the time between the first variant call and the occurrence of a specific epidemiological benchmark event b. In this study, we used two benchmark events: the variant's designation by WHO ($b = des$) and the moment its prevalence reaches 1% ($b = prev$, the similar benchmark was used in[117])[1]. The value of $FD^b(h)$ can be positive or negative, thus indicating early or late prediction, respectively.

[1]The event was assigned to the last time point if the variant's prevalence always stays below 1%

It's worth noting that the presence of a viral variant as a dense community does not necessarily indicate its circulation at that time. In the context of this study's model, this fact should be rather interpreted as an indication that the corresponding SAVs are linked densely enough to suggest the variant's viability. In particular, detecting the variant as a dense community in a particular country at an early time point does not necessarily mean that the variant originated there. As demonstrated below, while there are instances where this is true, more often the variants are detected earlier in countries with larger sample sizes that provide greater statistical power for inferring coordinated substitutions.

3.2.3.1 VOCs/VOIs as communities in coordinated substitution networks

To validate hypotheses (H1)-(H3), we estimated density-based p-values of known VOCs and VOIs for each country and each time point using the algorithm described in Subsection 3.1.3. The algorithm produces uniform samples of connected communities of each temporal epistatic network, and compares their densities with those of the VOCs/VOIs to calculate p-values. As a result, for each country and each VOC/VOI we obtained a time series of p-values. The series were adjusted by calculating FDR and applying the Benjamini-Hochberg procedure[17]. The resulting time series of adjusted p-values are illustrated in Figure 3.4A and Supplemental Figures A.1-A.10.

Our analysis of time series data showed that a significant proportion of cases exhibited variant expansion either succeeding or concurrent with a decrease in density-based p-values. To quantify this relationship, we employed sample cross-correlation[19] to measure the connection between p-values and variant prevalences throughout the growth period of the variant. We considered a range of positive and negative lags for the prevalence series in relation to p-value series and identified the

optimal lag l^* with the maximum absolute cross-correlation.

In 75% of all test cases, we detected a non-negative optimal lag and a medium-to-strong statistically significant negative correlation between p-values and lagged prevalences (95% CI for ρ: $(-0.93, -0.41)$, 95% CI for l^* (in days): $(0, 140)$). Focusing solely on VOCs, we observed this effect in 86% of cases (95% CI for ρ: $(-0.88, -0.37)$, 95% CI for l^*: $(0, 140)$).

We defined a variant as *"significantly dense"* when its adjusted p-value falls below 0.05 and at least 80% of its SAVs belong to the network's giant component. In our analysis, 52% of VOCs/VOIs, analyzed separately for different countries, became significantly dense at some moment in time. This percentage increased to 76% when only considering VOCs. Moreover, these variants were identified as significantly dense at low cumulative frequencies (median value $\mu = 4.2 \cdot 10^{-4}$, Figure 3.4d) and low prevalences ($\mu = 1.4 \cdot 10^{-3}$, Figure 3.4e).

We assessed forecasting depths, FD^{prev} and FD^{des}, with respect to times when the variants reached significant density. In general, VOCs/VOIs that achieved significant density tended to do so early. In particular, such variants were identified before reaching 1% prevalence in 64% of cases and before WHO designation times in 46% of cases. For early calls (i.e. given that $FD^{\text{prev}} > 0$ or $FD^{\text{des}} > 0$), the median forecasting depths were 120 and 78 days, respectively.

In genomic surveillance, decisions are typically made based on agglomerate information from multiple countries. In this context, it is important to note that all Variants of Concern (VOCs) and Variants of Interest (VOIs) have positive forecasting depths (FD^{prev}) in at least one country (Figure 3.4a); the same applies to FD^{des} with the exception of the theta variant (Figure 3.4b).

In particular, the Omicron variant (lineage B.1.1.529.1) becomes significantly dense in 7 coun-

tries as early as the beginning of 2021, with forecasting depths ranging from 199 to 319 days for FD^{des} and 165 to 285 days for FD^{prev}. Notably, all predictions were made before the actual Omicron haplotypes emerged, at a cumulative frequency of 0. The Delta variant (B.1.617.2) serves as another example of multiple early predictions, as it becomes significantly dense in ten countries ($FD^{\text{des}} \in [15, 360]$ and $FD^{\text{prev}} \in [30, 375]$).

Sample size seems to significantly impact haplotype detection. A strong positive correlation exists between the number of significantly dense VOCs/VOIs and the number of sequences per country ($\rho = 0.71$, $p < 0.01$). Specifically, in the United States, which has the highest number of sequences, all 10 variants reached significant density.

3.2.3.2 Inference of viral variants as dense communities in coordinated substitution networks

The most straightforward way to partially assess the validity of hypotheses (H2)-(H3) is to retrieve the densest subnetworks of coordinated substitution networks and compare them to known SARS-CoV-2 variants. This task is made easier by the fact that finding the densest subgraphs, based on our density definition, is a polynomially solvable problem (see Subsection 3.1.4). We used the f-score as a metric for detection accuracy, which in our context is defined as:

$$R_{t,i} = \frac{|C_t \cap V_i|}{|V_i|}, \quad P_{t,i} = \frac{|C_t \cap V_i|}{|C_t|}, \quad F_{t,i} = 2\frac{R_{t,i} \cdot P_{t,i}}{R_{t,i} + P_{t,i}} \tag{3.20}$$

Here $R_{t,i}$, $P_{t,i}$ and $F_{t,i}$ are the recall, precision and f-score for the SAVs of the VOC V_i found within the dense community C_t at the time t.

In total, 28% of densest communities were at least 80% identical to the known variants, all of

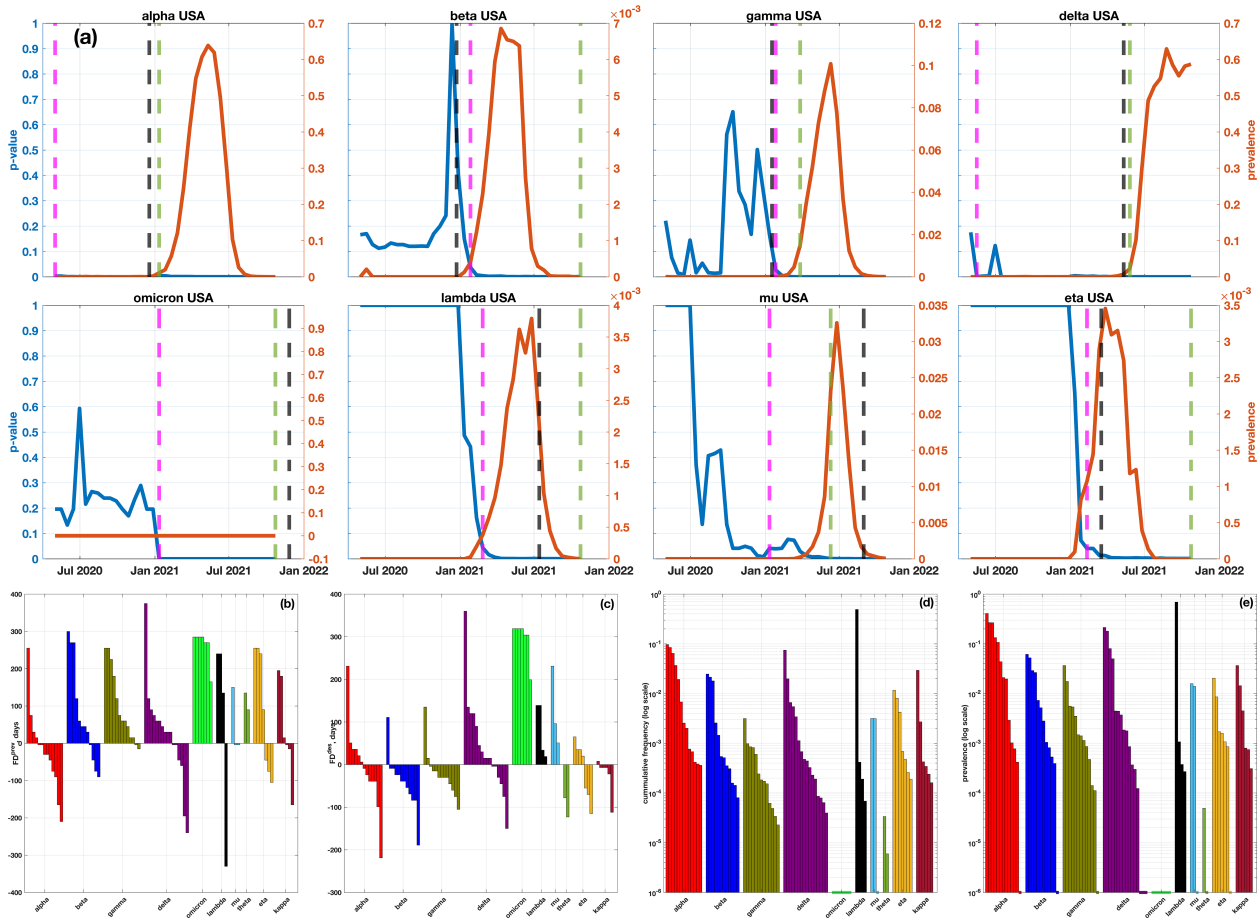

Figure 3.4 Density-based adjusted p-values of VOCs/VOIs. (a) p-values (blue) and prevalences (red) of 8 VOCs and VOIs in the USA coordinated substitution networks (refer to appendix A for plots of all VOCs/VOIs across all countries). Black, green, and magenta lines represent the times of VOC designation, achieving 1% prevalence, and becoming significantly dense, respectively. (b) and (c): Forecasting depths (y-axis) in relation to the 1% prevalence time and WHO designation time for each analyzed VOC/VOI across different countries. (d) and (e): Cumulative frequencies and prevalences for VOCs/VOIs across various countries at the times when they become significantly dense (in a logarithmic scale). Dashed lines at the bottom of the plot indicate that the variants reached significant density at frequencies/prevalences of 0.

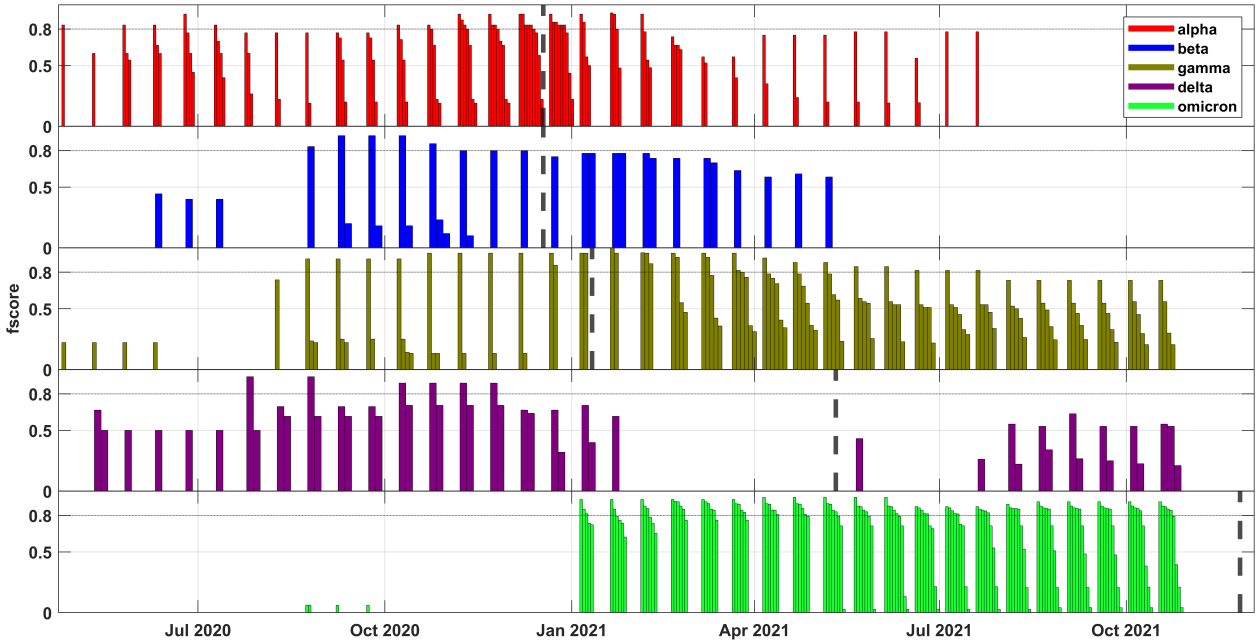

Figure 3.5 Comparison between VOCs and densest subnetworks of temporal coordinated substitution networks (results for individual countries are shown in Figure A.11-A.13). Each bar plot depicts the comparison results for a particular VOC; at each time point, bars correspond to the densest subgraphs from different countries closest to that VOC, and the bar heights are equal to the respective f-scores. Colored dashed lines mark times when the VOCs were designated by WHO.

which were VOCs. Notably, 86% of these communities were identified before the VOCs were officially designated by WHO, and 67% before the variants reached a 1% prevalence (Figure 3.5 and Figure A.14). Furthermore, these communities emerged when the corresponding VOC haplotypes had low cumulative frequencies (median value $\mu_f = 3.8 \cdot 10^{-4}$, Figure A.14c) and low prevalences (median value $\mu_p = 7.8 \cdot 10^{-4}$, Figure A.14d). Every VOC was detected with at least 0.8 accuracy in at least one country as early as $231, 111, 135, 285$ and 319 days before their designation times, and $255, 30, 255, 300$ and 319 days before achieving 1% prevalences, respectively (median values $FD^{\text{des}} = 111$ and $FD^{\text{prev}} = 60$, Figure 3.5 and Figure A.11-A.14). The most prominent example is the Omicron haplotype, which corresponds to 94 of the densest subnetworks across six countries.

While the examination of the densest subnetworks lends support to hypotheses (H2)-(H3), a more advanced algorithmic approach is essential for a comprehensive forecasting framework, as well as for stronger hypotheses confirmation. Indeed, generally, only a single densest subnetwork can be constructed per time point, even though multiple haplotypes with altered phenotypes might coexist at each specific moment. Additionally, we observed that, as coordinated substitution networks become denser over time, the densest subnetworks expand and may ultimately encompass several haplotypes, leading to decreased variant inference accuracy.

To overcome these problems, we developed a more complex algorithm for inferring viral haplotypes as dense network communities (Subsection 3.1.4). Briefly, the algorithm generates a pool of distinct dense subnetworks of varied sizes, partitions them into clusters, and assembles a haplotype from each cluster using graph-theoretical techniques. For every assembled haplotype, the algorithm also returns its *support* defined as the percentage of candidate subnetworks corresponding to that haplotype. As before, we used a 80% f-score threshold to declare variant detection.

The proposed algorithm demonstrated greater sensitivity in detecting known SARS-CoV-2 variants compared to the densest subgraph-based method (Figure 3.6). Specifically, it identified 90% (9 out of 10) of the analyzed variants in at least one country, with the Theta variant being the only exception. All Variants of Concern (VOCs) that were spreading during the study period (Alpha, Beta, Gamma, and Delta variants) were detected in 12-16 (out of 16) countries, while the Omicron variant was found in 6 countries.

A significant proportion of these detections were early, with 67% of VOCs/VOIs first identified before reaching a 1% prevalence in their respective countries and 49% detected prior to the WHO

designation times. In absolute terms, this represents a 2-fold and 2.9-fold increase in early detections compared to the densest subgraph-based method. When first detected, the median variant frequency was $\mu_f = 4.8 \cdot 10^{-4}$ (Figure 3.6d) and the median variant prevalence was $\mu_d = 2.2 \cdot 10^{-3}$, Figure 3.6e).

Concerning forecasting depths, 8 out of 10 known variants exhibited non-negative FD^{prev}, and 9 out of 10 showed non-negative FD^{des} in at least one country (Figure 3.6b,c). Specifically, all VOCs had positive forecasting depths and were detected as early as 231, 111, 150, 360, and 319 days before their designation times and 255, 300, 255, 375, and 319 days before reaching 1% prevalence, respectively (median values given the early prediction: $FD^{\text{des}} = 108$ and $FD^{\text{prev}} = 75$).

While the forecasting results for VOIs were somewhat less remarkable, Lambda, Mu, Eta, and Kappa variants were first identified as early as 124, 36, 5, and 23 days before WHO designation and 195, -45, 210, and 0 days before attaining a prevalence of 1% (median values given the early prediction: $FD^{\text{des}} = 29.5$ and $FD^{\text{prev}} = 195$).

Similar to the case with significantly dense subgraphs, sample sizes, and geographic diversity influence variant detection. A strong positive correlation was observed between the number of sequences per country and the number of variants with positive forecasting depths ($\rho = 0.80$, $p < 0.01$ for FD^{des} and $\rho = 0.69$, $p < 0.01$ for FD^{prev}). Some of the earliest forecasts, although not all, were made in the countries of origin for specific variants: notably, Beta, Gamma, and Lambda variants were detected in South Africa, Brazil, and Peru 111, 150, and 124 days before their designation times (Figure A.15-A.20).

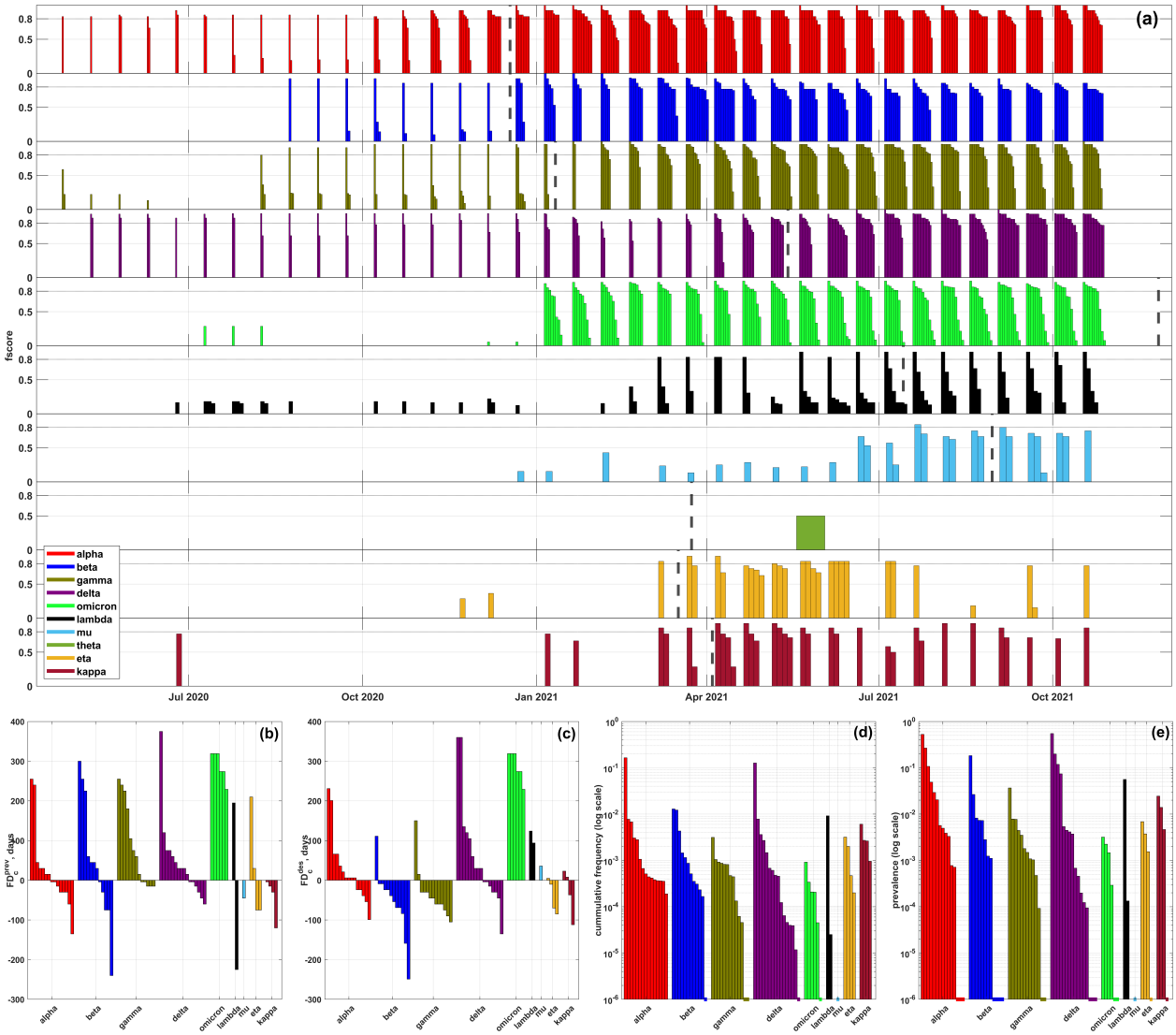

Figure 3.6 **(a) Summary of comparison between VOCs/VOIs and inferred haplotypes** (results for individual countries are shown on Figure S15-S20). Each bar plot depicts the comparison results for a particular VOC/VOI; at each time point, bars correspond to inferred haplotypes from different countries closest to that VOC, and the bar heights are equal to the respective f-scores. Colored dashed lines mark times when the VOCs were designated by WHO. **(b) and (c): forecasting depths** (y-axis) with respect to the 1% prevalence time and WHO designation time for each analyzed VOCs/VOIs over different countries. **(d) and (e): cumulative frequencies and prevalences** of VOCs/VOIs over different countries at first variant call times (in logarithmic scale). Dashed lines at the bottom of the plot signify that the corresponding variants were detected at cumulative frequencies or prevalences 0.

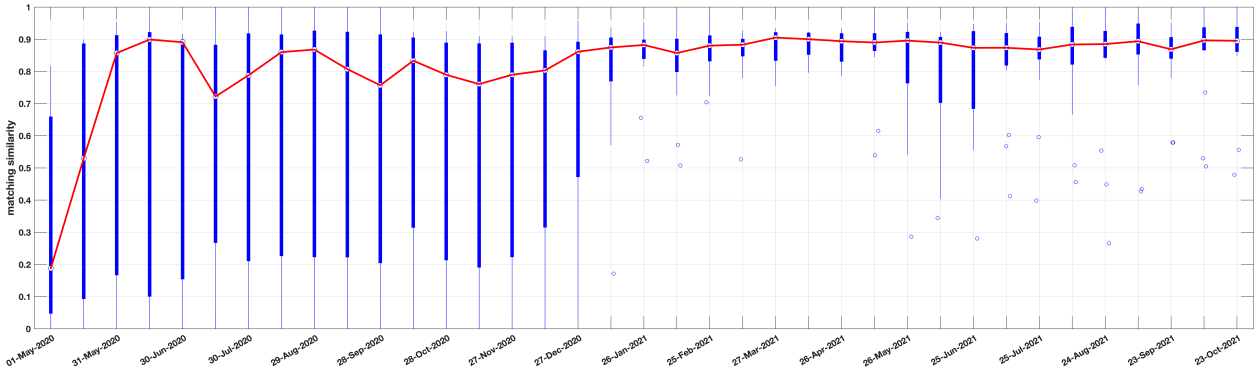

Figure 3.7 Precision of haplotype inference. Blue box plot: summary statistics of matching similarity at each time point over different countries. Red: median matching similarity over time.

To assess the precision of HELEN, it is important to consider that the true positive network communities identified by the algorithm might not only correspond to known VOCs/VOIs but also to variants exhibiting increased transmissibility that failed to become VOC/VOI due to factors such as genetic drift or containment through public health measures before achieving a high global prevalence. Consequently, we classify a haplotype v identified by HELEN at a specific time as *spreading*, if v is a known VOC/VOI or if the prevalence of variants highly similar to v has increased or will increase by a factor of 10 in the past or future. Note that a similar fold-based criterion was employed to define spreading mutations in[117]. A variant v' is considered highly similar to v if it contains at least 80% of v's SAVs; this definition encompasses variants genetically close to v and their descendants.

We measure precision using the *matching similarity* metric, denoted as $A_{I \to S}$. This metric evaluates the agreement between inferred haplotypes (I) and spreading haplotypes (S) by taking into account haplotype support as a proxy for haplotype call confidence and measuring the extent to which inferred haplotypes, weighted by their support ($\sigma_i : i \in I$), are matched by their

nearest spreading haplotypes. Formally, the matching similarity is the average f-score for inferred haplotypes in relation to their closest spreading haplotypes:

$$A_{I \to S} = \sum_{i \in I} \sigma_i \max_{s \in S} f_{i,s} \tag{3.21}$$

A similar measure, in the reverse form of a *matching error*, was used, e.g., in[100].

The summary statistics for matching similarity at each time point across different countries are summarized in Figure 3.7. HELEN achieved a median matching similarity above 75% from July 30, 2020, and above 85% - from December 27, 2020. Initially, there was a considerable variation in matching accuracy among countries, but it noticeably declined by early 2021. These observations are associated with the density dynamics of coordinated substitution networks in different countries, whereas the precision increases as more epistatically linked SAVs are identified.

Finally, it is noteworthy to compare the accuracy results of this study with those of[128], that similarly identified clusters of concordantly evolving spike protein sites in coordinated substitution networks using an alternative approach. That study identified 13 clusters, with f-scores ranging from 0% to 66.7% (median value $\mu = 4.8\%$) in relation to the nearest VOCs. Conversely, VOCs' f-scores in relation to the closest clusters spanned from 13.3% to 66.7% (median value $\mu = 16.7\%$). Specifically, 30% and 50% of Alpha variant sites were part of two clusters. Beta and Delta variant sites did not cluster together. Two distinct clusters covered 33.3% and 17.7% of Gamma variant sites, while 11%, 5.6% and 5.6% of Omicron sites were distributed across three clusters.

3.2.4 Running time and scalability

The computational methods employed in this study are reasonably efficient and scale to millions of sequences. The largest dataset analyzed, consisting of approximately $1.66 \cdot 10^6$ USA sequences sampled up to the final time point, provides an upper bound for the running times. For this dataset, constructing the coordinated substitution network took ~ 1 hour, estimating the p-values of 10 VOCs/VOIs took ~ 1.8 hours, and inferring viral haplotypes took ~ 38.6 hours. These computations were carried out on a workstation equipped with a 3GHz Intel Xeon E5 CPU and 64GB of RAM.

3.3 Discussion.

This study explores the hypothesis that viral variants with higher transmissibility can be associated with dense communities in coordinated substitution networks. Specifically, we investigated this idea in the context of SARS-CoV-2 spike protein genomic variants and found strong support for it. Our results indicate that network density can serve as a dependable indicator for the timely detection or prediction of emerging SARS-CoV-2 variants. As a result, we proposed an accurate, interpretable, and scalable method that can anticipate emerging SARS-CoV-2 haplotypes several months in advance, leading to early detection and improved forecasting.

These results were obtained using a synthetic approach that combines methods from statistics, combinatorial optimization, and population genetics. Firstly, we employed a sensitive statistical test that relies on a quasispecies population genetics model to identify linked pairs of SAVs that are jointly observed more often than expected if the corresponding 2-haplotype is inviable. This

method allowed us to construct coordinated substitution networks with rich community structures, providing a foundation for meaningful network-based inference. Secondly, we validated our hypothesis by estimating network density-based p-values of SARS-CoV-2 haplotypes. This allowed us to identify haplotypes with low p-values as potential variants of concern and demonstrate that known VOCs achieve low p-values significantly earlier than they reach frequencies high enough to be detected using conventional methods. Lastly, we utilized these findings to design an algorithm for the early detection of viral variants that identifies dense communities of SAV alleles and combines them into haplotypes. We demonstrate the efficacy of this algorithm by retrospectively identifying known VOCs and VOIs with high accuracy up to 10-12 months before they reached high prevalence and were designated by the WHO.

Compared to traditional phylogenetic lineage tracing, the proposed methodology offers several advantages. In particular, it can detect viral variants as dense communities at very low frequencies or even when actual variant sequences are not sampled - the latter is possible when there are sufficiently many well-covered variant's SAV pairs. This feature is naturally inherited from our prior methods[8,100] for reconstructing intra-host viral populations from noisy NGS data, which have demonstrated the ability to accurately detect viral haplotypes with frequencies as low as the level of sequencing noise. Additionally, the computational complexity of our network-based methods is a function of genome length rather than a sequence number. For SARS-CoV-2 data, the number of available sequences in GISAID is up to 4 orders of magnitude larger than the number of amino acid positions in the SARS-CoV-2 s-gene ($\sim 1.5 \cdot 10^7$ sequences versus $1.27 \cdot 10^3$ amino acid positions). This feature makes the proposed algorithms considerably more scalable than phylogenetic

methods.

It is important to note that there are limitations to this study, as the comprehensive forecasting of viral evolution is inherently an intractable problem. While the proposed methods have shown promising prediction results, caution should be exercised when interpreting them. Our findings by no means suggest that viral evolution is a deterministic process that can be predicted using mechanistic models. Instead, they demonstrate how to identify several potential evolutionary trajectories among exponentially many possibilities. These trajectories can guide further investigation and prioritization of functional screening. Moreover, the links between SAVs identified by HELEN represent *putative or potential* positive epistatic interactions[152], and their primary purpose is to serve as features for our prediction model. These links should be viewed as a statistical ensemble rather than individually, with our findings suggesting that haplotypes with altered phenotypes exhibit a significantly higher number of potential epistatic pairs compared to background haplotypes. Consequently, research focused on examining the biological mechanisms of specific SARS-CoV-2 epistatic interactions should incorporate more comprehensive structural data.

The utilized coordinated substitution/epistasis model is another limitation of this study as it only considers the interactions between SAV pairs, thus reflecting *"pairwise" or "second-order epistasis"*. Although combinations of mutations can have more complex fitness effects involving higher orders of epistasis[188], this model is justifiable for several computational reasons. Firstly, it is the minimal model that enables the detection of multiple overlapping haplotypes, which is an improvement over the mutation independence assumption used in other studies[117] that, in general, only allows ranking and prioritization of mutations. Secondly, k-haplotypes with $k \geq 3$ may

not have sufficiently high frequencies to be detected, thereby affecting the method's predictive power. In contrast, pairs are always covered by more sequences and can be detected earlier. Lastly, accounting for higher-order combinations of mutations can increase the computational complexity of the problem while the second-order model remains computationally tractable.

Furthermore, our method is based solely on genomic data, and its effectiveness could be enhanced by incorporating epidemiological and structural biology data and models. Additionally, our results highlight the significance of robust and diverse sampling practices, as early detections were predominantly made in countries with larger sample sizes, and some variants were only detected early in their countries of origin.

We believe that the methodology proposed in this study is not limited to SARS-CoV-2 and can be extended to other pathogens. The high sensitivity of HELEN should make it particularly suitable for detecting emerging and circulating strains of pandemic viruses, such as HIV or Hepatitis C.

CHAPTER 4

Viral outbreak investigation and transmission history reconstruction

Continuing advances in sequencing technologies are vitalizing *genomic epidemiology* – an interdisciplinary area of research that uses analysis of pathogen genomes to understand how they evolve and spread[6,18]. Inference of transmission histories is one of the fundamental problems of genomic epidemiology and a major driving force behind new developments in the field. The list of existing transmission network reconstruction tools includes Outbreaker and Outbreaker 2[91,25], SeqTrack[92], SCOTTI[46], Phybreak[95], Bitrugs[195], BadTrIP[47], Phyloscanner[196], StrainHub[45], TransPhylo[53], STraTUS[81], TreeFix-TP[171], QUENTIN[170], VOICE[74], HIVTrace[102], GHOST[112], MicrobeTrace[24], SharpTNI[162], TiTUS[163], TNeT[51] and others[200,121,52,122,41,26,80]. These tools have been successfully applied for investigation of outbreaks and surveillance of transmission dynamics of HIV, hepatitis C (HCV), SARS, MERS, SARS-CoV-2 and other viruses[191,150,205,146,99,23]. The majority of existing methods (although by no means all of them) utilize the phylogenetic approach, where transmission network reconstruction is considered as a character optimization problem, with characters being infected hosts. This paper also follows this paradigm (see Figure 4.1).

The hallmark of viruses as species is an extremely high genomic diversity originating from their error-prone replication. As a result, each infected individual usually hosts a heterogeneous population of numerous genomic variants. The first generation of transmission inference methods largely ignored intra-host viral diversity and considered only a single sequence per host (usually consensus). Later, it has been demonstrated that taking viral diversity into account greatly enhances the predictive power of transmission inference algorithms[196,170,5,156,99]. In particular, it allows to detect

the viral evolution directionality in situations when a reliable phylogenetic rooting is not possible[170,156,74] – such situation is very common for HIV, HCV, and other long-standing epidemics, as well as for the regional epidemics of SARS-CoV-2 characterized by multiple introductions of the virus. First phylogenetic approaches to infer transmission directions using viral genomic diversity appeared independently in[156] and[74]. Later, the ideas of[156] and[74] were incorporated into full transmission network inference frameworks Phyloscanner[196] and QUENTIN[170], respectively. These tools were followed by TNeT[51], TiTUS[163], SharpTNI[162], BadTrIP[47], all of which are specifically tailored to take into account intra-host viral diversity.

Despite the significant progress achieved with the appearance of the next generation of transmission inference methods, a number of computational, modeling, and algorithmic challenges still need to be addressed.

1) Most recent tools utilize a maximum parsimony principle and many of them are based on various extensions of classical Fitch and Sankoff labeling algorithms. This is in part due to the need to scale to the sizes of large genomic datasets. The maximum likelihood or Bayesian phylogenetic models are richer and incorporate additional inferred temporal information that can be used for more accurate reconstruction of transmission links[125]. However, the methods based on parameter-rich models lead to computationally hard optimization problems. To find transmission networks and estimate other parameters, such methods mostly rely on Markov Chain Monte Carlo (MCMC) sampling from the model parameter space[47,170,46]. Given that the parameter spaces are enormous[81], such a strategy is computationally expensive and may produce sub-optimal results.

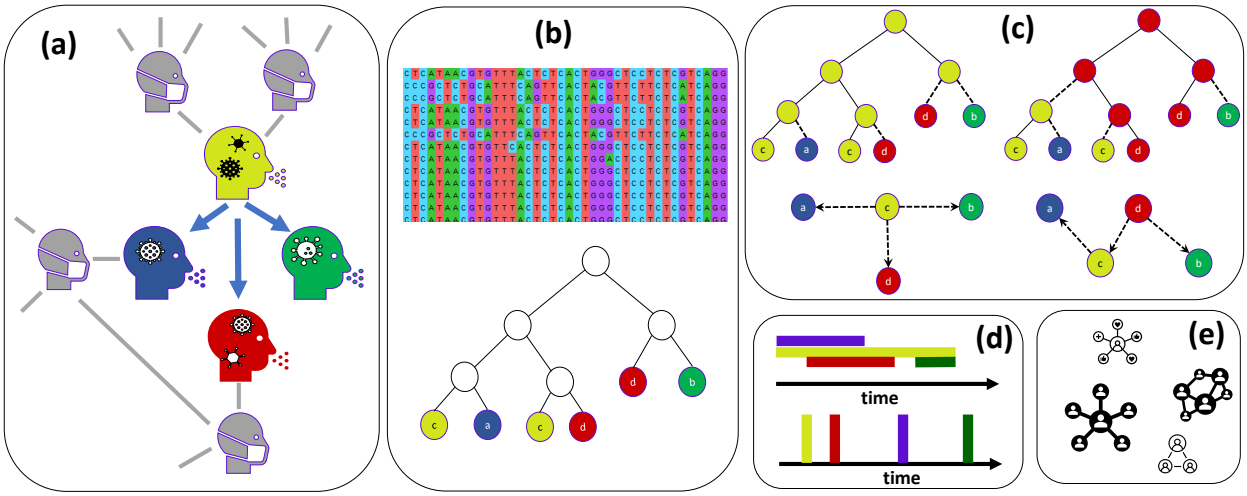

Figure 4.1 Approaches and challenges for transmission history reconstruction using genomic data. **(a)** *Example of a viral outbreak and its transmission network* consisting of 4 individuals (highlighted in light green, blue, dark green, and red) and 3 transmission links (blue arrows). The transmission network is part of a larger unobserved social network of contacts between susceptible individuals (the unobserved part is highlighted in gray). Social networks serve as conduits for the infection spread, and thus transmission networks reflect the properties of social networks. Due to the high virus mutation rates, each infected individual hosts a population of related but distinct viral genomic variants. **(b)** *First step of genomic epidemiology investigation.* Intra-host viral variants are sequenced, de-noised and aligned; the obtained viral haplotypes are used to construct a viral phylogeny. Leaves of this phylogeny correspond to sampled viral variants and labeled by their hosts (colors of the leaves correspond to the colors in (a)). **(c)** *Phylogenetic inference of transmission networks.* Labels of leaves are extended to internal nodes, and every tree edge with multi-labeled end nodes defines a transmission between the corresponding hosts. Two possible ancestral label assignments are depicted. Tree edges defining transmissions are dashed, the corresponding transmission network is shown below each assignment. Note that both assignments have the same number of such edges, i.e. the same parsimony score. Thus, parsimony does not allow to rank the obtained transmission networks. **(d)** *Resolution of phylogenetic ambiguities using case-specific epidemiological data* proposed in prior studies. One possibility is to consider patient exposure intervals (upper figure): in this example, the intervals for the red and green patients do not overlap, thus ruling out the second network containing a link between these patients. Another possibility is to take into account sampling times (lower figure): the light green patient was sampled earlier thus making more probable the first network, where it is a root. Unfortunately, such information often has limited use for many real outbreaks of HIV, HCV, SARS-CoV-2, etc. **(e)** *Resolution of phylogenetic ambiguities using the prior knowledge about social network properties*. We propose to integrate phylogenetic and random graph models: first, we sample transmission networks from the phylogeny-based distribution and then measure their agreement with the expected properties of the distribution of inter-host social networks. In this example, the depicted social network distribution favors the first candidate transmission network that has more "star-like" structure.

2) Several studies demonstrated that in many cases genomic data alone do not allow to resolve ambiguities in transmission network inference, and so the incorporation of additional evidence is necessary[89,182,91]. Such evidence most often comes in the form of case-specific

epidemiological information. However, the most common types of such information are useful only in particular settings. For example, many tools use sample collection times to identify the order of infections. However, HIV, HCV, and many other infections tend to be initially asymptomatic, and consequently, sampling times may not accurately reflect the actual infection times. Other tools rely on exposure intervals for the same purpose. However, in outbreaks with high transmission rates (e.g. in HIV/HCV outbreaks associated with injection drug use or during the global pandemic of SARS-CoV-2/Influenza), many susceptible hosts are almost constantly exposed to the virus, thus effectively making exposure intervals useless.

3) Most methods implicitly assume that transmission network edges are independent. Such an assumption is associated with *random mixing* models, which suppose that differences between individuals are negligible and any person can infect any other person with the same probability. However, this is not always the case, as, for example, certain hosts infect more people than an average individual[71].

In this study, we propose to address these challenges by integrating phylogenetic and random graph models. Our major idea is to bring into consideration the social component of the epidemics. Infectious diseases spread over the social networks of contacts between susceptible individuals, and transmission networks to a significant degree mirror the properties of these social networks[107,190,86,155]. Social networks are almost never known explicitly; however, their general features are well defined in network theory, sociology and classical epidemiology[129]. In light of this, we propose to infer transmission networks by integrating two components: the evolutionary

relationships between viral genomes represented by their phylogenies and the expected structural properties of inter-host social networks. Frequently cited properties of social contact networks include power law degree distribution, small diameter, modularity and presence of hubs[13,129]. All of them are reflected by network vertex degrees. Thus, we model social networks as random graphs with given expected degree distributions (EDDs). They are commonly scale-free[190,20], but our method can handle more specific EDDs of needle-sharing networks, sexual-contact networks or networks obtained by epidemiological contact tracing or respondent-driven sampling. The goal is to find transmission networks that are consistent with observed genomic data and have the highest probability to be subnetworks of random contact networks.

This methodology is implemented within a maximum likelihood algorithmic framework SOPHIE (SOcial and PHilogenetic Investigation of Epidemics). SOPHIE samples from the joint distribution of phylogeny ancestral traits defining transmission networks, estimates the probabilities that sampled networks are subgraphs of a random contact network and summarize them accordingly into the consensus network. This approach is scalable, accounts for intra-host diversity and accurately infers transmissions without case-specific epidemiological data. We applied SOPHIE to synthetic data simulated under different epidemiological and evolutionary scenarios, as well as to experimental data from epidemiologically curated HCV outbreaks. The experiments confirm the effectiveness of the proposed methodology.

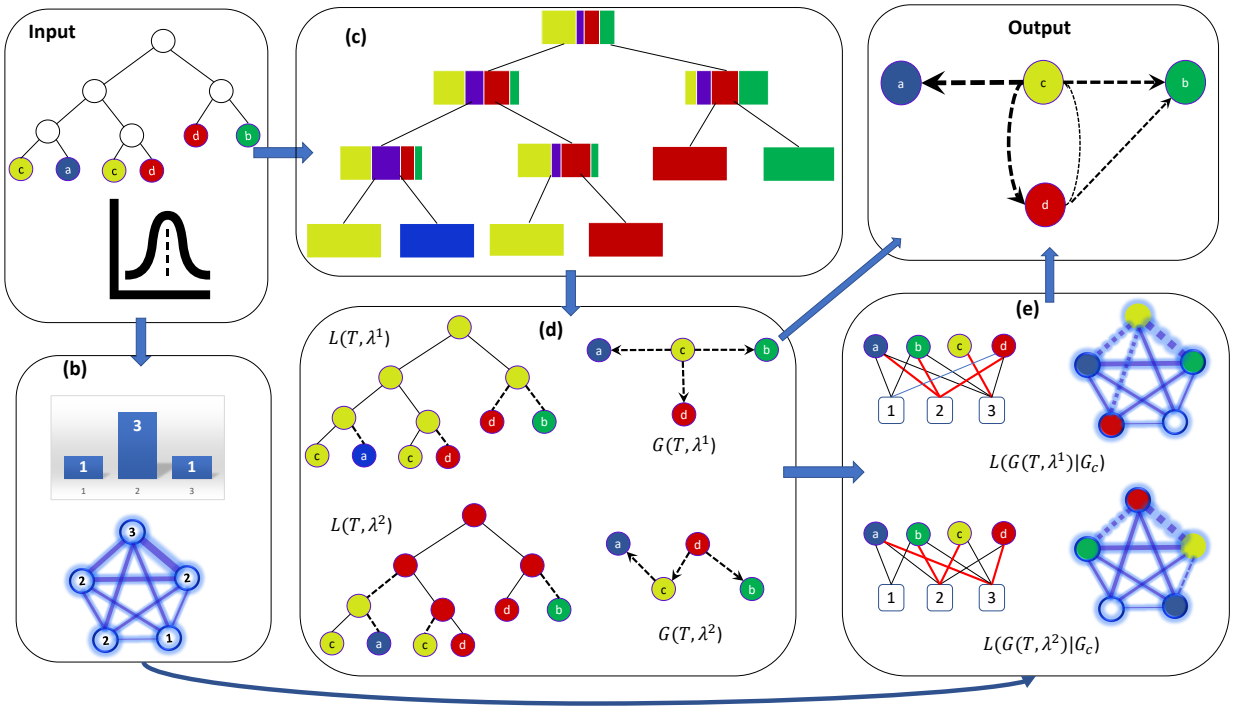

Figure 4.2 Joint phylogenetic and random graph-based approach for transmission history reconstruction implemented in SOPHIE. **Input:** a labeled phylogeny with leaves corresponding to viral haplotypes from 4 infected hosts (highlighted in different colors); expected degree distribution of a contact network that contains the true transmission network as a subgraph. **(b)** Generalized Random Graph (GRG) model of a contact network depicted as a complete graph with edge thicknesses proportional to their probabilities. It is accompanied by the expected degree counts of contact network vertices. **(c)** SOPHIE samples from the joint distribution of ancestral label assignments using dynamic programming. First, the algorithm performs a post-order traversal and calculates, for each internal node, conditional likelihoods of observing the labels of its descendants given a label of this node. On a figure, the widths of colored strips are proportional to the conditional likelihoods given the hosts with the corresponding color-codes. After all conditional likelihoods are calculated, the algorithm performs a pre-order traversal and samples a label for each node from the corresponding posterior distribution given its parent's sampled state (see Subsection 4.5.2.1). **(d)** Two sampled ancestral label assignments λ_1 and λ_2, the corresponding transmission networks and their phylogenetic likelihoods. Tree edges defining transmissions are dashed. The networks are obtained by contracting the tree nodes with the same labels. **(e)** SOPHIE calculates network likelihoods of sampled transmission networks by embedding them into random contact networks. To find an embedding, SOPHIE maps the transmission network vertices to their degrees in the contact network. It is done via the reduction to a generalized uncapacitated facility location problem with convex costs, where the hosts serve as clients and their possible expected degrees in – as facilities. On the left side of the panel, the instances of the facility location problem for two sampled networks are depicted. Optimal client assignments are highlighted in red, next to them the corresponding embeddings of transmission networks into contact networks are shown. See Subsection 4.5.2.2 for details. **Output:** a consensus of sampled transmission networks. Edges represent possible transmission links, their thicknesses are proportional to their inferred likelihood supports. See Subsection 4.5.2.3 for details.

.

4.1 Methods

We developed SOPHIE - a modeling and algorithmic framework to infer viral transmission networks from genomic data by integrating phylogenetic and random graph models. Within this framework, we define the *transmission network inference* problem as follows. We are given a time-labelled phylogeny $T = (V(T), E(T))$ with n_l leafs corresponding to viral haplotypes sampled from n_h infected hosts; each leaf u is assigned the label $\lambda_u \in [n_h]$ corresponding to its host. Such tree can be constructed using standard phylogenetic tools such as RAxML[173], PhyML[77] and IQ-Tree[132]. The goal is to extend $\boldsymbol{\lambda}$ to internal nodes in an optimal way. In this model, every multi-labelled tree edge uv corresponds to a direct or indirect transmission between the hosts λ_u and λ_v. Thus, the transmission network $G = G(T, \boldsymbol{\lambda})$ with the vertex set $V(G) = [n_h]$ can be constructed by contracting the vertices with the same label[80] (Figure 4.2). The simplest variant of this problem is the *maximum parsimony label inference* where the goal is to minimize the number of transmission events. It can be easily solved using e.g. Fitch or Sankoff algorithms[161,66] and their modifications. However, straightforward maximum parsimony approach alone often leads to epidemiologically unrealistic results[196]; furthermore, there are usually many most parsimonious solutions[50,163]. Within maximum likelihood framework, ancestral labels can be inferred using so-called "migration model"[159]. In this case, Fitch or Sankoff algorithms can be replaced by the dynamic programming algorithm of Pupko et.al.[145] or its extensions[159]. However, as mentioned above, phylogenetic signal alone can be insufficient for accurate transmission network reconstruction[89,182,91]. In particular, in the absence of reliable estimations of transmission rates between individual hosts, migration-based approaches have to rely on simple substitution models; as a result, similarly to the

case of maximum parsimony, the numbers of near-optimal solutions can be high.

In light of this, we extend a maximum likelihood approach by integrating a phylogenetic model with a model of social networks of susceptible individuals. Under this methodology, a transmission network is defined by two properties: it is a contraction of the phylogeny and, at the same time, a subgraph of a inter-host contact network of susceptible individuals (Figure 4.2). In reality, the contact network is not directly observed. Therefore, we model it as a random graph with the *expected degree distribution (EDD)*. EDD carries information about structural and spectral properties of contact networks[33,129,34], and can be adjusted to reflect specific epidemiological settings.

The general scheme of our approach is as follows (Figure 4.2):

1) We consider phylogeny node labels as discrete traits and sample from the joint distribution of label assignments under the selected substitution model (Subsection 4.5.2.1).

2) For each sampled label assignment $\boldsymbol{\lambda}$, we construct the corresponding transmission network $G(T, \boldsymbol{\lambda})$ and estimate its *network likelihood*, which is defined as the maximum probability that this network is a subgraph of a random contact network with the given EDD (Subsection 4.5.2.2)

3) Estimate the final transmission network as a weighted consensus of sampled networks. The edge weights here represent the inferred joint likelihood network-based and phylogeny-based likelihood support for the corresponding transmission links.

Each of these steps is described in detail in Star Methods section 4.5.

4.1.1 Algorithm benchmarking

We validated SOPHIE on synthetic and experimental data with known transmission networks. To evaluate the accuracy of inferred networks, we estimated sensitivity (i.e. the fraction of inferred transmission edges among true transmission edges), specificity (i.e. the fraction of true transmission edges among inferred transmission edges) and f-score (i.e. the harmonic mean of sensitivity and specificity). The latter parameter has been used as the principal evaluation metric.

In this study, SOPHIE was compared with Phyloscanner and TNet. Both methods are based on maximum parsimony principle: Phyloscanner reconstructs ancestral labels using a Sankoff algorithm with specially adjusted parsimony scores, while TNet uniformly samples from the space of most parsimonious label assignments that minimize the number of back transmissions. Other published phylogeny-based tools that account for intra-host viral diversity, TiTUS and BadTrIP, utilize case-specific exposure intervals as an additional source of information. Theoretically, in the absence of exact exposure dates, both tools can work with arbitrarily large exposure intervals. However, as noted by the authors of BadTrIP[47], such an assumption has a significant negative effect on in the accuracy of their method. We observed the similar effect for TiTUS: its average f-score was quite low (mostly within a range of $\sim 0.10 - 0.20$), thus suggesting that non-trivial exposure intervals are essential for it. Therefore, for the sake of fairness TiTUS and BadTrIP were excluded from further comparison.

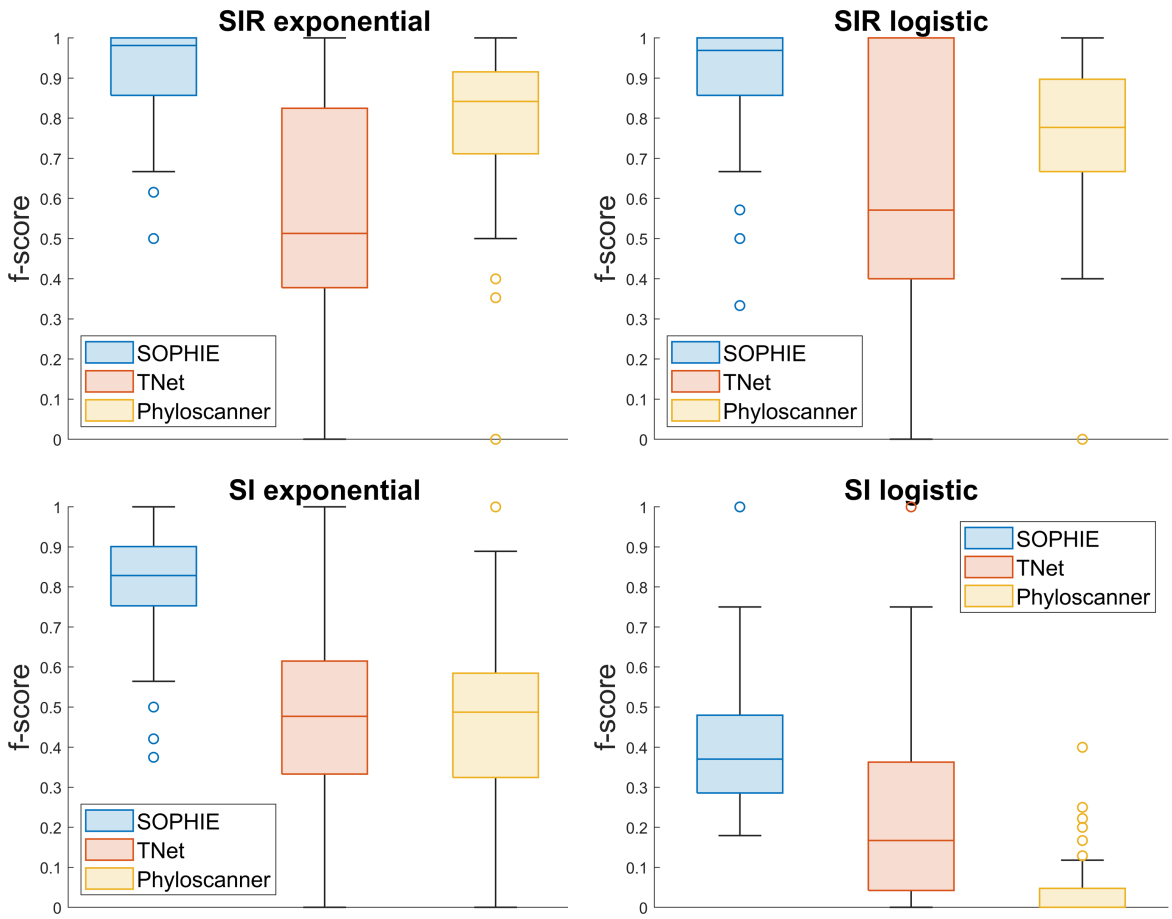

Figure 4.3 Comparative results of SOPHIE (best exponent), TNet and Phyloscanner on simulated data under different epidemiological and evolutionary scenarios with the true tree simulated by FAVITES

4.2 Data

4.2.0.1 Simulated data

To generate synthetic data, we used FAVITES[123] – a flexible tool that can simultaneously simulate viral sequences, phylogenies, contact networks and transmission networks under different evolutionary and epidemiological scenarios. In our case, we assumed that the virus spread over a contact

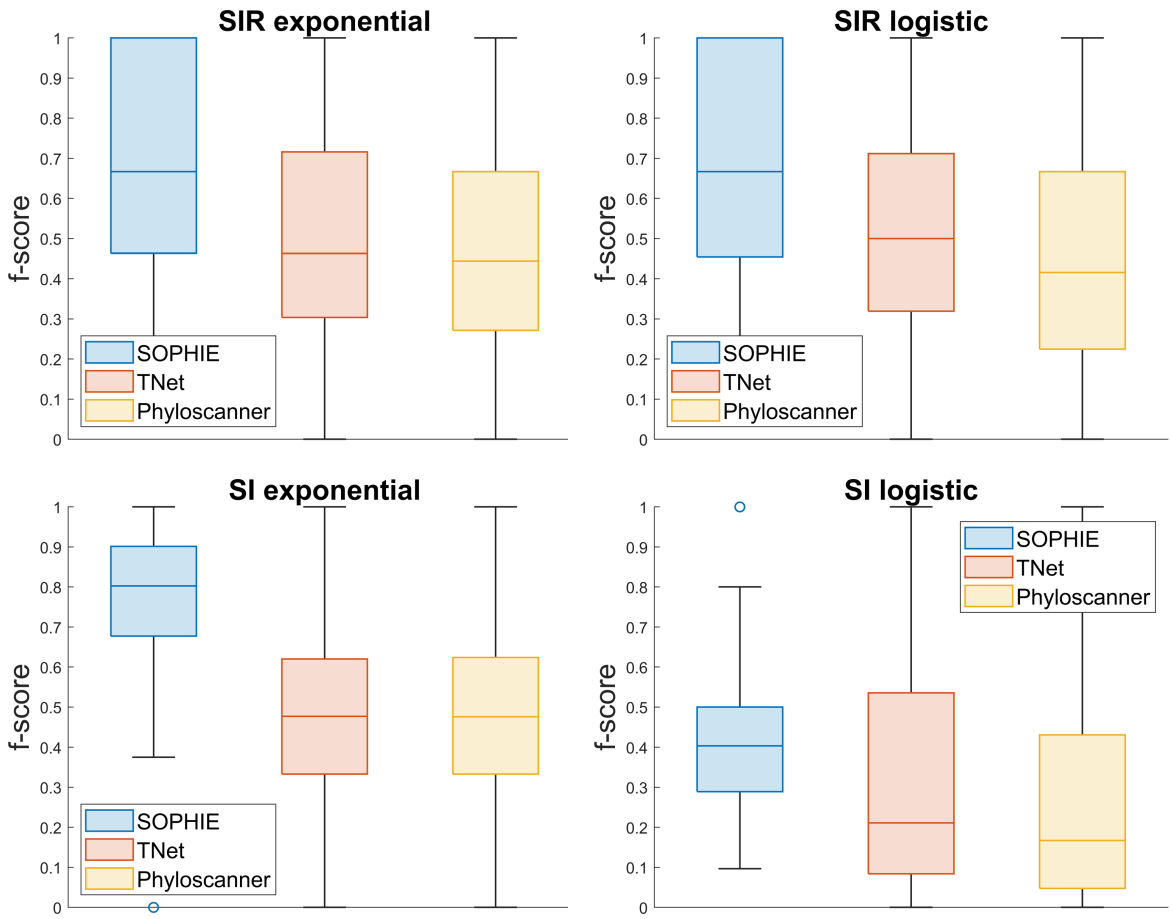

Figure 4.4 Comparative results of SOPHIE (best exponent), TNet and Phyloscanner on simulated data under different epidemiological and evolutionary scenarios with the tree reconstructed by RAxML

network of 100 susceptible individuals generated using the Barabasi-Albert model[13]. Transmission networks and data sampling were simulated under two epidemiological scenarios:

E1) Susceptible-Infected (SI) transmission model and simultaneous sampling of all infected individuals at the end of the simulation. This scenario corresponds to the typical settings of HIV or HCV outbreaks[146,141].

E2) Susceptible-Infected-Recovered (SIR) transmission model, with each individual sampling time being chosen from its infection time window. This scenario describes epidemics and surveillance of Influenza, SARS-CoV-2 and other viruses that are associated with acute rather than chronic infections.

Inside each host, viral phylogenies evolved under a coalescent model with two effective population size growth modes:

I1) Exponential effective population growth.

I2) Logistic effective population growth.

For each of the four combinations of scenarios E1-E2 and I1-I2, 100 simulated datasets have been generated, with 10 genomes sampled per infected host. For each dataset, we applied SOPHIE, Phyloscanner and TNet to two trees: a true phylogeny provided by FAVITES and a phylogeny reconstructed by RAxML[173]. For a network likelihood calculation with SOPHIE, we used a power-law distribution as an expected degree distribution. In this case, the algorithm has a power-law degree exponent α as a hyperparameter. We analyzed SOPHIE performance with the best exponent from the interval $(1, 2]$ and with the exponent randomly drawn from the gamma distribution with the mean 1.6. For each test instance, $100,000$ internal label assignments were sampled, and the final network calculated as a maximum-weight arborescence of the consensus network (see Subsection 4.5.2.3). Further details can be found in Star Methods section.

The results of SOPHIE evaluation and comparison with other methods are shown in Tables 4.1-4.2 and on Figures 4.3 - 4.4. First, the value of the exponent α does not significantly affect

the results. This demonstrates that accounting for the general shape of the expected degree distribution plays the most important role here, while guessing the best exponent allows for a moderate improvement. Second, for all eight experiments (four combinations of scenarios and 2 types of trees), we found that SOPHIE allows for a statistically significant improvement over TNet and Phyloscanner ($p < 0.05$, Kruskal-Wallis test). The average f-score of SOPHIE over all datasets is 0.71 (standard deviation 0.17) and can be as high as 0.92 and 0.90 (for SIR transmission models with the exponential and logistic coalescent and the true phylogeny). The average best absolute f-score improvement with respect to existing methods were 0.22 (standard deviation 0.09) over TNet and 0.25 (standard deviation 0.07) over Phyloscanner.

The accuracy of SOPHIE was negatively affected by the phylogenetic inference noise and was generally lower when RAxML tree was used. This effect is less pronounced for TNet and similarly pronounced for Phyloscanner. It is not surprising, since TNet, as a strictly parsimony-based method, depends only on the tree topology, while SOPHIE and Phyloscanner also utilize branch lengths. Still, the accuracy of SOPHIE for RAxML trees remains higher than for other tools.

The results of SOPHIE for different evolutionary and epidemiological scenarios are comparable, with the exception of the Susceptible-Infected transmission model with the logistic intra-host population growth. In that case, the accuracies of all methods were significantly lower.

As described in Methods, all algorithmic subroutines of SOPHIE are polynomial. Thus, the method is not too computationally expensive: the experimental average running time of SOPHIE on the analyzed data was 106.5s (standard deviation 285.4s). It somewhat slower than TNet (with

the running times measured in seconds) and Phyloscanner (that stops within 1-2 minutes), but it is to be expected, given that the SOPHIE's model is richer than for other tools.

4.2.0.2 Experimental data

We used a "gold standard" experimental dataset that has been previously utilized for benchmarking of transmission network inference algorithms in several studies[170,74,51]. It consists of 74 intra-host HCV populations sampled and sequenced during the investigation of 10 outbreaks by the Centers for Disease Control and Prevention. Viral populations contain from several dozen to several hundred sequences of lengths 264bp covering Hypervariable Region 1 (HVR1) of the HCV genome. In each outbreak, a single primary host identified by the investigators using epidemiological evidence infected all other hosts. Thus, transmission networks for that outbreaks are known.

Similarly to simulated data, the algorithms under consideration were applied to phylogenies reconstructed by RAxML. For all outbreaks, the uniform equilibrium label distribution, the rate $\mu = 1$ and the power-law exponent $\alpha = 2$ has been used. SOPHIE yielded the average f-score of 0.70, while TNet and Phyloscanner showed f-scores of 0.58 and 0.37, respectively (Table 4.1).

	True tree				RAxML tree				
	SIR exp	SIR log	SI exp	SI log	SIR exp	SIR log	SI exp	SI log	Real
SOPHIE (best α)	0.92	0.90	0.82	0.41	0.68	0.67	0.78	0.48	0.70
SOPHIE ($\alpha \sim \Gamma$)	0.89	0.86	0.75	0.33	0.63	0.61	0.73	0.42	
TNet	0.57	0.63	0.50	0.24	0.53	0.55	0.50	0.36	0.58
Phyloscanner	0.75	0.71	0.48	0.03	0.49	0.45	0.50	0.29	0.37

Table 4.1 Mean f-scores of SOPHIE, TNet, and Phyloscanner for different simulated and real datasets.

	True tree				RAxML tree			
	SIR exp	SIR log	SI exp	SI log	SIR exp	SIR log	SI exp	SI log
SOPHIE vs TNet	1.6e-12	1.4e-7	2.1e-19	7.1e-7	3.0e-3	4.7e-2	8.4e-16	4.7e-3
SOPHIE vs Phyloscanner	1.2e-4	6.3e-5	4.7e-21	0	4.2e-4	1.2e-4	6.3e-16	3.0e-6
TNet vs Phyloscanner	5.3e-3	4.4e-1	9.3e-1	1.3e-13	8.6e-1	1.9e-1	9.9e-1	1.9e-1

Table 4.2 p-values of multiple comparison for Kruskal-Wallis test.

4.3 Case study: HCV/HIV outbreak in rural Indiana, 2015

We utilized SOPHIE to analyze genomic data from the large HIV/HCV outbreak in Indiana[141,146,75,23,38]. The first 11 HIV infection cases associated with this outbreak have been discovered by the Indiana State Department of Health (ISDH) in a small rural community in Scott County, IN in early 2015. This triggered a further investigation by the ISDH and the CDC[38] that led to the detection of several hundred HIV and HCV infections and precipitated a declaration of a public health emergency by the state of Indiana[38]. The investigation linked the outbreak to unsafe injection use of the opioid oxymorphone[141], providing an important example of the rapid spread of viral infections associated with the nationwide epidemic of prescription opioid abuse[206,176].

Deep sequencing of intra-host viral populations has been carried out only for HCV; therefore, we focused this evaluation on the HCV genomic data. Each HCV dataset consists of viral haplotypes covering the E1/E2 junction of the HCV genome, which contains the hypervariable region 1 (HVR1). We sampled and analyzed transmission networks of the largest HCV transmission cluster identified previously[146]. It includes 116 persons infected with the HCV subtypes 1a and 3a; some persons were infected with both subtypes. The HCV subtypes are phylogenetically distinct. Given that, we first constructed and analyzed maximum likelihood phylogenies for each subtype separately. In addition, these phylogenies were post-processed using TreeTime[159] to infer time labels

of their internal nodes. The obtained time-scaled phylogenetic trees were used as inputs of SOPHIE and, after obtaining sampled transmission networks and their probabilities, provided times of inferred transmissions. Finally, transmission networks for both subtypes sampled by SOPHIE were joined into a single network. Further data processing details can be found in Star Methods section.

The inferred joint consensus transmission network of both subtypes is shown in Figure 4.5(a). When reconstructed transmission links have both the person's metadata and phylogenetic data, they tend to agree with each other. The subcluster of persons infected with subtype 1a is large, established earlier, and is likely to serve as a source for the 3a subcluster. This finding matches the observation that the inferred primary case of the 3a subcluster (the only vertex with the expected in-degree below 1 in the 3a network) is coinfected with both subtypes. In addition, both persons with known acute infection from the analyzed cluster (detected by the HCV seroconversion test) have low expected outdegrees ($< 10^{-4}$), confirming that they carried secondary rather than primary infections.

The output from SOPHIE was used to estimate key epidemiological parameters directly from the inferred transmission networks. Such estimates can be more realistic than more traditional assessments based on random mixing models applied to incidence statistics[108]. Furthermore, we used time labels of the viral phylogenies to estimate timing of each link in each sampled transmission network to assess the outbreak dynamics.

The dynamics of incident case numbers (i.e. the numbers of inferred transmissions within a specified time interval, in our case, 1 month) suggest that the outbreak started in the middle of

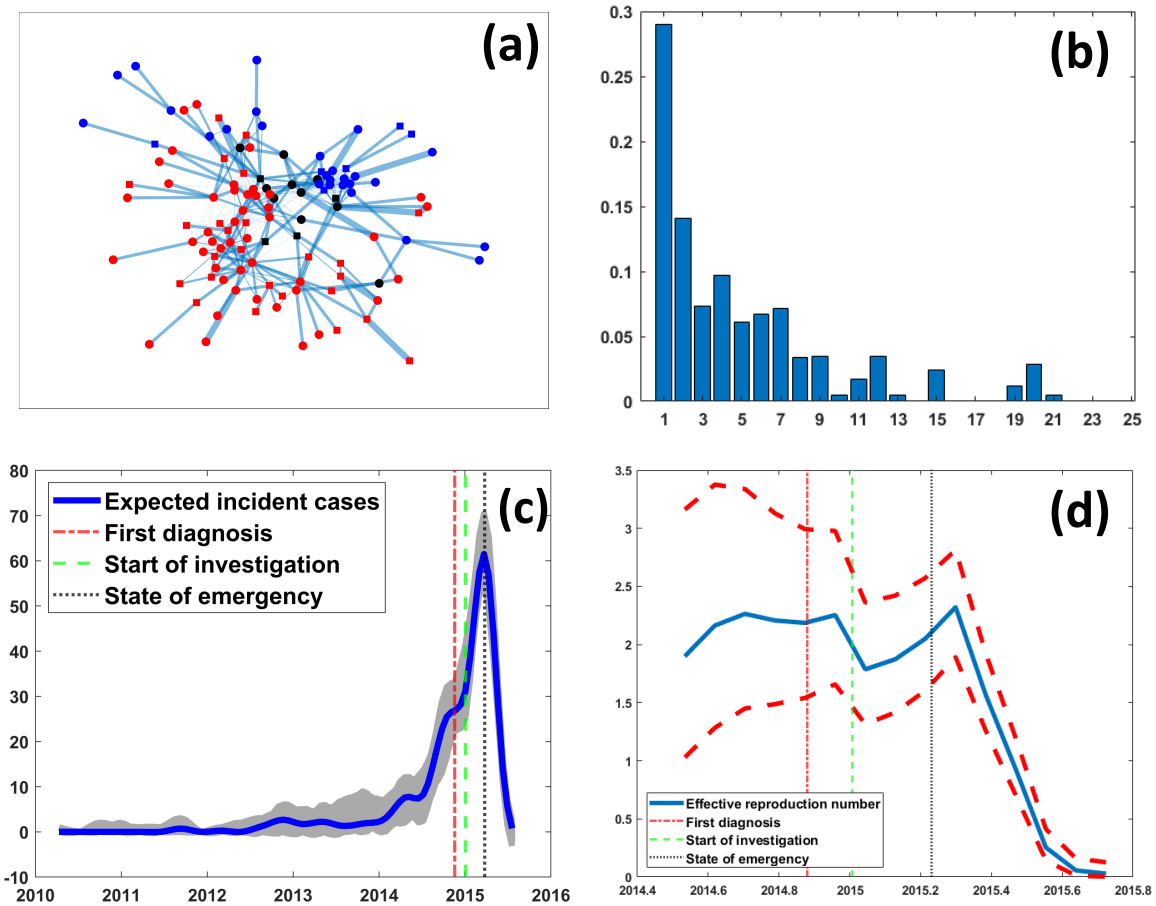

Figure 4.5 Computational analysis of the Indiana HCV outbreak. (a) Consensus transmission network. The thickness of each edge is proportional to its inferred likelihood support. Only edges with the support above 0.0005 are shown. Nodes infected with subtype 1a, 3a and both are shown in red, blue and black, respectively. Squared nodes are co-infected with HIV. (b) Distribution of the generation times by month. (c) The dynamics of incident cases over time. The blue line is the expected number of incident cases at a given time. The grey area shows incident cases for sampled networks. Vertical lines depict major public health events. (d) Effective reproduction numbers R_t for the exponential stage of the outbreak. Vertical lines depict major public health events.

2012, and transitioned to the exponential stage in 2014 (Figure 4.5(c)). The incidence rapidly declined after the declaration of the state public health emergency. The exponential stage largely coincides with the timeline of HIV spread in the same community[23]. In addition, 35 persons from the analyzed cluster were co-infected with HIV, and 25 of them form a connected subgraph of the consensus subnetwork formed by edges with the high support shown in Figure 4.5(a). These findings suggest that the HIV outbreak and the larger part of the HCV outbreak were triggered by the same epidemiological mechanism; however; HCV preceded HIV by several years, and the HIV spread might have been facilitated by the pre-established HCV transmission network.

The inferred incidence (Figure 4.5(c)) and the inferred distribution of generation times (time intervals between the infection times of the sources and recipients, Figure 4.5(b)), were used in EpiEstim[40] to estimate the effective reproduction number R_t (virus transmissibility at a given time) over a 1-month sliding window during the exponential phase of the outbreak. The mean values of R_t varied between 1.81 and 2.33 before the emergency declaration, indicating sustained transmissions. Following the declaration, they rapidly dropped below the epidemic threshold of $R_t = 1$. We also directly measured the basic reproduction number R_0 as an average degree of transmission sources in sampled networks. An estimation $R_0 = 2.71$ (95% CI: $(2.63, 2.79)$) close to the estimates for R_t was obtained. The estimates produced by SOPHIE are more moderate and seemingly more realistic than, for example, the values $R_0 = 6.6$ (95% CI: $(3.2, 9.9)$) and $R_0 = 5.1$ (95% CI: $(1.7, 9.2)$) produced by the birth-death skyline phylodynamics model[172] with the uniform reproduction number prior implemented in BEAST[57]. Moreover, the SOPHIE-based values agree better with the estimate of $R_0 = 3.8$ for the parallel HIV outbreak obtained using contact tracing[23].

4.4 Discussion

Analysis of viral transmission networks is essential for epidemiological and evolutionary studies of pathogens, as it allows to assess and monitor the transmission dynamics[6,18], understand the mechanisms of transmission, infection establishment and emergence of drug resistance and vaccine escape[143,114], as well as to design efficient public health intervention strategies[28]. Hence, inference of viral transmission networks is one of the most fundamental problem of genomic epidemiology and a major driving force behind new developments in the field.

In this paper, we introduce a method for transmission network reconstruction based on the integration of a phylogenetic maximum likelihood (ML) model and a random graph model. The idea to implant social networks into the phylogenetic framework was proposed in our prior study[170] and implemented in a tool QUENTIN. SOPHIE substantially differs from QUENTIN in several ways: (1) it is fully based on maximum likelihood paradigm, (2) it is phylogenetic rather than network-based, and (3) it uses more general and comprehensive random graph model. In general, SOPHIE re-evaluates phylogeny-based candidate transmission networks according to their match to the expected properties of an unobserved contact network and prioritizes the networks, which fit to both the viral phylogeny and these properties.

We showed that the proposed approach is capable of achieving a substantial accuracy improvement over the state-of-the-art phylogenetic transmission inference methods based on maximum parsimony principle, while retaining their scalability and speed. This improvement is likely associated with the relative sampling efficiency of parsimony and likelihood-based methods. Indeed, for most of the simulated test examples the total numbers of optimal parsimonious solutions (as

calculated by TiTUS[163]) were exceedingly large, with the median number of solutions over all tests being $3.9 \cdot 10^{15}$. Representative uniform sampling from such large set is challenging. In contrast, SOPHIE samples from a more informative distribution. Furthermore, the network-based part of the proposed model allows to optimize the search in the solution space by employing a polynomial-time combinatorial optimization machinery. This distinguishes SOPHIE from other phylodynamics models that are often less computationally tractable and have to rely on the MCMC sampling.

Our case study of HCV/HIV outbreak in rural Indiana demonstrates how SOPHIE can be used to analyze viral transmission dynamics and assess the effects of public health interventions. We expect that the our methodology will be useful for other studies of bloodborne and airborne pathogens that spread via human contacts. However, it should be noted that in general the advantages of our approach are more pronounced for bloodborne pathogens, where the role of social contacts is stronger, while for airborne pathogens many contacts could be episodic.

Despite the aforementioned advantages, the proposed methodology has a room for further expansion and improvement. First, its phylogenetic component is currently based on trait substitution models with fixed between-host transmission rates. Incorporation of rate inference via EM or other iterative algorithms can potentially enhance our approach. Such a technique proved to be useful for nucleotide substitution models within traditional phylogenetics and phylodynamics[159]. In our case, however, its application is more challenging due to smaller numbers of ancestral trait changes. Second, ideally the label sampling scheme should simultaneously account for both parts of the joint likelihood. However, use of MCMC or other similar approach for such sampling is

non-scalable, while development of the scheme based on combinatorial optimization seems to be challenging. One possible combinatorial approach envisioned by us is the utilization of spectral techniques. Third, as suggested by computational experiments, SOPHIE is sensitive to potential phylogenetic inference inaccuracy, especially, in respect to branch lengths estimation. This can be addressed by allowing for length updates, similarly to transmission rates. Finally, the experiments also revealed the decreased accuracy of SOPHIE (and other methods), when applied to data produced by the SIR transmission model with the intra-host logistic coalescent. This suggests that in this case the method's accuracy may benefit from replacement of the ML phylogenetic model with the Bayesian coalescent or other appropriate phylodynamic model. Such models are, however, less computationally tractable; therefore their incorporation into our framework will require innovative algorithmic solutions.

4.5 Star Methods

4.5.1 Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Simulated data	This paper	https://doi.org/10.5281/zenodo.6792964 https://github.com/compbel/SOPHIE/
HCV benchmarking data		https://doi.org/10.5281/zenodo.6792964 https://github.com/compbel/SOPHIE/
Case study data	[146]	Available from the CDC upon request
Software and Algorithms		
SOPHIE	This paper	https://doi.org/10.5281/zenodo.6792964 https://github.com/compbel/SOPHIE/
Phyloscanner	[196]	https://github.com/BDI-pathogens/phyloscanner
TNeT	[51]	https://compbio.engr.uconn.edu/software/TNet/
FAVITES	[123]	https://github.com/niemasd/FAVITES
RAxML	[173]	https://cme.h-its.org/exelixis/web/software/raxml/
TreeTime	[159]	https://github.com/neherlab/treetime
EpiEstim	[40]	https://github.com/mrc-ide/EpiEstim
BEAST	[57]	https://beast.community/

4.5.2 Method details

4.5.2.1 Sampling of ancestral label assignments

Suppose that the Markov chain-based substitution model for labels is fixed, i.e. we are given the equilibrium patient probabilities $(\pi_i)_{i=1}^{n_h}$ and the rate matrix $Q = (q_{i,j})_{i,j=1}^{n_h}$, where $q_{i,j}$ is the transmission rate between hosts i and j for $i \neq j$, and $q_{i,i} = -\sum_{j=1}^{n_h} q_{i,j}$. In most cases, transmission rates between specific hosts are unknown. Therefore usually the substitution model will be the fully-symmetric substitution model (similar to Jukes-Cantor model for DNA) with $\pi_i = 1/n_h$ and $q_{i,j} = \mu/(n_h - 1)$, where μ is the general transmission rate. In certain cases, however, between-host transmission rates can be assessed from epidemiological contact tracing or comparison of exposure intervals, if such information is available. In that case, more general substitution model can be employed.

Given the substitution model, we sample from the joint distribution of ancestral label assignments using an extension of the Felsenstein pruning[65] - a standard dynamic programming algorithm for phylogeny likelihood calculation. It is a dynamic programming algorithm that performs a post-order traversal of the phylogeny T and computes, at each node $v \in V(T)$ and for each host $i \in [n_h]$, the conditional likelihood $L(v, i)$ of observing the labels of leafs that are descendants of v, given that $\lambda_v = i$. The computations are based on the following recurrent relation[65]:

$$L(v,i) = \begin{cases} \left(\sum_{j=1}^{n_h} P_{vx}(i,j)L(x,j)\right) \times \left(\sum_{j=1}^{n_h} P_{vy}(i,j)L(y,j)\right), \text{ if } v \text{ is an internal node} \\ \text{with children } x \text{ and } y; \\ 1, \text{ if } v \text{ is a leaf and } \lambda_v = i; \\ 0, \text{ if } v \text{ is a leaf and } \lambda_v \neq i. \end{cases} \tag{4.1}$$

Here $P_{vx} = exp(t_{vx}Q)$ is an $n_h \times n_h$ transition matrix for an edge $vx \in E(T)$, where t_{vx} is the length of vx.

After all conditional likelihoods L are calculated, we perform a pre-order traversal of T and sample a label for each node from the corresponding posterior distribution given its parent's sampled state. The sampling is repeated n_s times. The sampling procedure is formally described by Algorithm 1. For each sampled label assignment $\boldsymbol{\lambda} = (\lambda_v)_{v \in V(T)}$, its *phylogenetic likelihood* $L(T, \boldsymbol{\lambda})$ is calculated as $L(T, \boldsymbol{\lambda}) = \pi_{\lambda_r} \prod_{uv \in E(T)} P_{uv}(\lambda_u, \lambda_v)$, where r is the root of T.

Algorithm 1 Ancestral label sampling

1: Calculate conditional node likelihoods $L(v, i)$ using Felsenstein pruning.
2: **for** $s = 1, ..., n_s$ **do**
3: **for** each internal node v in a pre-order traversal of P **do**
4: **if** v is a root **then**
5: assign v the label $\lambda_v = i$ with the probability $\frac{\pi_i L(v,i)}{\sum_{j=1}^{n_h} \pi_j L(v,j)}$
6: **else**
7: let p be the parent of v
8: assign v the label $\lambda_v = i$ with the probability $\frac{P_{pv}(\lambda_p,i)L(v,i)}{\sum_{j=1}^{n_h} P_{pv}(\lambda_p,j)L(v,j)}$

For large phylogenies, the number of ancestral label assignments with comparable likelihoods can be large. Thus, in order to facilitate sampling of the assignments that potentially produce transmission networks with high network likelihoods, we employ several heuristic adjustments of the general sampling scheme. First, we reduce the tree before sampling by iteratively removing sibling leafs with the same label and assigning that label to their parent. This procedure replaces all monophyletic clades with their most recent common ancestor. This modification decreases the dimensionality of the ancestral label space, thus allowing to obtain a representative sample with fewer iterations. In addition, it speeds up likelihood calculations and decreases the number of likelihood re-scalings[198] required to resolve the numerical precision issues. Next, it is known that

intra-host viral population diversity can serve as a marker of the population age[15], and therefore hosts with more diverse populations are more likely to be sources of transmissions[170,196,156]. We account for that by multiplying the likelihoods $L(v, i)$ calculated for the reduced tree by the number of descendants of v with the label i.

The total running time of the sampling step is $O(n_s n_l n_h^2)$

4.5.2.2 Estimation of the network likelihood

Likelihood definition. We assume that the transmission network $G = G(T, \boldsymbol{\lambda})$ is a subgraph (not necessarily induced) of a random contact network $\mathcal{G}_c$ on $n_c \geq n_h$ vertices. We model $\mathcal{G}_c$ as a random graph with the given degree distribution $\boldsymbol{p} = (p_1, p_2, ...)$, where p_k is the probability that a randomly selected vertex has a degree k.

Every vertex $i \in V(G)$ has a degree d_i in G and a degree $D_i \geq d_i$ in $\mathcal{G}_c$. Let us call a mapping $\mathcal{D} : V(G) \to [n_c - 1]$, that assigns a degree $\mathcal{D}(i) = D_i$ to a vertex i, an *embedding of* G *into* $\mathcal{G}_c$. Then we approximate the network likelihood $L(G|\mathcal{G}_c)$ via the probability of the best embedding:

$$L(G|\mathcal{G}_c) = \max_{\mathcal{D}} p(G, D|\mathcal{G}_c) \tag{4.2}$$

To define the conditional probability $p(G, \mathcal{D}|\mathcal{G}_c)$, we can factorize it as

$$p(G, \mathcal{D}|\mathcal{G}_c) \propto p(G|\mathcal{D})p(\mathcal{D}|\mathcal{G}_c). \tag{4.3}$$

The first factor $p(G|\mathcal{D})$ is the probability of the subgraph G given the degrees of its vertices in the contact network $\mathcal{G}_c$. It can be calculated by assuming that n_c is large enough and $\mathcal{G}_c$ follows

the *Generalized Random Graph (GRG) model*[33,34] – a general and widely used model of a random graph with given expected degrees. According to this model, edges are independently assigned to pairs of vertices (i, j) with probabilities $p_{ij} = \frac{D_i D_j}{2m_c}$, where $m_c = \frac{n_c}{2} \sum_{k=1}^{n_c-1} k p_k$ is the expected number of edges of $\mathcal{G}_C$. Using this definition, we get

$$p(G|\mathcal{D}) = \prod_{ij \in E(G)} \frac{D_i D_j}{2m_c} = \frac{1}{(2m_c)^{m_h}} \prod_{i=1}^{n_h} D_i^{d_i}, \tag{4.4}$$

where m_h is the number of edges of G.

To define the second factor $p(\mathcal{D}|\mathcal{G}_c)$, consider the vector of expected degree counts $C = (C_1, ..., C_{n_c-1})$ of $\mathcal{G}_c$, i.e. $C_j = \lceil p_j n_c \rceil$ is a rounded expected number of vertices of degree j. Then $p(\mathcal{D}|\mathcal{G}_c)$ is the probability that the degrees $(D_1, ..., D_{n_h})$ are sampled without replacement from the population C. Thus, $p(\mathcal{D}|\mathcal{G}_c)$ is described by the probability mass function of the multivariate hypergeometric distribution:

$$p(\mathcal{D}|\mathcal{G}_c) = \frac{1}{\binom{n_c}{n_h}} \prod_{k=1}^{n_c-1} \binom{C_k}{\sigma_k}, \tag{4.5}$$

where $\sigma_k = |\mathcal{D}^{-1}(k)| = |\{i : D_i = k\}|$.

Likelihood calculation. To calculate the network likelihood, we need to solve the optimization problem (4.2). After logarithmic transformation, it is equivalent to the following problem:

$$\max_{\mathcal{D}} \left(\sum_{i=1}^{n_h} d_i \log(D_i) + \sum_{k=1}^{n_c-1} \log \binom{C_k}{\sigma_k} \right). \tag{4.6}$$

In turn, this problem can be reduced to a *generalized uncapacitated facility location problem with*

convex costs[56], where the vertices of G serve as clients and their possible expected degrees in $\mathcal{G}_c$ – as facilities. More specifically, we consider the set of clients $K = [n_h]$ and the set of facilities $F = [n_c - 1]$; if the client i is served by the facility k (i.e. $D_i = k$), where $k \geq d_i$, then the profit $b_{ik} = d_i \log(k)$ is generated. Furthermore, the assignment of σ_k clients to a facility k produces a profit $f_k(\sigma_k) = \log \binom{C_k}{\sigma_k}$. The objective is to assign all clients to facilities in such a way that the total profit is maximized.

The crucial property of the obtained problem is the fact that the functions $f_k(\sigma)$ are concave (or, if we are using more standard minimization formulation, $-f_k(\sigma)$ are convex). Thus, we can use the scheme proposed in[79] to reduce our problem to the maximum-weight matching problem for bipartite graphs, which is solvable in polynomial time[165]. Namely, we construct a bipartite graph H with the parts (X, Y), where the part X coincides with the set of clients K, and the part Y contains C_k vertices $y_k^0, ..., y_k^{C_{k-1}}$ for each facility k. The vertices $i \in X$ and $y_k^j \in Y$ are adjacent whenever $d_i \leq k$, and the weight of this edge is set to $w_{iy_k^j} = b_{ik} + f_k(j+1) - f_k(j)$. Then maximum-weight matching of H gives us the solution of (4.6). This fact follows from the concavity of the function f_k, which implies that any maximum-weight matching that covers the vertex $y_k^j \in Y$ should also cover all vertices y_k^l for $l \leq j$.

It is easy to see that the number of edges in the bipartite graph H is $n_h(n_c + 1) - 2m_h$. Therefore, the described reduction approach combined with the generalized Hungarian algorithm for the matching problem[149] calculates the network likelihood in time $O(n_h^2 n_c - 2m_h)$.

Finally, it should be noted that the model (4.3) contains the size of the contact network n_c as a parameter. In our calculations, we used the value that is large enough to guarantee the existence of

a feasible solution of (4.6), i.e. $n_c = \max_i \lceil c_i/p_i \rceil$, where $c_i = |\{j : d_j = i\}|$ are degree counts of G. In particular, if the expected degree distribution of $\mathcal{G}_c$ follows the power law with the exponent α, then n_c can be estimated as $n_c = \max_i \lceil \zeta(\alpha) c_i i^\alpha \rceil$, where $\zeta(\alpha)$ is the Riemann zeta function.

4.5.2.3 Distribution and consensus of sampled networks

The output of the algorithms described above is the set of N sampled solutions, where each solution consists of the label assignment $\boldsymbol{\lambda}^i$, the corresponding transmission network $G(T, \boldsymbol{\lambda}^i)$ and the joint likelihood $L(T, \boldsymbol{\lambda}^i) L(G(T, \boldsymbol{\lambda}^i) | \mathcal{G}_c)$. The distributions of transmission networks and labels, as well as derivative epidemiological parameters, can be further analyzed directly – an example of such analysis for a particular case study is presented in Subsection 4.3. In particular, sampled networks can be summarized into the weighted *consensus network* with the adjacency matrix $\mathcal{W} = (w_{ij})_{i,j=1}^N = \sum_{i=1}^N p_i A_i$, where A_i is the adjacency matrix of the network $G(T, \boldsymbol{\lambda}^i)$, and $p_i = \frac{L(T,\boldsymbol{\lambda}^i) L(G(T,\boldsymbol{\lambda}^i)|\mathcal{G}_c)}{\sum_{j=1}^N L(T,\boldsymbol{\lambda}^j) L(G(T,\boldsymbol{\lambda}^j)|\mathcal{G}_c)}$ is the probability density value estimate for that network. In this case, w_{ij} is an inferred likelihood support for an edge ij, and $d^+(i) = \sum_{j=1}^n w_{ji}$ and $d^-(i) = \sum_{j=1}^n w_{ji}$ are expected in- and out-degrees of a vertex i, respectively. When a specific output network is needed (e.g. for benchmarking, see Subsections 4.2.0.1-4.2.0.2), then we calculate it as the maximum-weight arborescence of this weighted network.

4.5.3 Quantification and statistical analysis

4.5.3.1 Simulation and algorithm comparison details.

Synthetic data used in this study was generated by FAVITES[123]. Viral genomes of length 2640bp (that roughly corresponding to lengths of HIV gap and pol polyproteins) were assumed to evolve

under the GTR+Γ substitution model. The GTR rate matrix and gamma parameter were borrowed from[171], where they were estimated based on real HCV data. Inside each host, viral phylogenies evolved under a coalescent model with exponential or logistic effective population growth. We assumed that the virus spread over a contact network of 100 susceptible individuals, that was produced using the Barabasi-Albert model[13]. Two epidemiological scenarios were used: Susceptible-Infected (SI) transmission model and simultaneous sampling of all infected individuals and Susceptible-Infected-Recovered (SIR) transmission model, with each individual sampling time being chosen from a truncated normal distribution of the individual's infection time window. The full lists of FAVITES parameters are available in configuration files provided with simulated datasets in SOPHIE repository.

For each of four combinations of evolutionary and epidemiological models, 100 simulated datasets have been generated, with 10 genomes sampled per infected host. Simulations that produced no transmission links were discarded. For each dataset, we considered a true phylogeny provided by FAVITES and a phylogeny reconstructed by RAxML[173]. The latter was run with the GTR+Γ substitution model, and with optimization of substitution rates and site - specific evolutionary rates.

TNet was run with the default settings. For Phyloscanner, we set the within-host penalty parameter to 0 (otherwise, it produced no transmission links). For SOPHIE, at the label sampling stage we used the uniform equilibrium probability distribution and fixed transmission rates $\mu = 0.0001$ and $\mu = 0.005$ for all Favites and RAxML trees, respectively. For each test instance, $100,000$ internal label assignments were sampled. The f-score has been used as an evaluation metric.

To compare the distributions of f-scores for different algorithms, we utilized a non-parametric Kruskal–Wallis test.

4.5.3.2 Analysis of HCV outbreak in rural Indiana

Analyzed HCV datasets consist of viral haplotypes sampled from infected individuals and sequenced using GS FLX Titanium Sequencing Kit (454 Life Sciences, Roche, Branford, CT). The haplotypes cover the E1/E2 junction of the HCV genome (264 bp), which contains the hypervariable region 1 (HVR1). For our analysis, we used haplotypes that were sampled at least 5 times in each infected person. In total, 4167 viral haplotypes (or ≈ 36 haplotypes per person) have been considered. Prior to phylogenetic analysis, the sequences have been aligned using MAFFT[93]. Next, maximum likelihood phylogenies were constructed for each subtype; in addition, these phylogenies were time-labeled using TreeTime[159] run with default parameters. The obtained time-scaled phylogenetic trees were processed by SOPHIE, for which we used the uniform equilibrium label distribution, the rate $\mu = 1$, and the power-law exponent $\alpha = 2$. For each phylogeny, $2,000,000$ label assignments were sampled.

A Supplementary figures

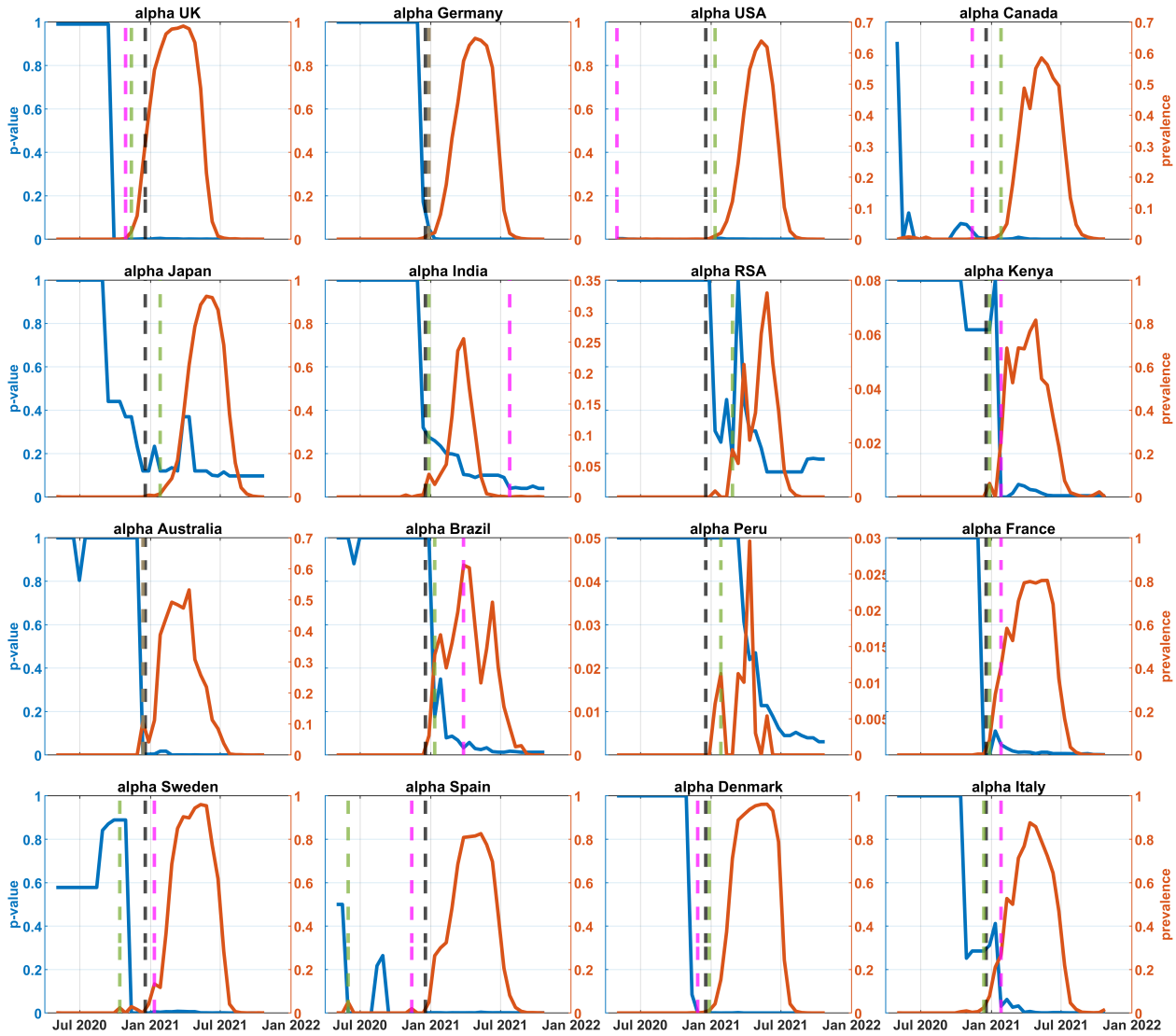

Figure A.1 p-values (blue) and prevalences (red) of Alpha variant in the analyzed countries. Black, green, and magenta lines represent the times of VOC designation, achieving 1% prevalence, and becoming significantly dense, respectively.

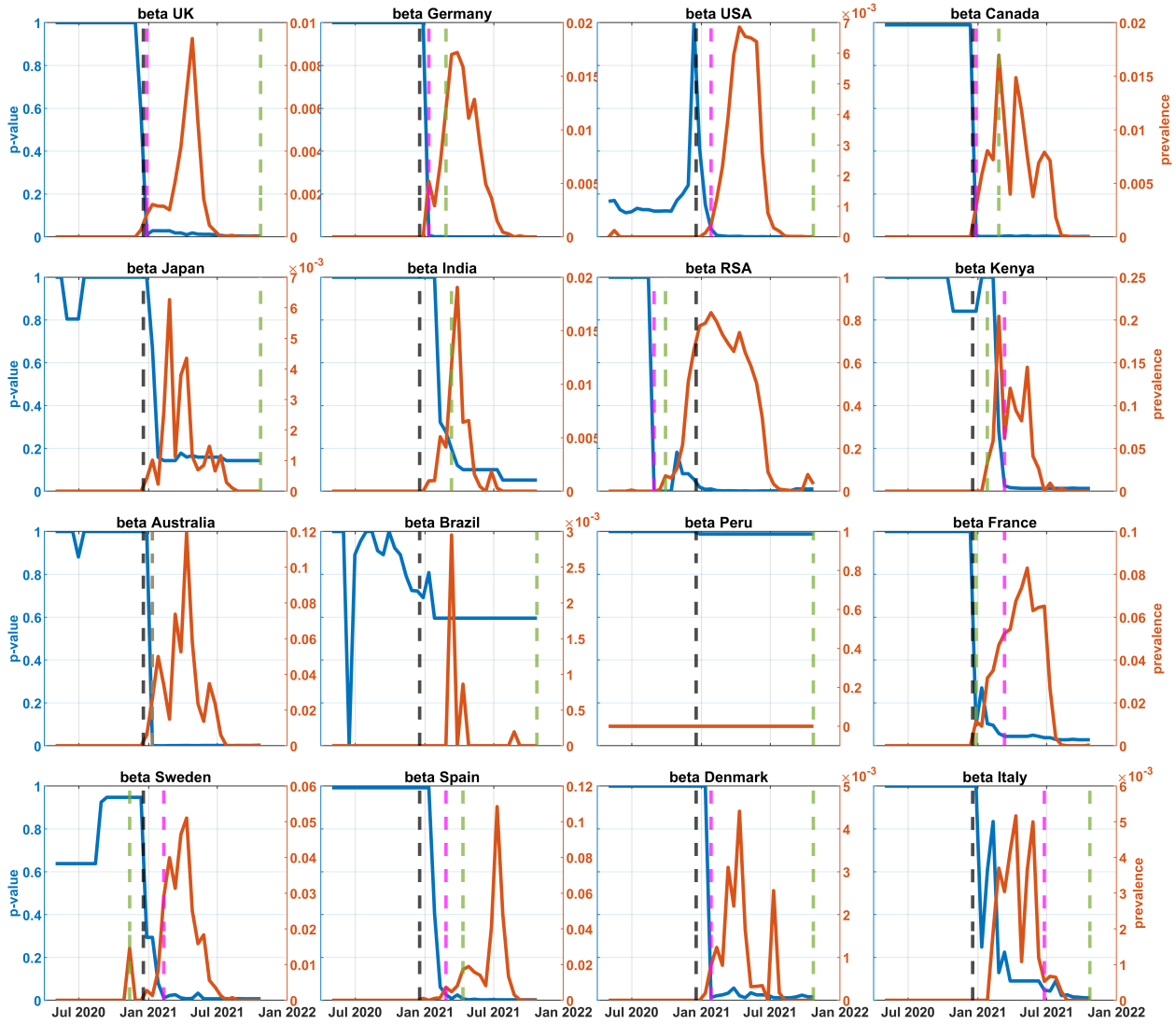

Figure A.2 p-values (blue) and prevalences (red) of Beta variant in the analyzed countries. Black, green, and magenta lines represent the times of VOC designation, achieving 1% prevalence, and becoming significantly dense, respectively.

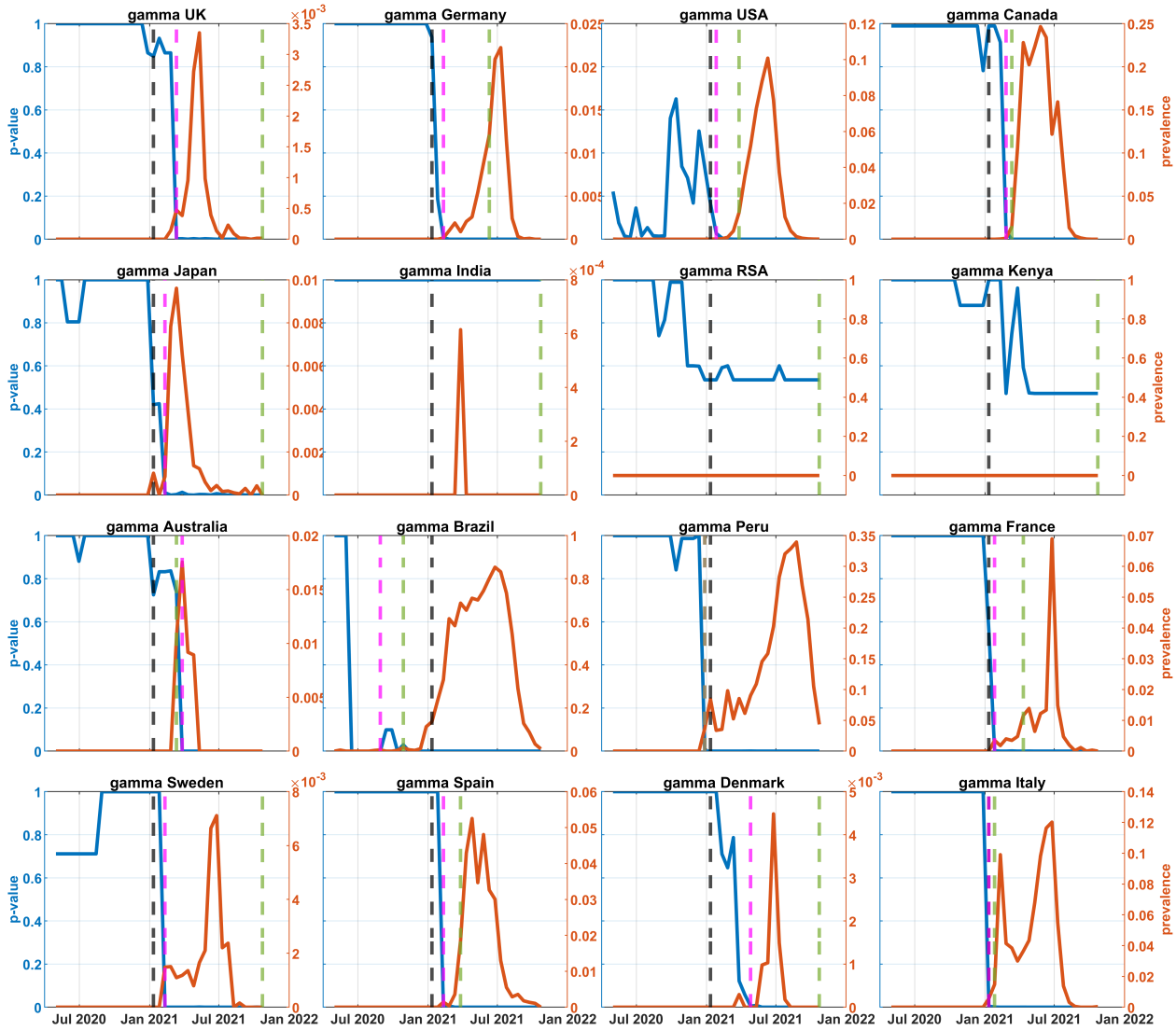

Figure A.3 p-values (blue) and prevalences (red) of Gamma variant in the analyzed countries. Black, green, and magenta lines represent the times of VOC designation, achieving 1% prevalence, and becoming significantly dense, respectively.

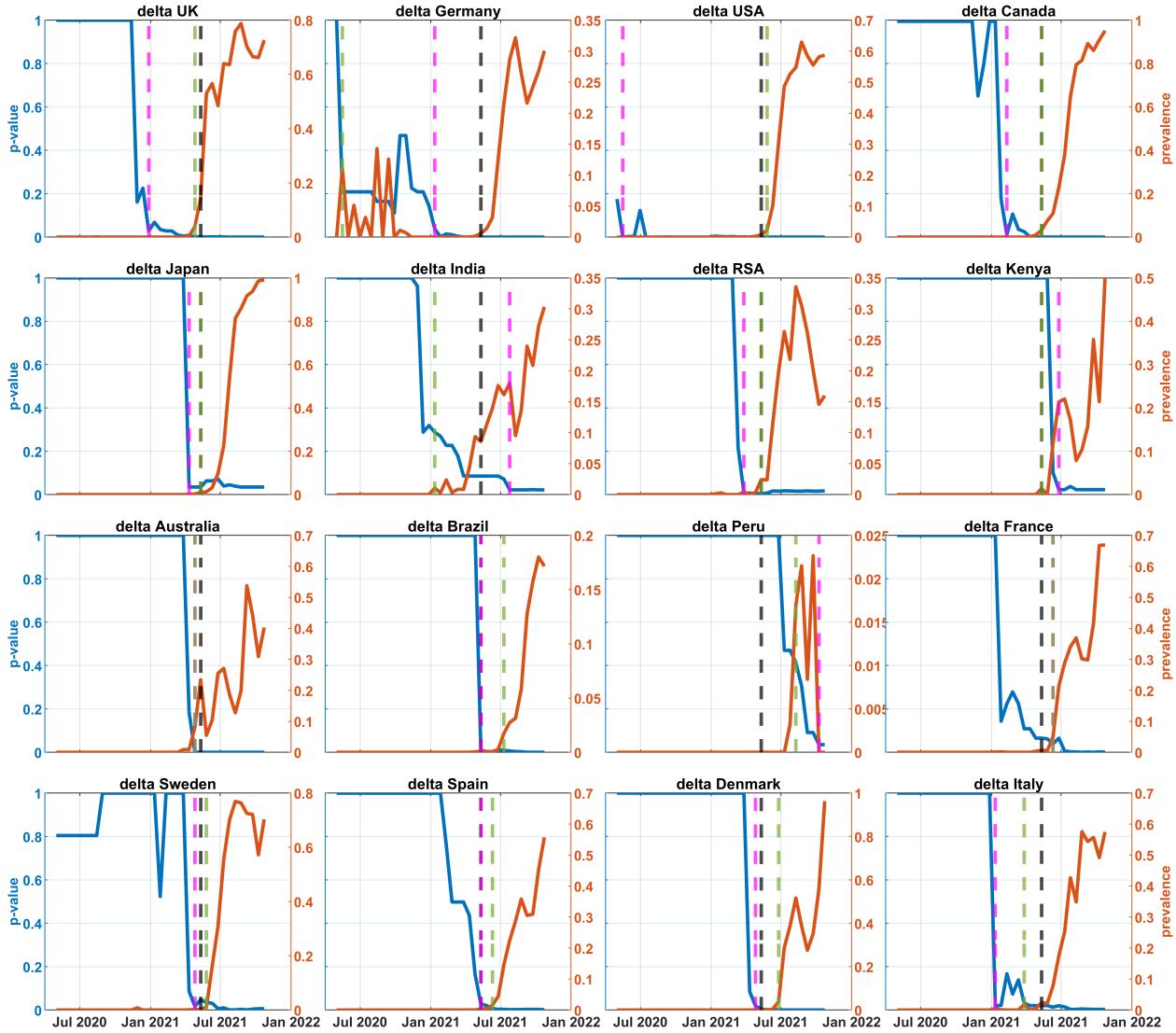

Figure A.4 p-values (blue) and prevalences (red) of Delta variant in the analyzed countries. Black, green, and magenta lines represent the times of VOC designation, achieving 1% prevalence, and becoming significantly dense, respectively.

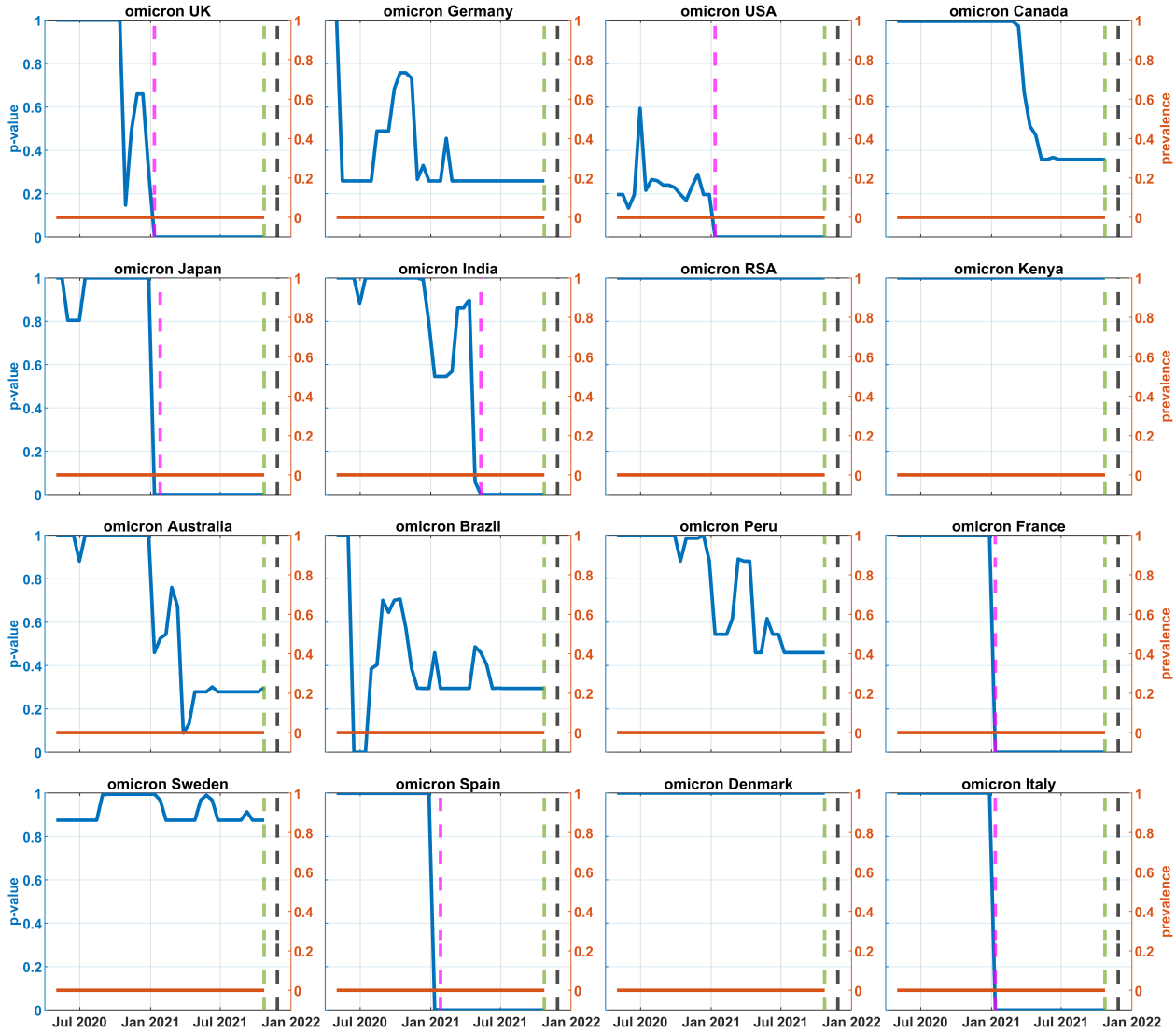

Figure A.5 p-values (blue) and prevalences (red) of Omicron variant in the analyzed countries. Black, green, and magenta lines represent the times of VOC designation, achieving 1% prevalence, and becoming significantly dense, respectively.

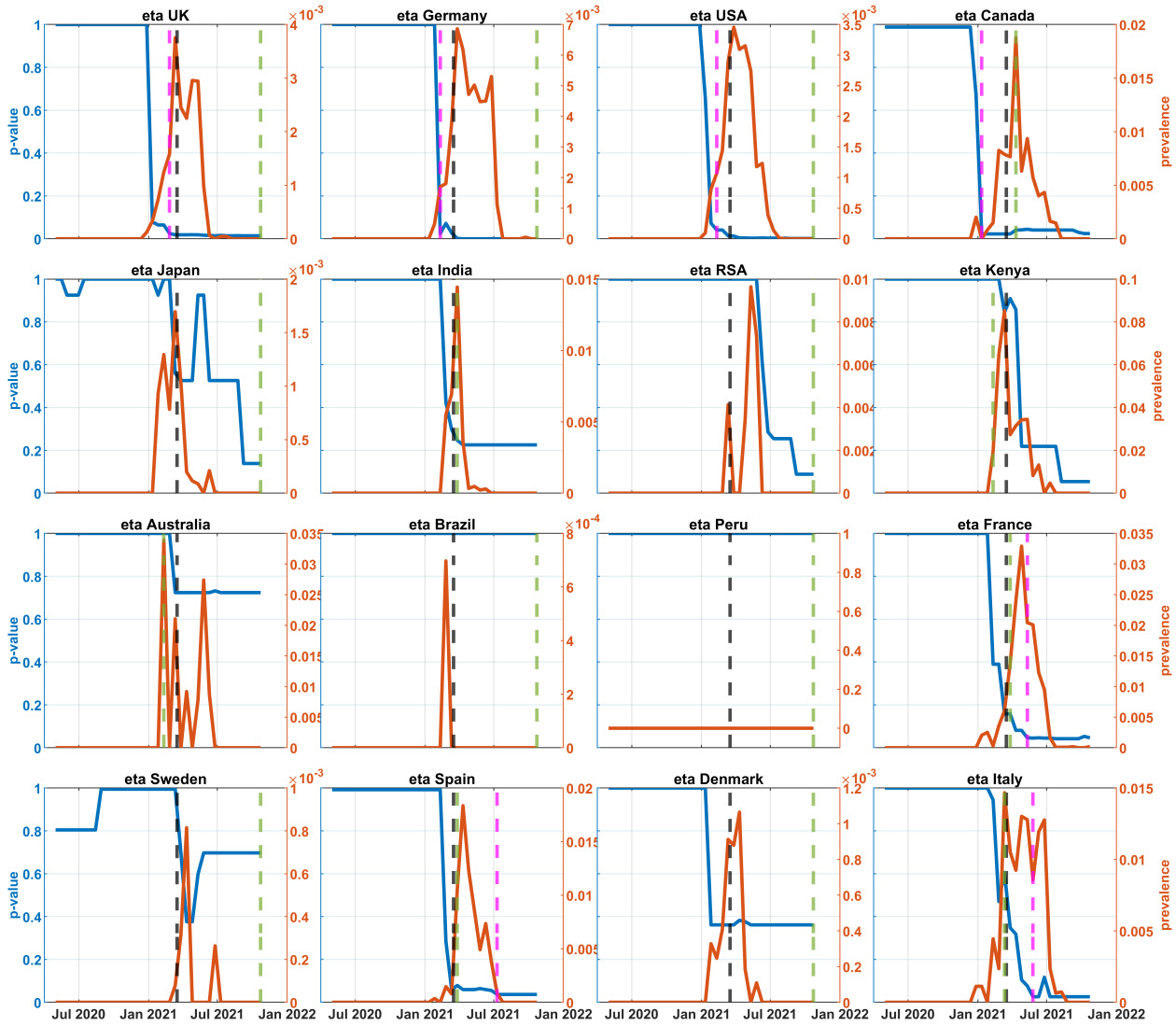

Figure A.6 p-values (blue) and prevalences (red) of Eta variant in the analyzed countries. Black, green, and magenta lines represent the times of VOC designation, achieving 1% prevalence, and becoming significantly dense, respectively.

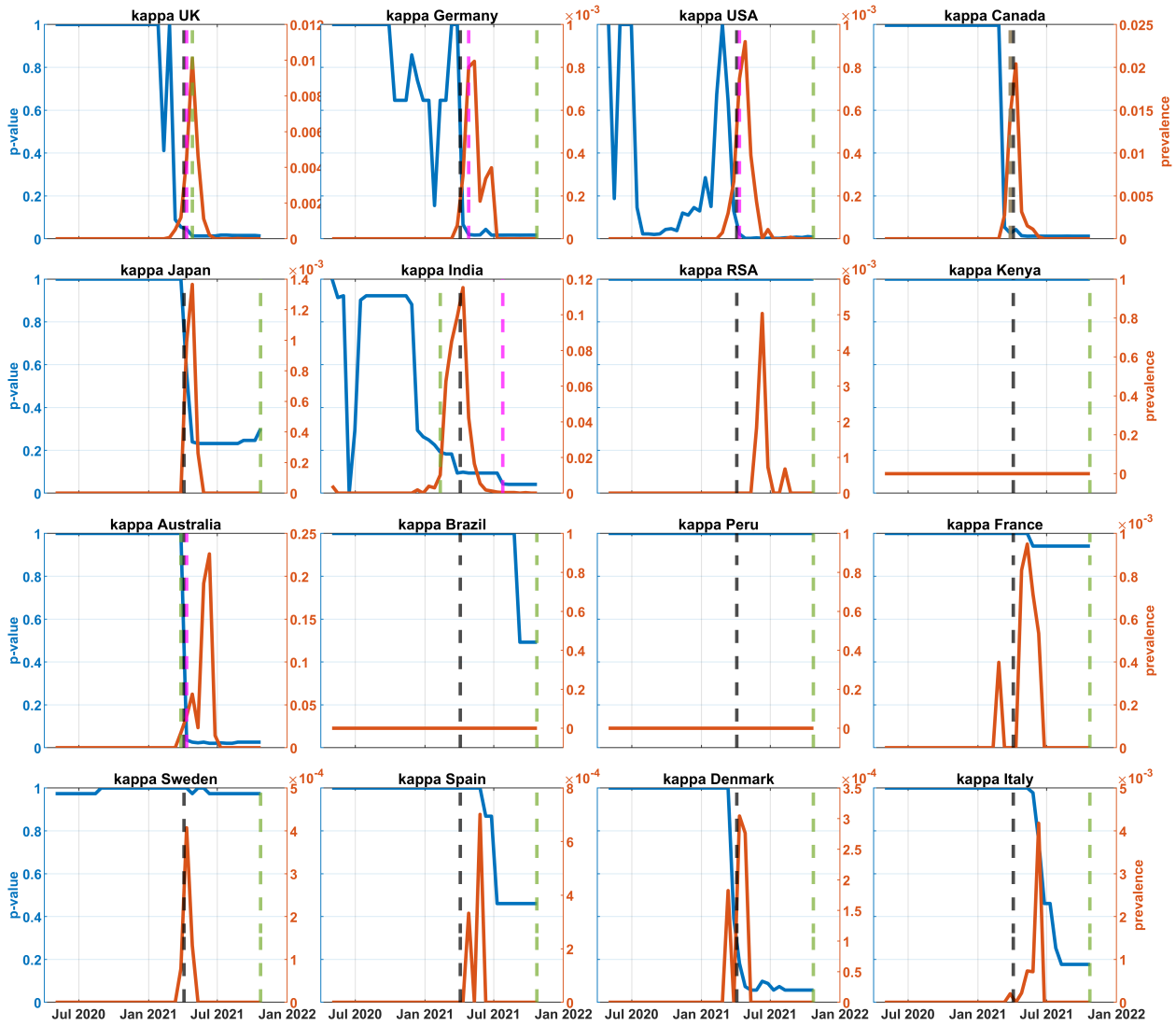

Figure A.7 p-values (blue) and prevalences (red) of Kappa variant in the analyzed countries. Black, green, and magenta lines represent the times of VOC designation, achieving 1% prevalence, and becoming significantly dense, respectively.

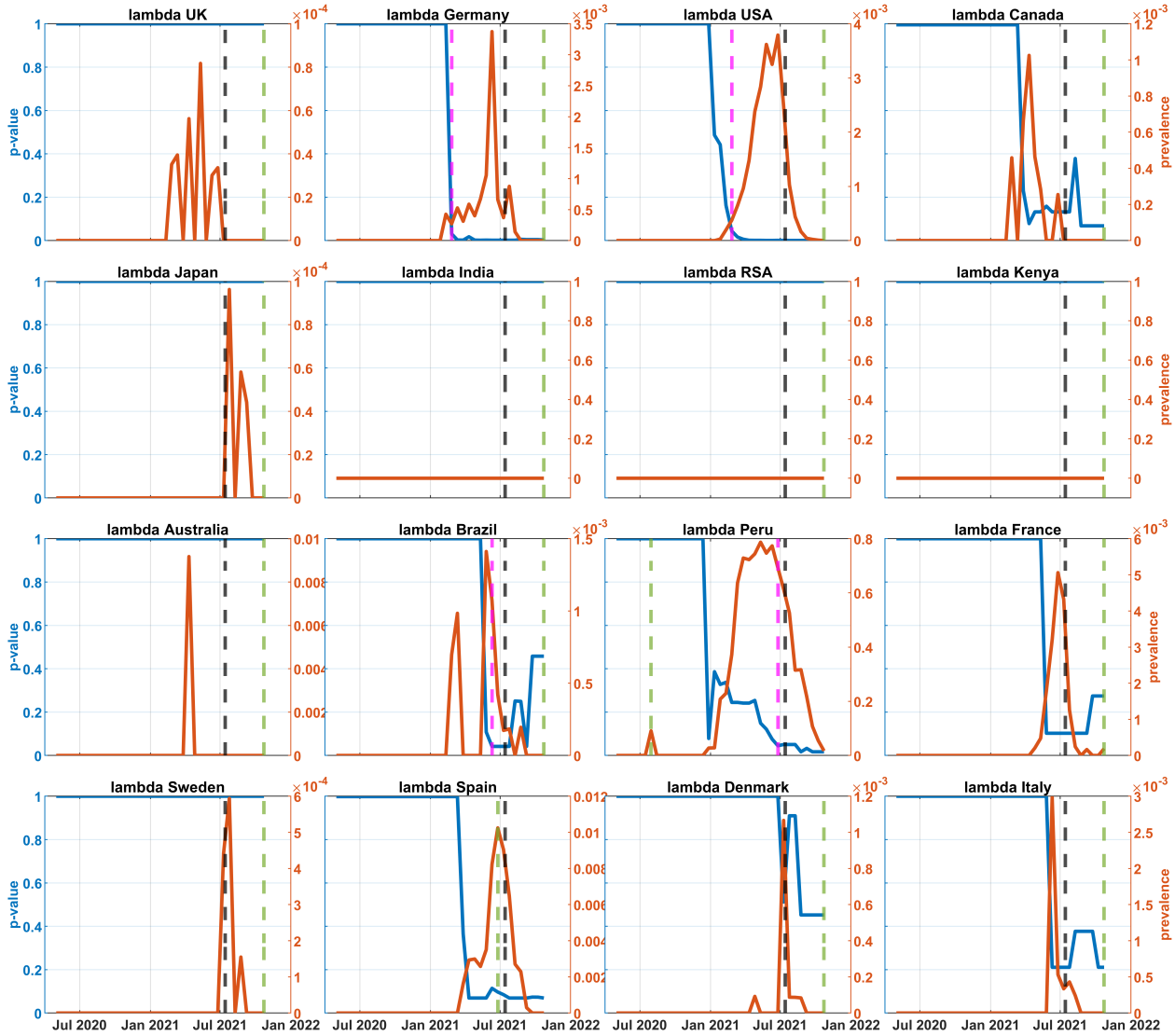

Figure A.8 p-values (blue) and prevalences (red) of Lambda variant in the analyzed countries. Black, green, and magenta lines represent the times of VOC designation, achieving 1% prevalence, and becoming significantly dense, respectively.

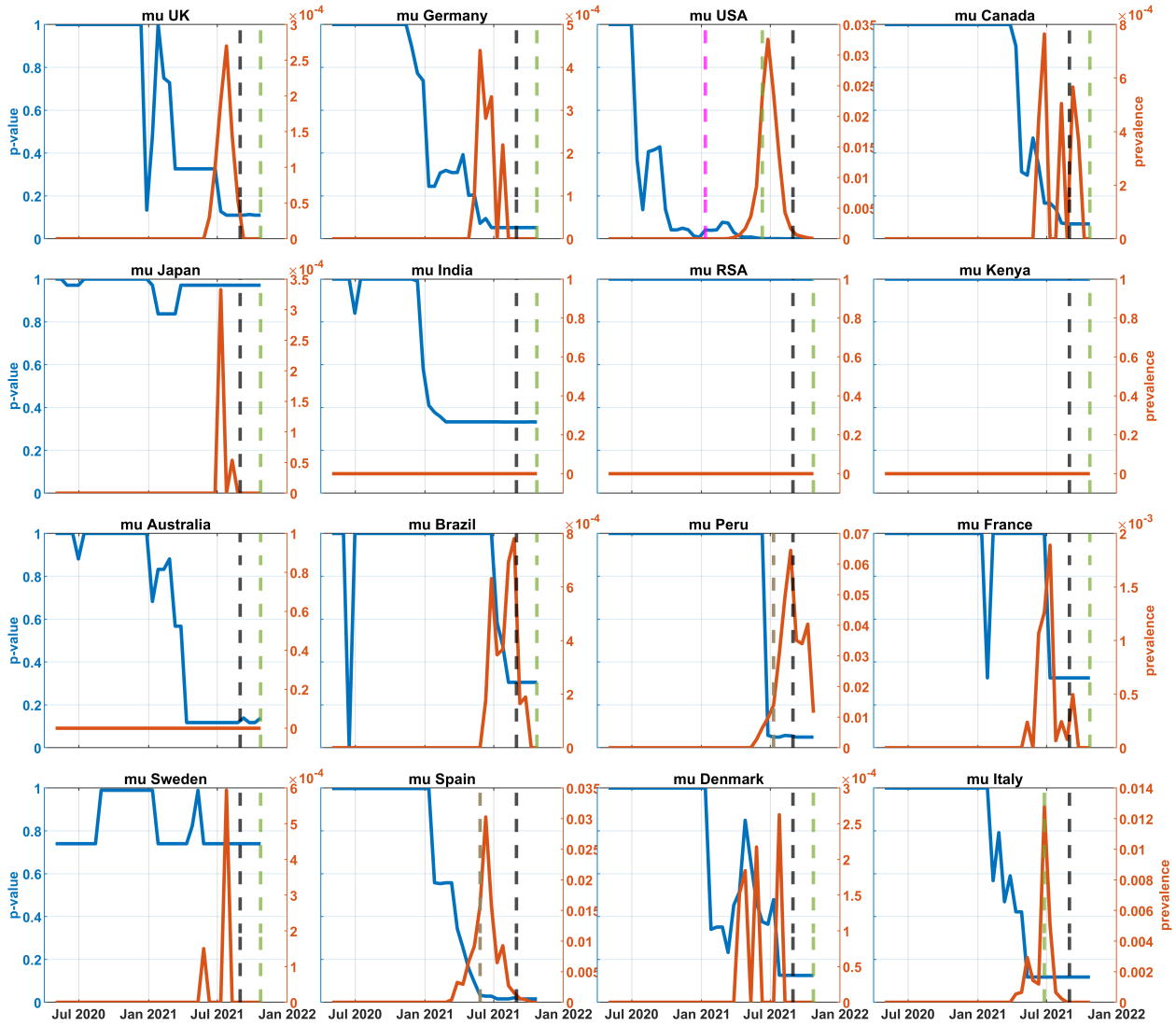

Figure A.9 p-values (blue) and prevalences (red) of Mu variant in the analyzed countries. Black, green, and magenta lines represent the times of VOC designation, achieving 1% prevalence, and becoming significantly dense, respectively.

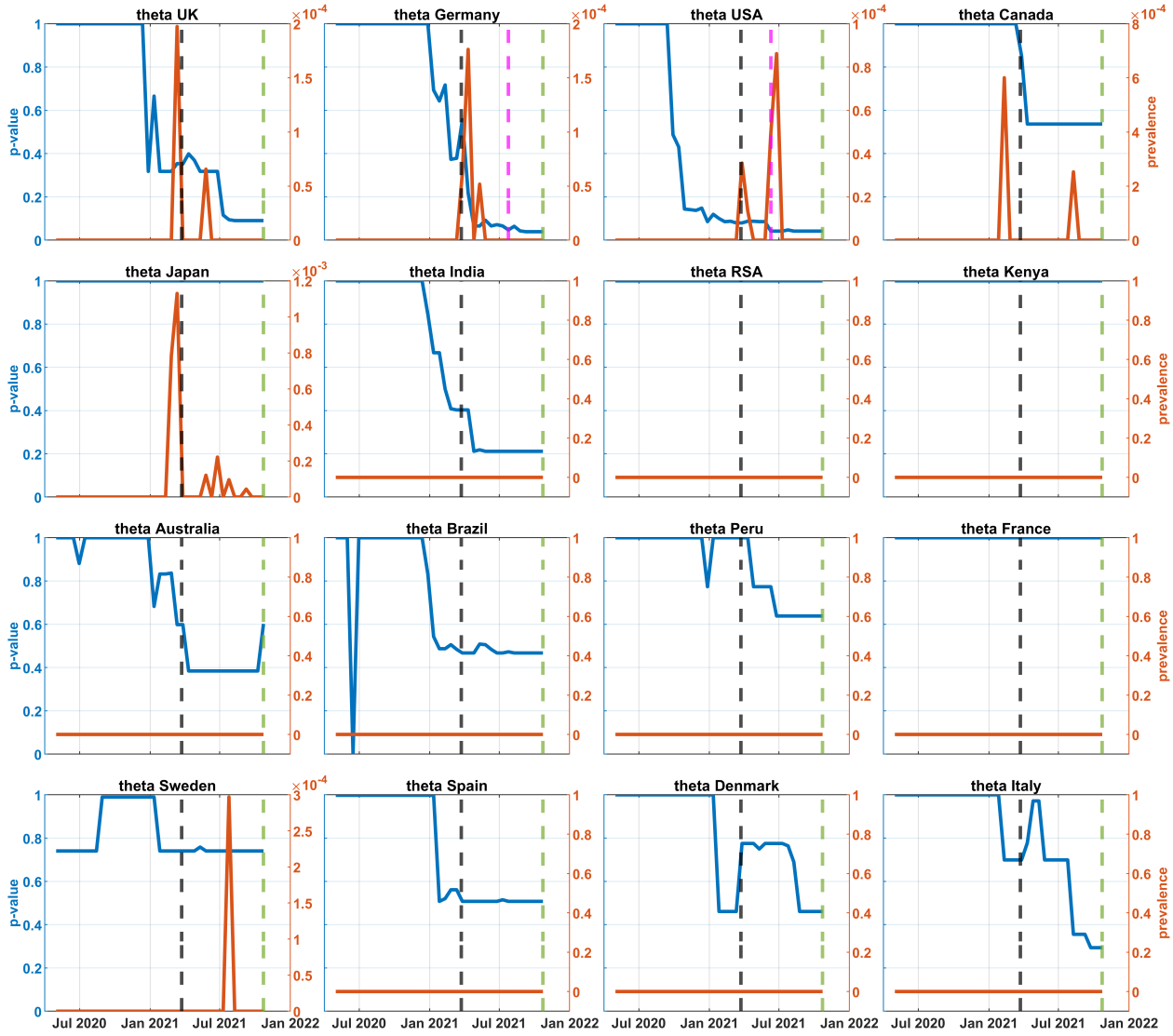

Figure A.10 p-values (blue) and prevalences (red) of Theta variant in the analyzed countries. Black, green, and magenta lines represent the times of VOC designation, achieving 1% prevalence, and becoming significantly dense, respectively.

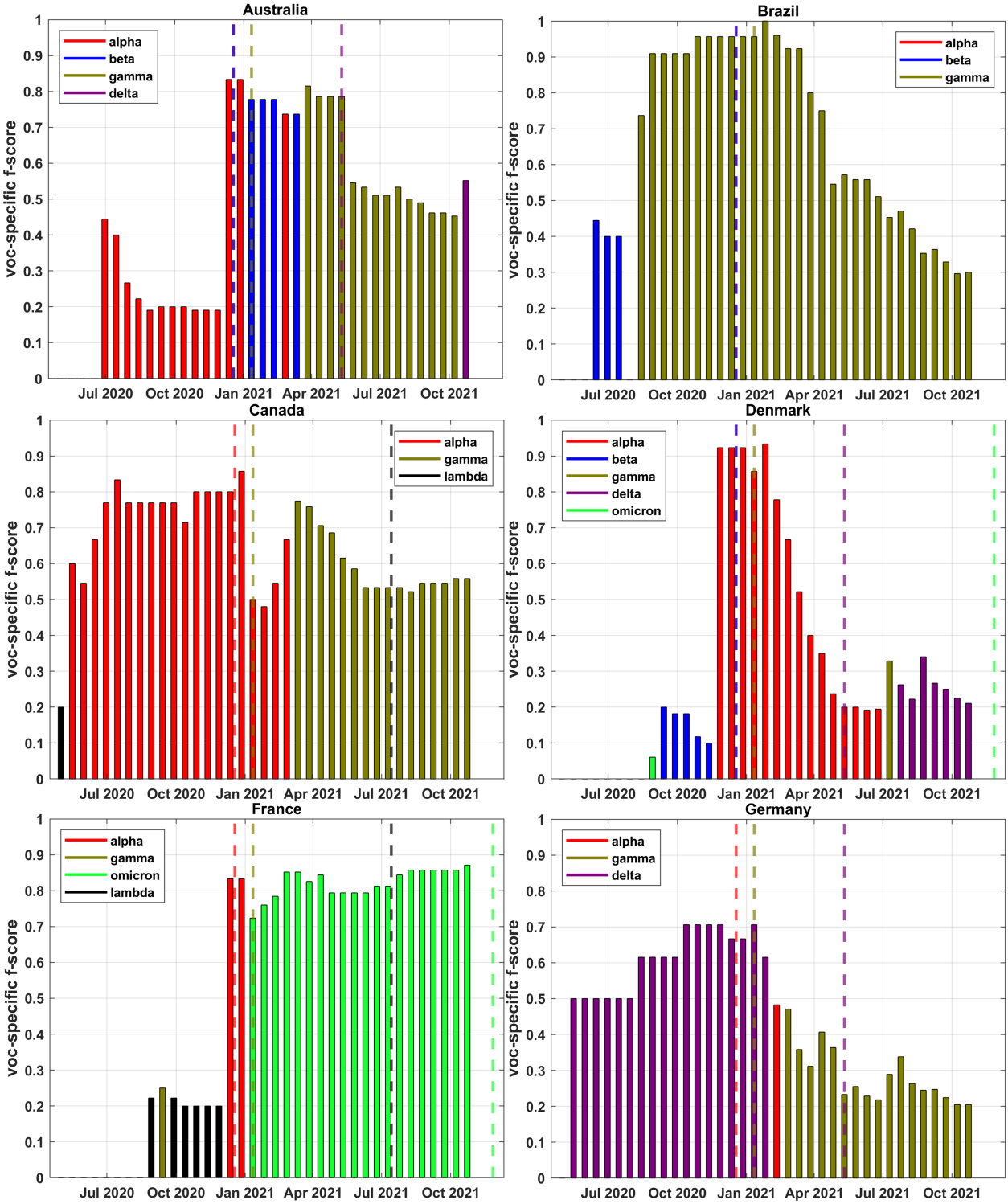

Figure A.11 Comparison between VOCs and densest subnetworks of temporal epistatic networks for selected countries (part 1). At each time point, bar color code corresponds to the VOC closest to the inferred densest subnetwork, and the bar hight is equal to the respective f-score. The number at the top of each bar is the frequency of the corresponding VOC among sequences sampled at the current time interval, measured in percent and rounded to closest integer value. Colored dashed lines mark times when specific VOCs were designated by WHO.

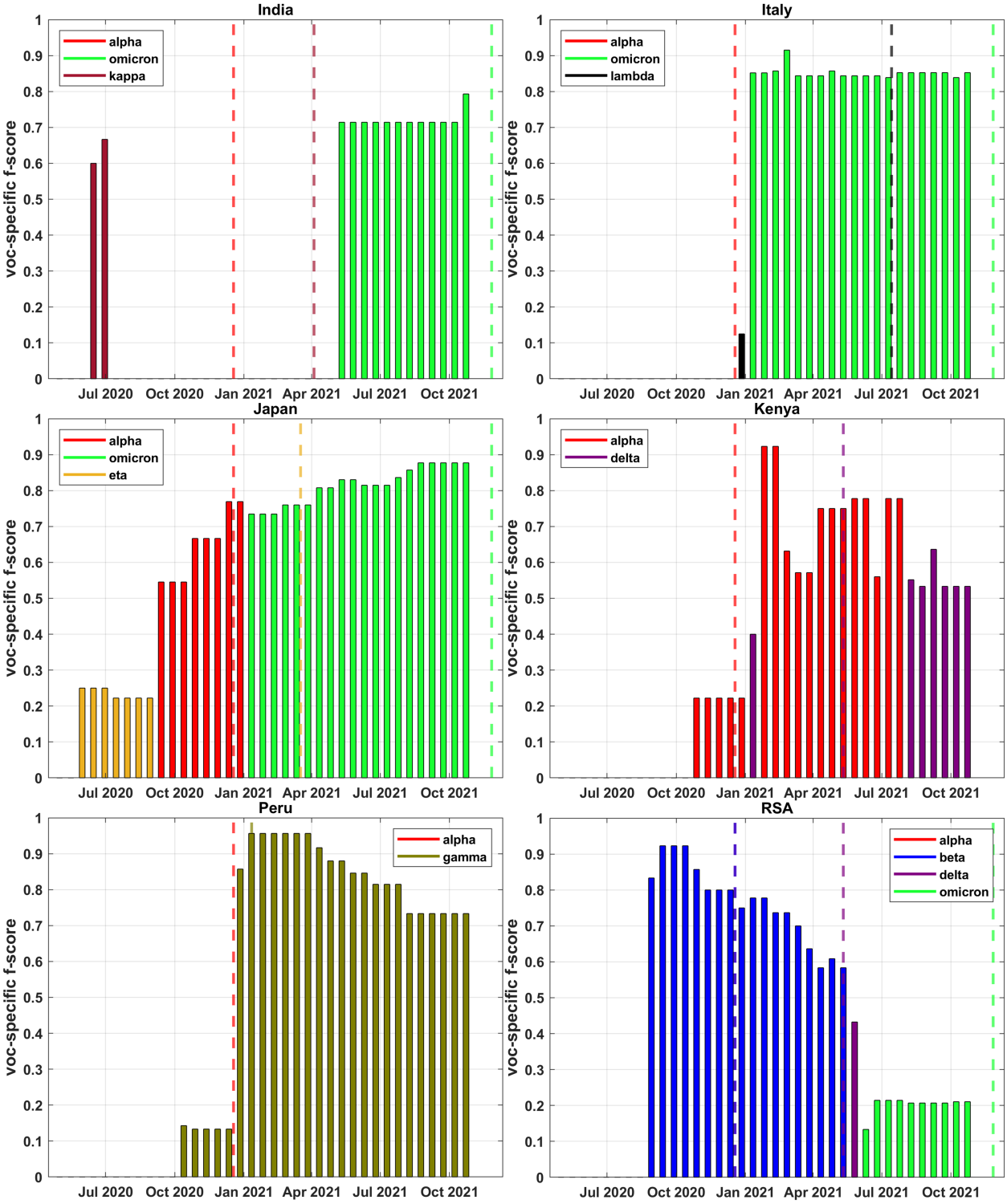

Figure A.12 Comparison between VOCs and densest subnetworks of temporal epistatic networks for selected countries (part 2). At each time point, bar color code corresponds to the VOC closest to the inferred densest subnetwork, and the bar hight is equal to the respective f-score. The number at the top of each bar is the frequency of the corresponding VOC among sequences sampled at the current time interval, measured in percent and rounded to closest integer value. Colored dashed lines mark times when specific VOCs were designated by WHO.

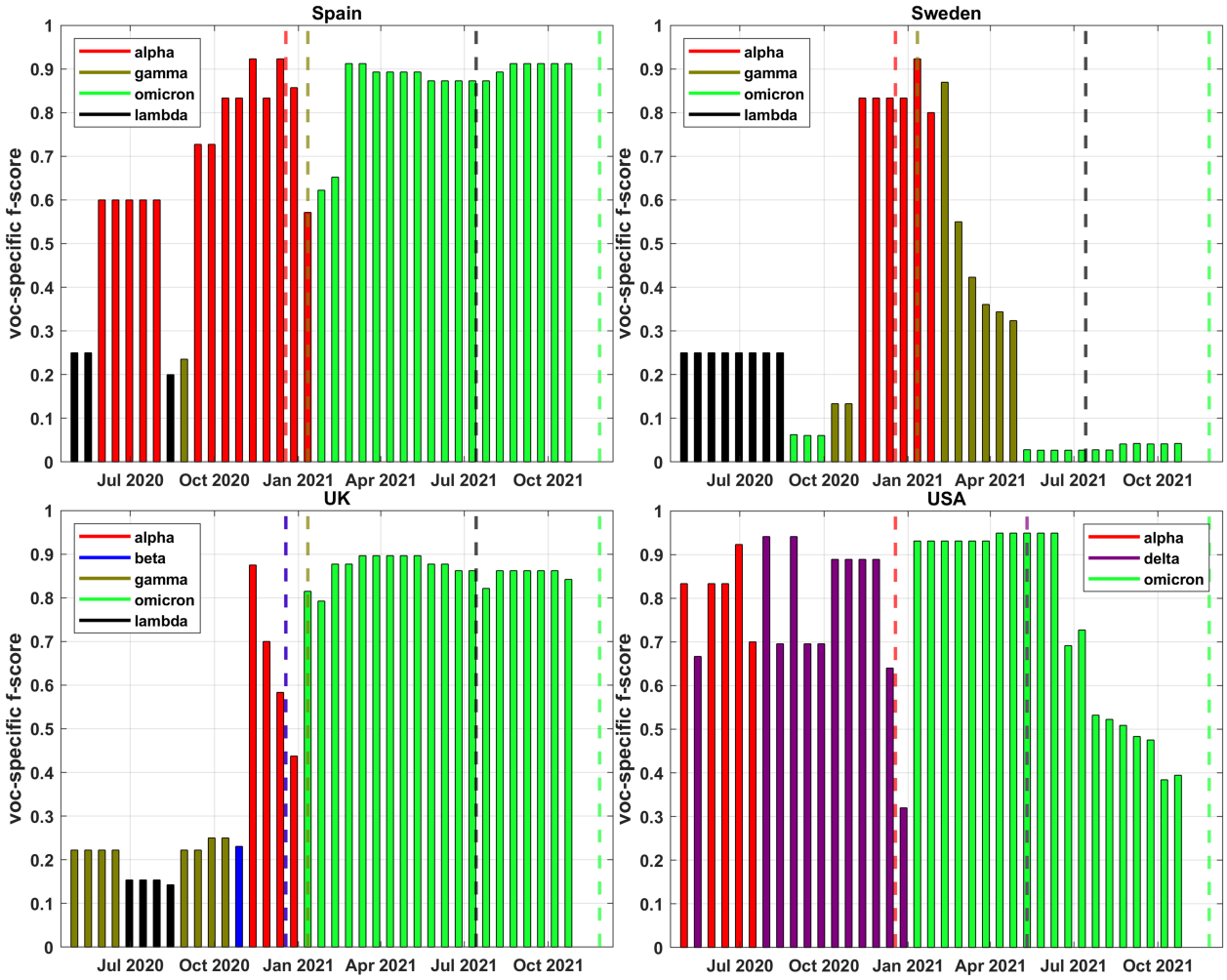

Figure A.13 Comparison between VOCs and densest subnetworks of temporal epistatic networks for selected countries (part 3). At each time point, bar color code corresponds to the VOC closest to the inferred densest subnetwork, and the bar hight is equal to the respective f-score. The number at the top of each bar is the frequency of the corresponding VOC among sequences sampled at the current time interval, measured in percent and rounded to closest integer value. Colored dashed lines mark times when specific VOCs were designated by WHO.

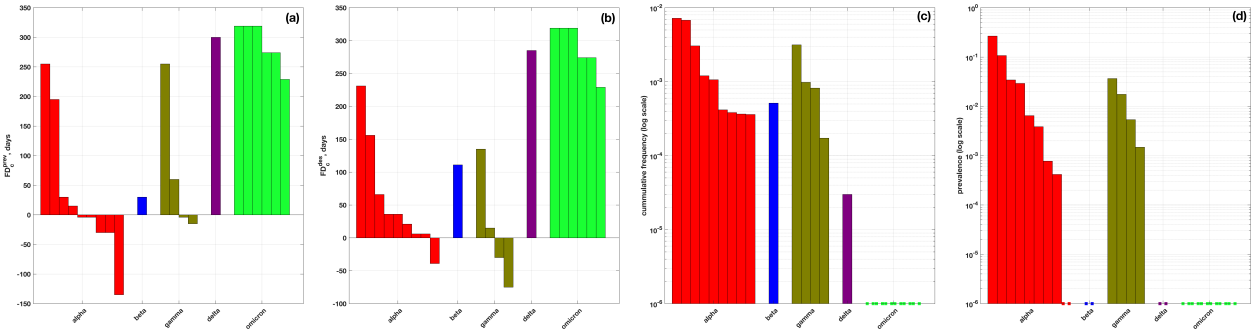

Figure A.14 Summary of comparison between VOCs and densest subnetworks of temporal epistatic networks for all countries. (a) and (b): forecasting depths (y-axis) with respect to the 1% prevalence time and WHO designation time for each analyzed VOCs over different countries. (c) and (d): cumulative frequencies and prevalences of VOCs over different countries at earliest times when they are at least 80% identical to densest subgraphs of epistatic networks (in logarithmic scale). Dashed lines at the bottom of the plot signify that the variants were found at frequencies/prevalences 0.

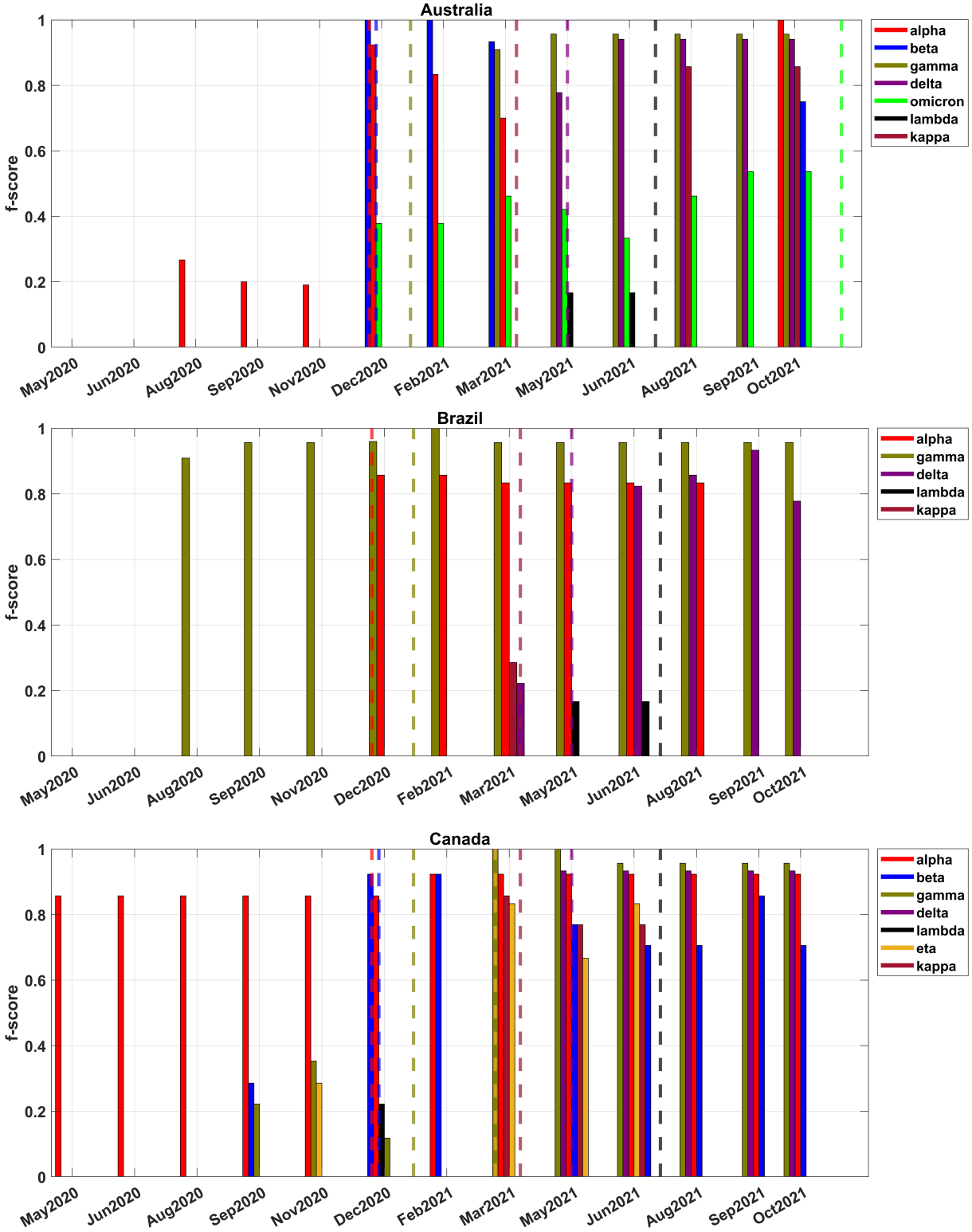

Figure A.15 **Comparison between VOCs and inferred haplotypes for selected countries** (Part 1). At each time point, each bar represents an inferred haplotype closest to a particular VOC, the bar height is equal to the respective f-score. The results are displayed for 13 uniformly distributed timepoints to avoid overcrowding of the figure. Colored dashed lines mark times when specific VOCs were designated by WHO.

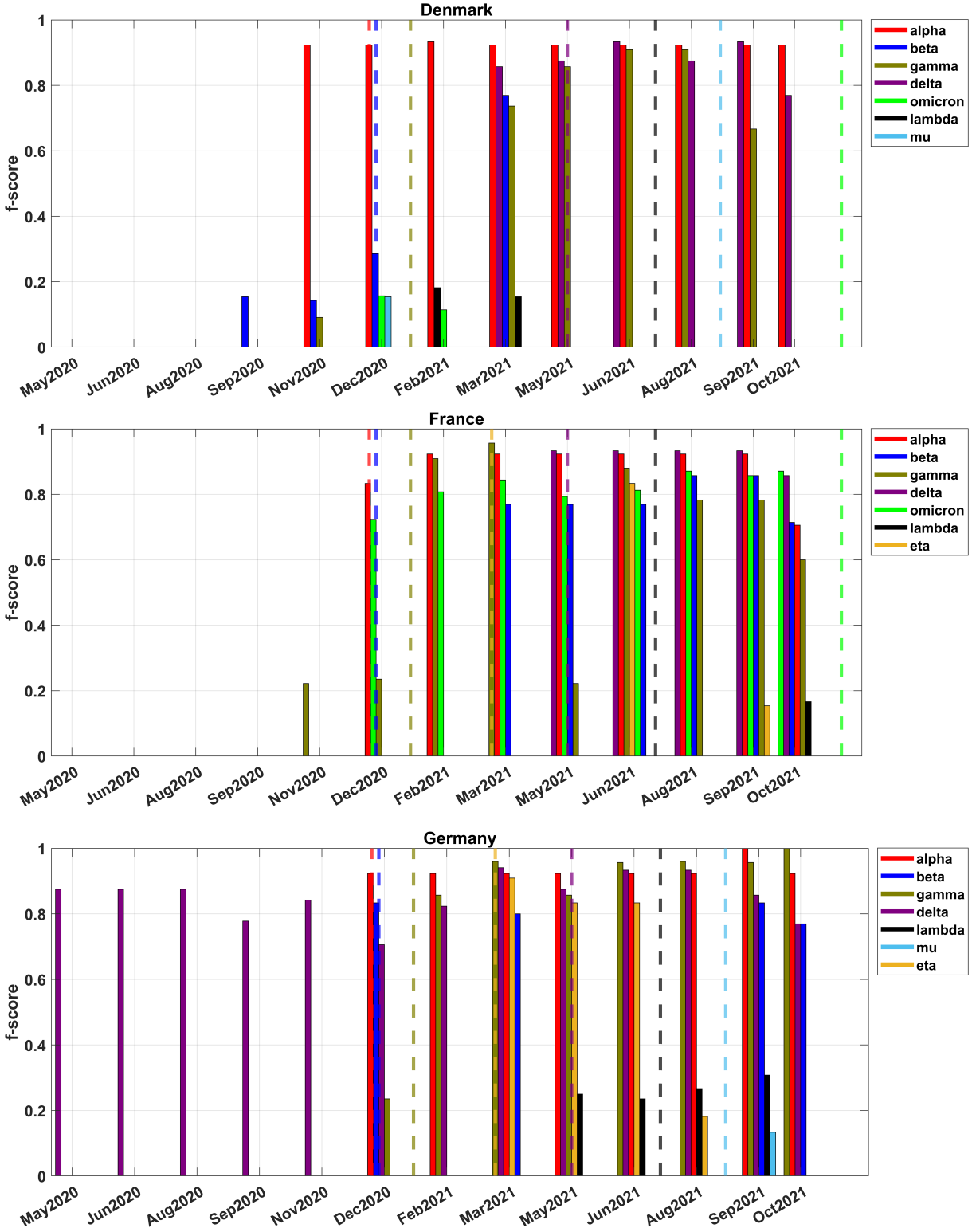

Figure A.16 **Comparison between VOCs and inferred haplotypes for selected countries** (Part 2). At each time point, each bar represents an inferred haplotype closest to a particular VOC, the bar height is equal to the respective f-score. The results are displayed for 13 uniformly distributed timepoints to avoid overcrowding of the figure. Colored dashed lines mark times when specific VOCs were designated by WHO.

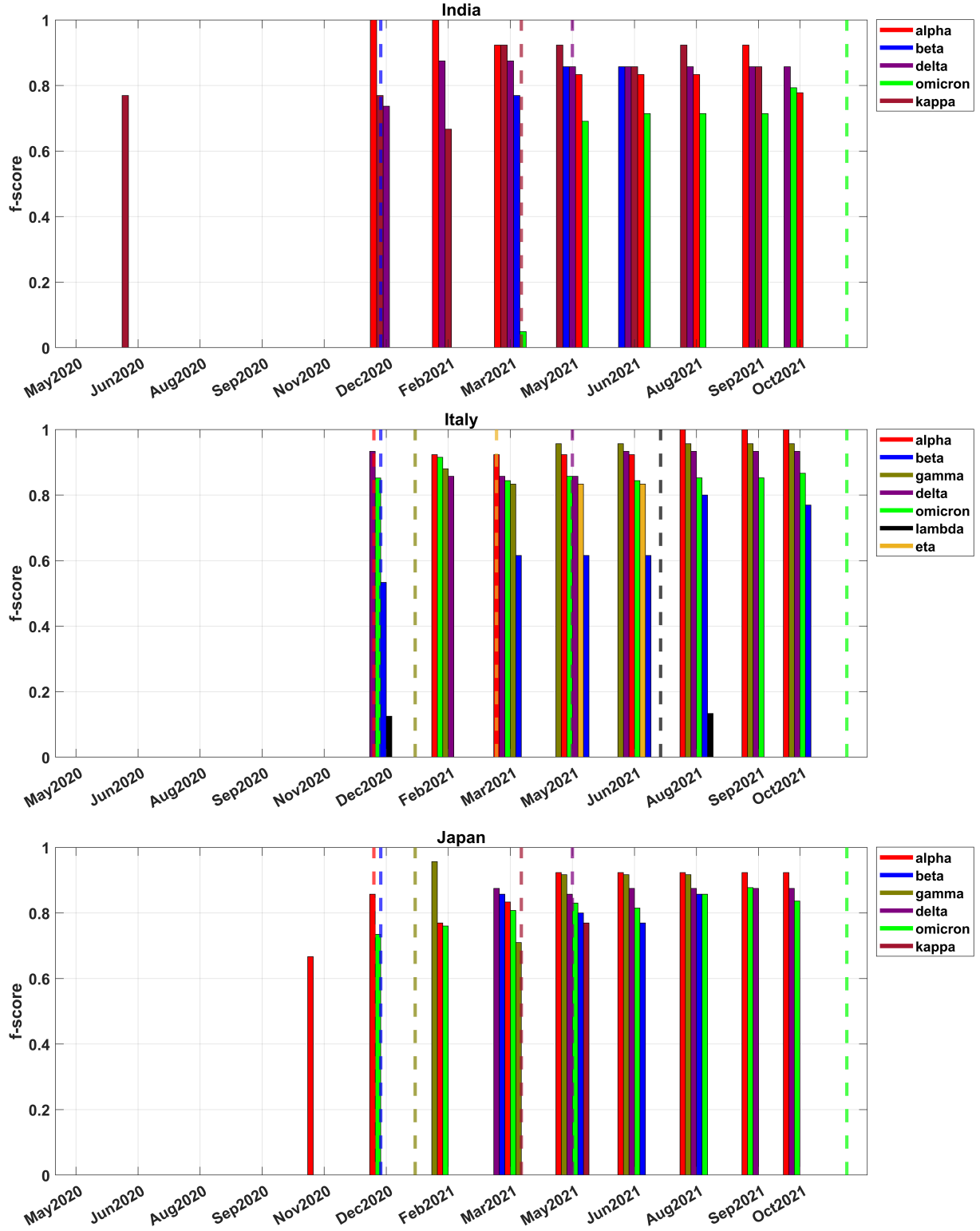

Figure A.17 **Comparison between VOCs and inferred haplotypes for selected countries** (Part 3). At each time point, each bar represents an inferred haplotype closest to a particular VOC, the bar height is equal to the respective f-score. The results are displayed for 13 uniformly distributed timepoints to avoid overcrowding of the figure. Colored dashed lines mark times when specific VOCs were designated by WHO.

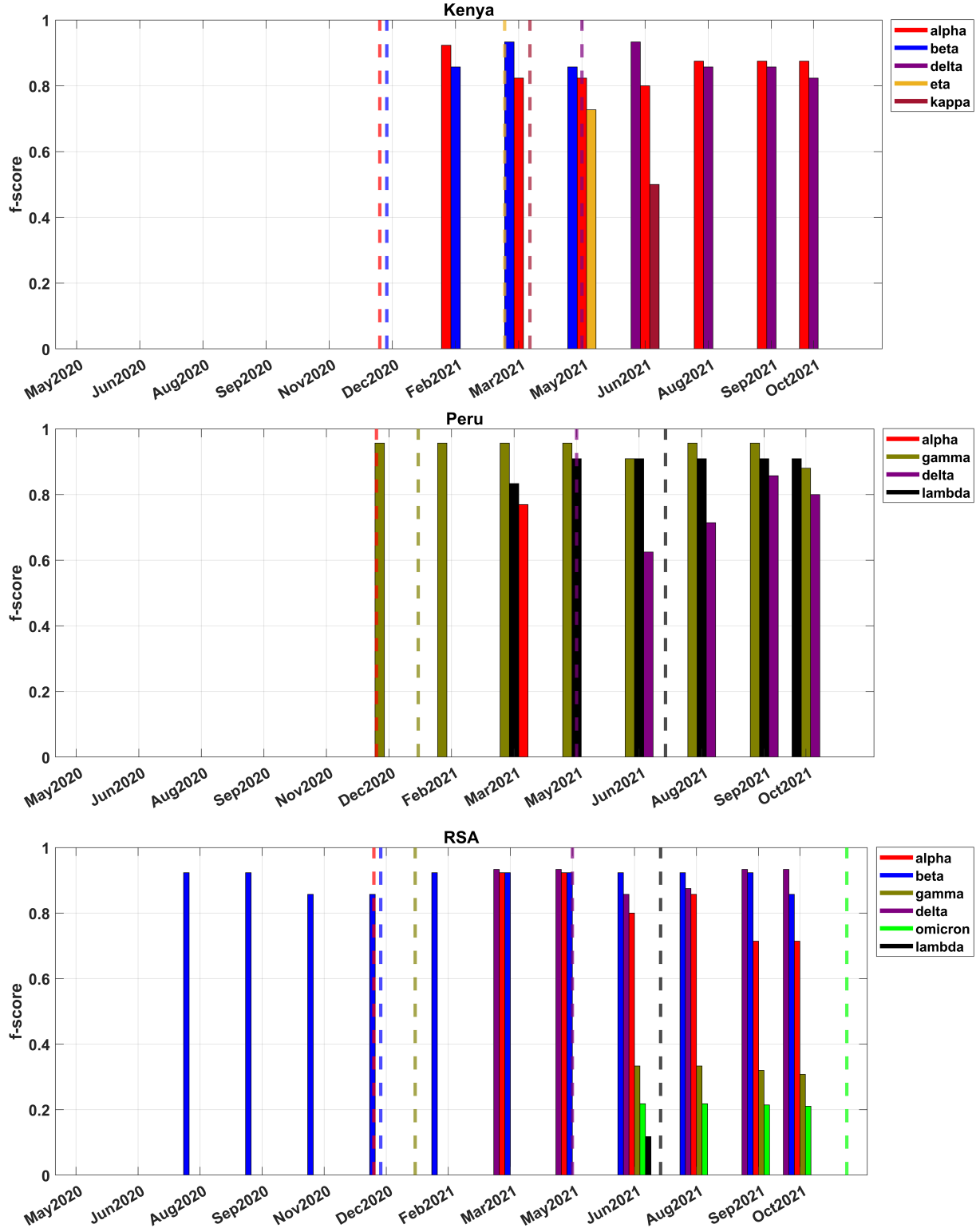

Figure A.18 **Comparison between VOCs and inferred haplotypes for selected countries** (Part 4). At each time point, each bar represents an inferred haplotype closest to a particular VOC, the bar height is equal to the respective f-score. The results are displayed for 13 uniformly distributed timepoints to avoid overcrowding of the figure. Colored dashed lines mark times when specific VOCs were designated by WHO.

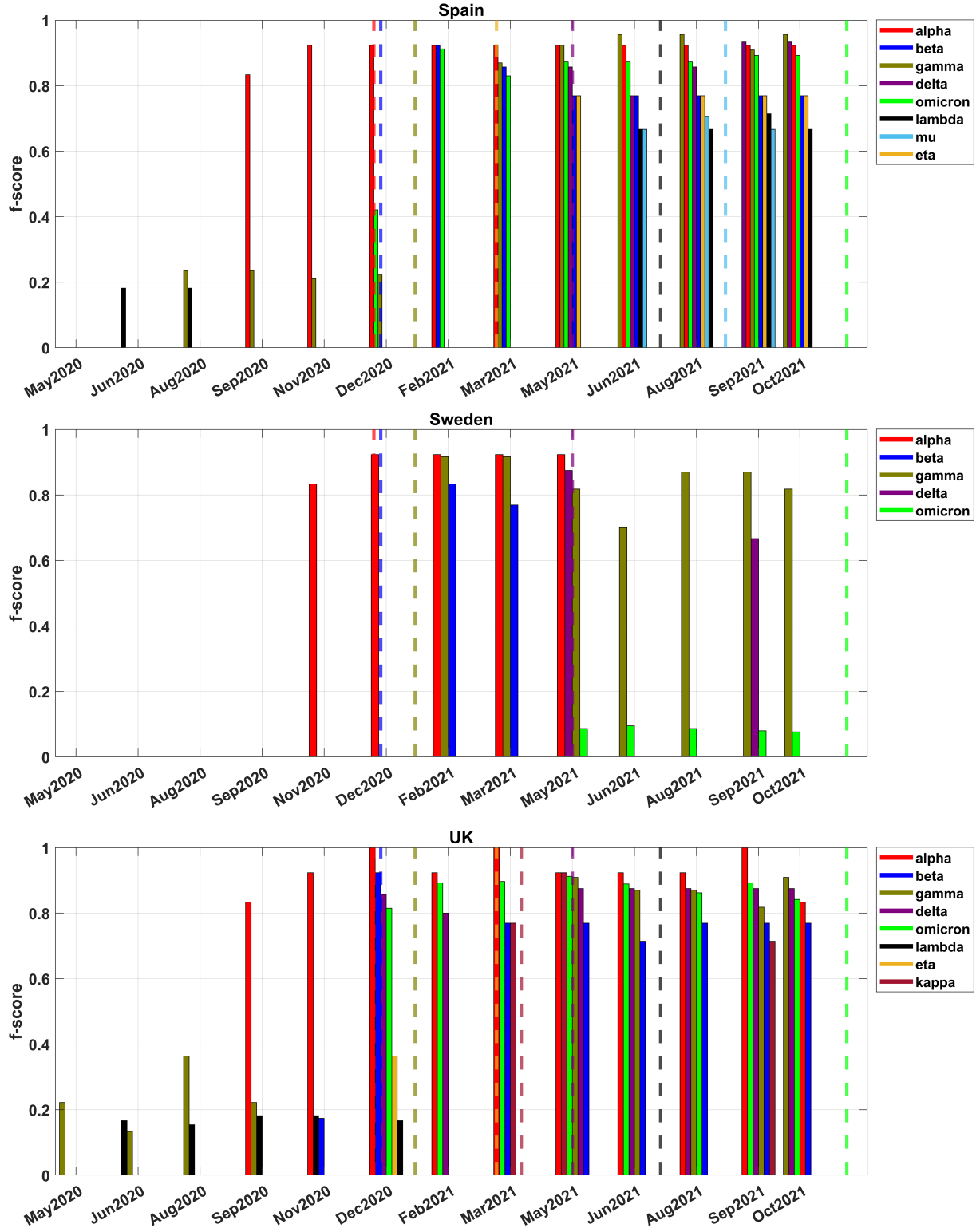

Figure A.19 **Comparison between VOCs and inferred haplotypes for selected countries** (Part 5). At each time point, each bar represents an inferred haplotype closest to a particular VOC, the bar height is equal to the respective f-score. The results are displayed for 13 uniformly distributed timepoints to avoid overcrowding of the figure. Colored dashed lines mark times when specific VOCs were designated by WHO.

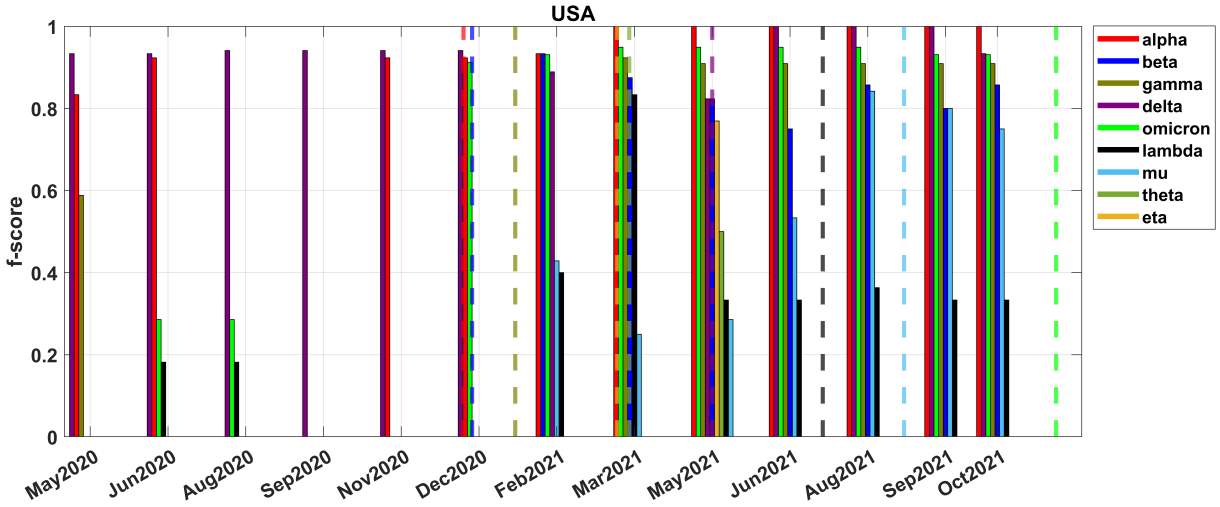

Figure A.20 **Comparison between VOCs and inferred haplotypes for selected countries** (Part 6). At each time point, each bar represents an inferred haplotype closest to a particular VOC, the bar height is equal to the respective f-score. The results are displayed for 13 uniformly distributed timepoints to avoid overcrowding of the figure. Colored dashed lines mark times when specific VOCs were designated by WHO.

REFERENCES

1. S. F. Ahmed, A. A. Quadeer, and M. R. McKay. Covidep: a web-based platform for real-time reporting of vaccine target recommendations for sars-cov-2. *Nature Protocols*, 15(7): 2141–2142, 2020.
2. S. Ahn and H. Vikalo. abayesqr: A bayesian method for reconstruction of viral populations characterized by low diversity. *Journal of Computational Biology*, 25:637–648, 2018.
3. F. Ali, A. Kasry, and M. Amin. The new sars-cov-2 strain shows a stronger binding affinity to ace2 due to n501y mutant. *Medicine in drug discovery*, 10:100086, 2021.
4. M. Anderberg. *Cluster Analysis for Applications*. Academic Press, 1973.
5. A. Apostolou, M. L. Bartholomew, R. Greeley, S. M. Guilfoyle, M. Gordon, C. Genese, J. P. Davis, B. Montana, and G. Borlaug. Transmission of hepatitis c virus associated with surgical procedures-new jersey 2010 and wisconsin 2011. *MMWR. Morbidity and mortality weekly report*, 64(7):165–170, 2015.
6. G. L. Armstrong, D. R. MacCannell, J. Taylor, H. A. Carleton, E. B. Neuhaus, R. S. Bradbury, J. E. Posey, and M. Gwinn. Pathogen genomics in public health. *New England Journal of Medicine*, 381(26):2569–2580, 2019.
7. A. Artyomenko, N. Mancuso, A. Zelikovsky, P. Skums, and I. Mandoiu. kgem: An em-based algorithm for local reconstruction of viral quasispecies. In *Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference on*, pages 1–1. IEEE, 2013.

8. A. Artyomenko, N. C. Wu, S. Mangul, E. Eskin, R. Sun, and A. Zelikovsky. Long single-molecule reads can resolve the complexity of the influenza virus composed of rare, closely related mutant variants. *Journal of Computational Biology*, 24(6):558–570, 2017.

9. Y. Asahiro, R. Hassin, and K. Iwama. Complexity of finding dense subgraphs. *Discrete Applied Mathematics*, 121(1-3):15–26, 2002.

10. C. Audet and J. E. Dennis Jr. Analysis of generalized pattern searches. *SIAM Journal on optimization*, 13(3):889–903, 2002.

11. J. Baaijens, A. Aabidine, E. Rivals, and A. Schönhuth. De novo assembly of viral quasispecies using overlap graphs. *Genome research*, 27:835–848, 2017.

12. C. Bai, J. Wang, G. Chen, H. Zhang, K. An, P. Xu, Y. Du, R. D. Ye, A. Saha, A. Zhang, et al. Predicting mutational effects on receptor binding of the spike protein of sars-cov-2 variants. *Journal of the American Chemical Society*, 143(42):17646–17654, 2021.

13. A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439): 509–512, 1999.

14. P. Barrat-Charlaix, J. Huddleston, T. Bedford, and R. A. Neher. Limited predictability of amino acid substitutions in seasonal influenza viruses. *Molecular Biology and Evolution*, 38 (7):2767–2777, 2021.

15. P. B. I. Baykal, J. Lara, Y. Khudyakov, A. Zelikovsky, and P. Skums. Quantitative differences between intra-host hcv populations from persons with recently established and persistent infections. *Virus Evolution*, 6(2):veaa103, 2021.

16. N. Beerenwinkel, L. Pachter, and B. Sturmfels. Epistasis and shapes of fitness landscapes.

Statistica Sinica, pages 1317–1342, 2007.

17. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

18. A. Black, D. R. MacCannell, T. R. Sibley, and T. Bedford. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nature Medicine*, pages 1–10, 2020.

19. G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

20. A. J. L. Brown, S. J. Lycett, L. Weinert, G. J. Hughes, E. Fearnhill, and D. T. Dunn. Transmission network parameters estimated from hiv sequences for a nationwide epidemic. *Journal of Infectious Diseases*, page jir550, 2011.

21. C. Burch, P. Turner, and K. Hanley. Patterns of epistasis in rna viruses: a review of the evidence from vaccine design. *Journal of evolutionary biology*, 16(6):1223–1235, 2003.

22. D. Cai and Y. Sun. Reconstructing viral haplotypes using long reads. *Bioinformatics*, 38(8): 2127–2134, 2022.

23. E. M. Campbell, H. Jia, A. Shankar, D. Hanson, W. Luo, S. Masciotra, S. M. Owen, A. M. Oster, R. R. Galang, M. W. Spiller, et al. Detailed transmission network analysis of a large opiate-driven outbreak of hiv infection in the united states. *The Journal of infectious diseases*, 216(9):1053–1062, 2017.

24. E. M. Campbell, A. Boyles, A. Shankar, J. Kim, S. Knyazev, R. Cintron, and W. M. Switzer.

Microbetrace: retooling molecular epidemiology for rapid public health response. *PLoS computational biology*, 17(9):e1009300, 2021.

25. F. Campbell, X. Didelot, R. Fitzjohn, N. Ferguson, A. Cori, and T. Jombart. outbreaker2: a modular platform for outbreak reconstruction. *BMC bioinformatics*, 19(11):1–8, 2018.
26. F. Campbell, A. Cori, N. Ferguson, and T. Jombart. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS computational biology*, 15(3):e1006930, 2019.
27. D. Campo, Z. Dimitrova, R. J. Mitchell, J. Lara, and Y. Khudyakov. Coordinated evolution of the hepatitis c virus. *Proceedings of the National Academy of Sciences*, 105(28):9685–9690, 2008.
28. D. S. Campo and Y. Khudyakov. Intelligent network disruption analysis (indra): A targeted strategy for efficient interruption of hepatitis c transmissions. *Infection, Genetics and Evolution*, 63:204–215, 2018.
29. M. Charikar. Greedy approximation algorithms for finding dense components in a graph. In *International workshop on approximation algorithms for combinatorial optimization*, pages 84–95. Springer, 2000.
30. H. Cherifi, G. Palla, B. K. Szymanski, and X. Lu. On community structure in complex networks: challenges and opportunities. *Applied Network Science*, 4(1):1–35, 2019.
31. G. Chowell, L. Simonsen, C. Viboud, and Y. Kuang. Is west africa approaching a catastrophic phase or is the 2014 ebola epidemic slowing down? different models yield different answers for liberia. *PLoS currents*, 6, 2014.

32. G. Chowell, A. Tariq, and J. M. Hyman. A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. *BMC medicine*, 17(1):164, 2019.

33. F. Chung and L. Lu. The average distance in a random graph with given expected degrees. *Internet Mathematics*, 1(1):91–113, 2004.

34. F. Chung, L. Lu, and V. Vu. Spectra of random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 100(11):6313–6318, 2003. ISSN 0027-8424. doi: 10.1073/pnas.0937490100. URL `https://www.pnas.org/content/100/11/6313`.

35. S. Ciccolella*, M. Patterson*, P. Bonizzoni, and G. D. Vedova. Effective clustering for single cell sequencing cancer data. In *the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB, Niagara Falls, NY, USA, 2019)*, ACM, pages 437–446, 2019. doi: 10.1145/3307339.3342149.

36. S. Ciccolella, C. Ricketts, M. S. Gomez, M. Patterson, D. Silverbush, P. Bonizzoni, I. Hajirasouliha, and G. D. Vedova. Inferring cancer progression from single-cell sequencing while allowing mutation losses. *Bioinformatics*, btaa722, 2020. doi: 10.1093/bioinformatics/btaa722.

37. S. Ciccolella, M. Soto Gomez, M. D. Patterson, G. Della Vedova, I. Hajirasouliha, and P. Bonizzoni. gpps: an ilp-based approach for inferring cancer progression with mutation losses from single cell data. *BMC bioinformatics*, 21(1):1–16, 2020.

38. C. Conrad, H. M. Bradley, D. Broz, S. Buddha, E. L. Chapman, R. R. Galang, D. Hillman, J. Hon, K. W. Hoover, M. R. Patel, et al. Community outbreak of hiv infection linked to

injection drug use of oxymorphone—indiana, 2015. *MMWR. Morbidity and mortality weekly report*, 64(16):443, 2015.

39. L. Corey, C. Beyrer, M. S. Cohen, N. L. Michael, T. Bedford, and M. Rolland. Sars-cov-2 variants in patients with immunosuppression, 2021.

40. A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology*, 178(9):1505–1512, 2013.

41. E. M. Cottam, G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley, D. J. Paton, D. P. King, and D. T. Haydon. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society of London B: Biological Sciences*, 275(1637):887–895, 2008.

42. K. Crona, A. Gavryushkin, D. Greene, and N. Beerenwinkel. Inferring genetic interactions from comparative fitness data. *eLife*, 6:e28629, 2017.

43. N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. Pearson, T. Russell, D. Tully, A. D. Washburne, et al. Estimated transmissibility and severity of novel sars-cov-2 variant of concern 202012/01 in england. 2020.

44. N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne, et al. Estimated transmissibility and impact of sars-cov-2 lineage b. 1.1. 7 in england. *Science*, 372(6538):eabg3055, 2021.

45. A. de Bernardi Schneider, C. T. Ford, R. Hostager, J. Williams, M. Cioce, Ü. V. Çatalyürek, J. O. Wertheim, and D. Janies. Strainhub: A phylogenetic tool to construct pathogen trans-

mission networks. *Bioinformatics*, 36(3):945–947, 2020.

46. N. De Maio, C.-H. Wu, and D. J. Wilson. Scotti: efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS computational biology*, 12(9): e1005130, 2016.

47. N. De Maio, C. J. Worby, D. J. Wilson, and N. Stoesser. Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS computational biology*, 14(4):e1006117, 2018.

48. J. Deasy, E. Rocheteau, K. Kohler, D. J. Stubbs, P. Barbiero, P. Liò, and A. Ercole. Forecasting ultra-early intensive care strain from covid-19 in england. 2020. doi:10.1101/2020.03.19.20039057.

49. P. L. Delamater, E. J. Street, T. F. Leslie, Y. T. Yang, and K. H. Jacobsen. Complexity of the basic reproduction number (r0). *Emerging infectious diseases*, 25(1):1, 2019.

50. S. Dhar, C. Zhang, I. Mandoiu, and M. S. Bansal. Tnet: Phylogeny-based inference of disease transmission networks using within-host strain diversity. In *International Symposium on Bioinformatics Research and Applications*, pages 203–216. Springer, 2020.

51. S. Dhar, C. Zhang, I. Mandoiu, and M. S. Bansal. Tnet: Transmission network inference using within-host strain diversity and its application to geographical tracking of covid-19 spread. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.

52. X. Didelot, J. Gardy, and C. Colijn. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular biology and evolution*, 31(7):1869–1879, 2014.

53. X. Didelot, C. Fraser, J. Gardy, and C. Colijn. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular biology and evolution*, 34(4):997–1007, 2017.

54. P. T. Dolan, Z. J. Whitfield, and R. Andino. Mapping the evolutionary potential of rna viruses. *Cell host & microbe*, 23(4):435–446, 2018.

55. E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

56. Z. Drezner and H. W. Hamacher. *Facility location: applications and theory*. Springer Science & Business Media, 2001.

57. A. J. Drummond and A. Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7(1):214, 2007.

58. M. Eigen, J. McCaskill, and P. Schuster. Molecular quasi-species. *The Journal of Physical Chemistry*, 92(24):6881–6891, 1988.

59. S. Elbe and G. Buckland-Merrett. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1:33–46, 2017. doi: 10.1002/gch2.1018.

60. S. F. Elena, R. V. Solé, and J. Sardanyés. Simple genomes, complex interactions: epistasis in rna virus. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(2):026106, 2010.

61. EMBL-EBI. Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK, 2020.

62. A. H. Esfahanian and S. Louis Hakimi. On computing the connectivities of graphs and digraphs. *Networks*, 14(2):355–366, 1984.

63. S. Even and R. E. Tarjan. Network flow and testing graph connectivity. *SIAM journal on*

computing, 4(4):507–518, 1975.

64. U. Feige, D. Peleg, and G. Kortsarz. The dense k-subgraph problem. *Algorithmica*, 29(3): 410–421, 2001.

65. J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, 2003.

66. W. Fitch. Towards defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology*, 20:406–416, 1971.

67. C. Fraïsse and J. J. Welch. The distribution of epistasis on simple fitness landscapes. *Biology letters*, 15(4):20180881, 2019.

68. C. Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PloS one*, 2(8):e758, 2007.

69. T. M. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

70. S. E. Galloway, P. Paul, D. R. MacCannell, M. A. Johansson, J. T. Brooks, A. MacNeil, R. B. Slayton, S. Tong, B. J. Silk, G. L. Armstrong, et al. Emergence of sars-cov-2 b. 1.1. 7 lineage—united states, december 29, 2020–january 12, 2021. *Morbidity and Mortality Weekly Report*, 70:95–99, 2021.

71. A. P. Galvani and R. M. May. Dimensions of superspreading. *Nature*, 438(7066):293–295, 2005.

72. T. Ganyani, C. Kremer, D. Chen, A. Torneri, C. Faes, J. Wallinga, and N. Hens. Estimating the generation interval for coronavirus disease (covid-19) based on symptom onset data, march 2020. *Eurosurveillance*, 25(17):2000257, 2020.

73. W. F. Garcia-Beltran, E. C. Lam, K. S. Denis, A. D. Nitido, Z. H. Garcia, B. M. Hauser, J. Feldman, M. N. Pavlovic, D. J. Gregory, M. C. Poznansky, et al. Multiple sars-cov-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell*, 184(9):2372–2383, 2021.

74. O. Glebova, S. Knyazev, A. Melnyk, A. Artyomenko, Y. Khudyakov, A. Zelikovsky, and P. Skums. Inference of genetic relatedness between viral quasispecies from sequencing data. *BMC genomics*, 18(10):918, 2017.

75. G. S. Gonsalves and F. W. Crawford. Dynamics of the hiv outbreak and response in scott county, in, usa, 2011–15: a modelling study. *The lancet HIV*, 5(10):e569–e577, 2018.

76. S. J. Gould. *Wonderful life: the Burgess Shale and the nature of history*. WW Norton & Company, 1990.

77. S. Guindon, J. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.*, 59(3):307–321, 2010. doi: 10.1093/sysbio/syq010.

78. L. Gurobi Optimization. Gurobi optimizer reference manual, 2019. URL `http://www.gurobi.com`.

79. M. T. Hajiaghayi, M. Mahdian, and V. S. Mirrokni. The facility location problem with general cost functions. *Networks: An International Journal*, 42(1):42–47, 2003.

80. M. Hall, M. Woolhouse, and A. Rambaut. Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. *PLoS computational biology*, 11(12): e1004613, 2015.

81. M. D. Hall and C. Colijn. Transmission trees on a known pathogen phylogeny: Enumeration and sampling. *Molecular biology and evolution*, 36(6):1333–1343, 2019.

82. M. Hoffmann, H. Kleine-Weber, and S. Pöhlmann. A multibasic cleavage site in the spike protein of sars-cov-2 is essential for infection of human lung cells. *Molecular Cell*, 2020.

83. M. Hoffmann, P. Arora, R. Groß, A. Seidel, B. F. Hörnich, A. S. Hahn, N. Krüger, L. Graichen, H. Hofmann-Winkler, A. Kempf, et al. Sars-cov-2 variants b. 1.351 and p. 1 escape from neutralizing antibodies. *Cell*, 184(9):2384–2393, 2021.

84. Z. Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 1–8, 1997.

85. Z. Huang. Extensions to the k-modes algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998.

86. G. J. Hughes, E. Fearnhill, D. Dunn, S. J. Lycett, A. Rambaut, A. J. L. Brown, and U. H. D. R. Collaboration. Molecular phylodynamics of the heterosexual hiv epidemic in the united kingdom. *PLoS pathogens*, 5(9):e1000590, 2009.

87. P. B. Icer Baykal, J. Lara, Y. Khudyakov, A. Zelikovsky, and P. Skums. Quantitative differences between intra-host hcv populations from persons with recently established and persistent infections. *Virus evolution*, 7(1):veaa103, 2021.

88. K. Jahn, J. Kuipers, and N. Beerenwinkel. Tree inference for single-cell data. *Genome Biology*, 17(1):86, 2016. doi: 10.1186/s13059-016-0936-x.

89. D. Jha, P. Skums, A. Zelikovsky, Y. Khudyakov, and R. Singh. Modeling the spread of hiv and

hcv infections based on identification and characterization of high-risk communities using social media. In *International Symposium on Bioinformatics Research and Applications*, pages 425–430. Springer, Cham, 2017.

90. X. Jiao, H. Imamichi, B. T. Sherman, R. Nahar, R. L. Dewar, H. C. Lane, T. Imamichi, and W. Chang. Quasiseq: profiling viral quasispecies via self-tuning spectral clustering with pacbio long sequencing reads. *Bioinformatics*, 2022.

91. Jombart, A. Cori, X. Didelot, S. Cauchemez, C. Fraser, and N. Ferguson. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput Biol*, 10(1):e1003457, 2014.

92. T. Jombart, R. Eggo, P. Dodd, and F. Balloux. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390, 2011.

93. K. Katoh and D. M. Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

94. L. Kirola. Genetic emergence of b. 1.617. 2 in covid-19. *New Microbes and New Infections*, 43:100929, 2021.

95. D. Klinkenberg, J. A. Backer, X. Didelot, C. Colijn, and J. Wallinga. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS computational biology*, 13(5):e1005495, 2017.

96. S. Knyazev, V. Tsyvina, A. Melnyk, A. Artyomenko, T. Malygina, Y. B. Porozov, E. Campbell, W. M. Switzer, P. Skums, and A. Zelikovsky. Cliquesnv: Scalable reconstruction of

intra-host viral populations from ngs reads. *bioRxiv*, page 264242, 2019.

97. S. Knyazev, L. Hughes, P. Skums, and A. Zelikovsky. Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. *Briefings in Bioinformatics*, 2020.

98. S. Knyazev, L. Hughes, P. Skums, and A. Zelikovsky. Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. *Briefings in bioinformatics*, 22:96–108, 2021.

99. S. Knyazev, L. Hughes, P. Skums, and A. Zelikovsky. Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. *Briefings in bioinformatics*, 22(1): 96–108, 2021.

100. S. Knyazev, V. Tsyvina, A. Shankar, A. Melnyk, A. Artyomenko, T. Malygina, Y. B. Porozov, E. M. Campbell, W. M. Switzer, P. Skums, et al. Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction. *Nucleic acids research*, 49(17):e102–e102, 2021.

101. S. Knyazev, K. Chhugani, V. Sarwal, R. Ayyala, H. Singh, S. Karthikeyan, D. Deshpande, P. I. Baykal, Z. Comarova, A. Lu, et al. Unlocking capacities of genomics for the covid-19 response and future pandemics. *Nature Methods*, 19(4):374–380, 2022.

102. S. L. Kosakovsky Pond, S. Weaver, A. J. Leigh Brown, and J. O. Wertheim. Hiv-trace (transmission cluster engine): a tool for large scale molecular epidemiology of hiv-1 and other rapidly evolving pathogens. *Molecular biology and evolution*, 35(7):1812–1819, 2018.

103. K. Kupferschmidt. Where did ‘weird’omicron come from?, 2021.

104. M. Lässig, V. Mustonen, and A. M. Walczak. Predicting evolution. *Nature ecology & evolu-*

tion, 1(3):1–9, 2017.

105. K. Leung, M. H. Shum, G. M. Leung, T. T. Lam, and J. T. Wu. Early empirical assessment of the n501y mutant strains of sars-cov-2 in the united kingdom, october to november 2020. *medRxiv*, 2020.

106. T. Li, S. Ma, and M. Ogihara. Entropy-based criterion in categorical clustering. In *Proc. 21st International Conference on Machine Learning (ICML), 2004*, volume 3, pages 536–543, 2004.

107. F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Åberg. The web of human sexual contacts. *Nature*, 411(6840):907–908, 2001.

108. Q.-H. Liu, M. Ajelli, A. Aleta, S. Merler, Y. Moreno, and A. Vespignani. Measurability of the epidemic reproduction number in data-driven contact networks. *Proceedings of the National Academy of Sciences*, 115(50):12680–12685, 2018.

109. Y. Liu, J. Liu, B. A. Johnson, H. Xia, Z. Ku, C. Schindewolf, S. G. Widen, Z. An, S. C. Weaver, V. D. Menachery, et al. Delta spike p681r mutation enhances sars-cov-2 fitness over alpha variant. *bioRxiv*, 2021.

110. Y. Liu, J. Kearney, M. Mahmoud, B. Kille, F. J. Sedlazeck, and T. J. Treangen. Rescuing low frequency variants within intra-host viral populations directly from oxford nanopore sequencing data. *Nature communications*, 13(1):1–9, 2022.

111. Y. Liu, J. Liu, B. A. Johnson, H. Xia, Z. Ku, C. Schindewolf, S. G. Widen, Z. An, S. C. Weaver, V. D. Menachery, et al. Delta spike p681r mutation enhances sars-cov-2 fitness over alpha variant. *Cell Reports*, 39(7):110829, 2022.

112. A. G. Longmire, S. Sims, I. Rytsareva, D. S. Campo, P. Skums, Z. Dimitrova, S. Ramachandran, M. Medrzycki, H. Thai, L. Ganova-Raeva, et al. Ghost: global hepatitis outbreak and surveillance technology. *BMC genomics*, 18(10):916, 2017.

113. B. Luan, H. Wang, and T. Huynh. Enhanced binding of the n501y-mutated sars-cov-2 spike protein to the human ace2 receptor: insights from molecular dynamics simulations. *FEBS letters*, 595(10):1454–1461, 2021.

114. K. A. Lythgoe, M. Hall, L. Ferretti, M. de Cesare, G. MacIntyre-Cockett, A. Trebes, M. Andersson, N. Otecko, E. L. Wise, N. Moore, et al. Sars-cov-2 within-host diversity and transmission. *Science*, 372(6539):eabg0821, 2021.

115. O. A. MacLean, R. J. Orton, J. B. Singer, and D. L. Robertson. No evidence for distinct types in the evolution of sars-cov-2. *Virus Evolution*, 6(1):veaa034, 2020.

116. J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297, 1967.

117. M. C. Maher, I. Bartha, S. Weaver, J. Di Iulio, E. Ferri, L. Soriaga, F. A. Lempp, B. L. Hie, B. Bryson, B. Berger, et al. Predicting the mutational drivers of future sars-cov-2 variants of concern. *Science translational medicine*, 14(633):eabk3445, 2022.

118. M. McCallum, J. Bassi, A. Marco, A. Chen, A. Walls, J. Iulio, M. Tortorici, M. Navarro, C. Silacci-Fregni, C. Saliba, M. Agostini, D. Pinto, K. Culap, S. Bianchi, S. Jaconi, E. Cameroni, J. Bowen, S. Tilles, M. Pizzuto, S. Guastalla, G. Bona, A. Pellanda, C. Garzoni, W. Van Voorhis, L. Rosen, G. Snell, A. Telenti, H. Virgin, L. Piccoli,

D. Corti, and D. Veesler. Sars-cov-2 immune evasion by variant b.1.427/b.1.429. 1, 2021. doi:10.1101/2021.03.31.437925.

119. A. Melnyk, F. Mohebbi, S. Knyazev, B. Sahoo, R. Hosseini, P. Skums, A. Zelikovsky, and M. Patterson. Clustering based identification of sars-cov-2 subtypes. In *International Conference on Computational Advances in Bio and Medical Sciences*, pages 127–141. Springer, 2020.

120. A. Melnyk, F. Mohebbi, S. Knyazev, B. Sahoo, R. Hosseini, P. Skums, A. Zelikovsky, and M. Patterson. From alpha to zeta: Identifying variants and subtypes of sars-cov-2 via clustering. *Journal of Computational Biology*, 28(11):1113–1129, 2021.

121. N. Mollentze, L. H. Nel, S. Townsend, K. Le Roux, K. Hampson, D. T. Haydon, and S. Soubeyrand. A bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1782):20133251, 2014.

122. M. J. Morelli, G. Thébaud, J. Chadœuf, D. P. King, D. T. Haydon, and S. Soubeyrand. A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput Biol*, 8(11):e1002768, 2012.

123. N. Moshiri, M. Ragonnet-Cronin, J. O. Wertheim, and S. Mirarab. Favites: simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics*, 35 (11):1852–1861, 2019.

124. A. Moulana, T. Dupic, A. M. Phillips, J. Chang, S. Nieves, A. A. Roffler, A. J. Greaney, T. N. Starr, J. D. Bloom, and M. M. Desai. Compensatory epistasis maintains ace2 affinity in

sars-cov-2 omicron ba. 1. *Nature Communications*, 13(1):7011, 2022.

125. S. A. Nadeau, T. G. Vaughan, J. Scire, J. S. Huisman, and T. Stadler. The origin and early spread of sars-cov-2 in europe. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

126. F. Naveca, C. da Costa, V. Nascimento, V. Souza, A. Corado, F. Nascimento, Á. Costa, D. Duarte, G. Silva, M. Mejía, et al. Sars-cov-2 reinfection by the new variant of concern (voc) p. 1 in amazonas, brazil. *Virological.org*, 2021.

127. F. Naveca, V. Nascimento, V. Souza, A. Corado, F. Nascimento, G. Silva, A. Costa, D. Duarte, K. Pessoa, L. Gonçalves, et al. Phylogenetic relationship of sars-cov-2 sequences from amazonas with emerging brazilian variants harboring mutations e484k and n501y in the spike protein. *Virological. org*, 2021.

128. A. D. Neverov, G. Fedonin, A. Popova, D. Bykova, and G. Bazykin. Coordinated evolution at amino acid sites of sars-cov-2 spike. *Elife*, 12:e82516, 2023.

129. M. Newman. *Networks: an introduction*. Oxford University Press, 2010.

130. M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

131. A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

132. L.-T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh. Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1):268–274, 2015.

133. H. Nishiura and G. Chowell. The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In *Mathematical and statistical estimation approaches in epidemiology*, pages 103–121. Springer, 2009.

134. M. A. Nowak. *Evolutionary dynamics*. Harvard University Press, 2006.

135. F. Obermeyer, M. Jankowiak, N. Barkas, S. F. Schaffner, J. D. Pyle, L. Yurkovetskiy, M. Bosso, D. J. Park, M. Babadi, B. L. MacInnis, et al. Analysis of 6.4 million sars-cov-2 genomes identifies mutations associated with fitness. *Science*, 376(6599):1327–1332, 2022.

136. T. Ohta. The nearly neutral theory of molecular evolution. *Annual review of ecology and systematics*, pages 263–286, 1992.

137. W. H. Organization et al. Scientific advisory group for the origins of novel pathogens. *nd, https://www. who. int/groups/scientific-advisory-group-on-the-origins-of-novel-pathogens-(sago)(accessed June 25, 2022)*, 2021.

138. Á. O'Toole, O. G. Pybus, M. E. Abram, E. J. Kelly, and A. Rambaut. Pango lineage designation and assignment using sars-cov-2 spike gene nucleotide sequences. *BMC genomics*, 23 (1):1–13, 2022.

139. G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *nature*, 435(7043):814, 2005.

140. B. Pell, Y. Kuang, C. Viboud, and G. Chowell. Using phenomenological models for forecasting the 2015 ebola challenge. *Epidemics*, 22:62–70, 2018.

141. P. J. Peters, P. Pontones, K. W. Hoover, M. R. Patel, R. R. Galang, J. Shields, S. J. Blosser, M. W. Spiller, B. Combs, W. M. Switzer, et al. Hiv infection linked to injection use of

oxymorphone in indiana, 2014–2015. *New England Journal of Medicine*, 375(3):229–239, 2016.

142. D. Planas, D. Veyer, A. Baidaliuk, I. Staropoli, F. Guivel-Benhassine, M. M. Rajah, C. Planchais, F. Porrot, N. Robillard, J. Puech, et al. Reduced sensitivity of sars-cov-2 variant delta to antibody neutralization. *Nature*, 596(7871):276–280, 2021.

143. A. Popa, J.-W. Genger, M. D. Nicholson, T. Penz, D. Schmid, S. W. Aberle, B. Agerer, A. Lercher, L. Endler, H. Colaço, et al. Genomic epidemiology of superspreading events in austria reveals mutational dynamics and transmission properties of sars-cov-2. *Science translational medicine*, 12(573), 2020.

144. S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, and V. Roth. Hiv haplotype inference using a propagating dirichlet process mixture model. *IEEE/ACM transactions on computational biology and bioinformatics*, 11:182–191, 2014. doi: 10.1109/TCBB.2013.145.

145. T. Pupko, I. Pe, R. Shamir, and D. Graur. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular biology and evolution*, 17(6):890–896, 2000.

146. S. Ramachandran, H. Thai, J. C. Forbi, R. R. Galang, Z. Dimitrova, G.-l. Xia, Y. Lin, L. T. Punkova, P. R. Pontones, J. Gentry, et al. A large hcv transmission network enabled a fast-growing hiv outbreak in rural indiana, 2015. *EBioMedicine*, 37:374–381, 2018.

147. A. Rambaut, N. Loman, O. Pybus, W. Barclay, J. Barrett, A. Carabelli, T. Connor, T. Peacock, D. Robertson, and E. Volz. Preliminary genomic characterisation of an emergent sars-cov-2 lineage in the uk defined by a novel set of spike mutations. Accessed: 2020-12-25.

148. J. D. Ramírez, M. Muñoz, L. H. Patiño, N. Ballesteros, and A. Paniz-Mondolfi. Will the

emergent sars-cov2 b. 1.1. 7 lineage affect molecular diagnosis of covid-19? *Journal of Medical Virology*, 93:2566–2568, 2021.

149. L. Ramshaw and R. E. Tarjan. On minimum-cost assignments in unbalanced bipartite graphs. *HP Labs, Palo Alto, CA, USA, Tech. Rep. HPL-2012-40R1*, 2012.

150. O. Ratmann, M. K. Grabowski, M. Hall, T. Golubchik, C. Wymant, L. Abeler-Dörner, D. Bonsall, A. Hoppe, A. L. Brown, T. de Oliveira, et al. Inferring hiv-1 transmission networks and sources of epidemic spread in africa with deep-sequence phylogenetic analysis. *Nature communications*, 10(1):1–13, 2019.

151. J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical review E*, 74(1):016110, 2006.

152. N. D. Rochman, Y. I. Wolf, G. Faure, P. Mutz, F. Zhang, and E. V. Koonin. Ongoing global and regional adaptive evolution of sars-cov-2. *Proceedings of the National Academy of Sciences*, 118(29), 2021.

153. N. D. Rochman, G. Faure, Y. I. Wolf, P. L. Freddolino, F. Zhang, and E. V. Koonin. Epistasis at the sars-cov-2 receptor-binding domain interface and the propitiously boring implications for vaccine escape. *Mbio*, 13(2):e00135–22, 2022.

154. J. Rodriguez-Rivas, G. Croce, M. Muscat, and M. Weigt. Epistatic models predict mutable sites in sars-cov-2 proteins and epitopes. *Proceedings of the National Academy of Sciences*, 119(4):e2113118119, 2022.

155. C. M. Romano, I. M. G. de Carvalho-Mello, L. F. Jamal, F. L. de Melo, A. Iamarino, M. Motoki, J. R. R. Pinho, E. C. Holmes, P. M. de Andrade Zanotto, V. Consortium, et al. Social

networks shape the transmission dynamics of hepatitis c virus. *PLoS One*, 5(6):e11170, 2010.

156. E. O. Romero-Severson, I. Bulla, and T. Leitner. Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences*, page 201522930, 2016.

157. K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. Hyman, P. Yan, and G. Chowell. Real-time forecasts of the covid-19 epidemic in china from february 5th to february 24th, 2020. *Infectious Disease Modelling*, 5:256–263, 2020.

158. K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. M. Hyman, P. Yan, and G. Chowell. Short-term forecasts of the covid-19 epidemic in guangdong and zhejiang, china: February 13–23, 2020. *Journal of clinical medicine*, 9(2):596, 2020.

159. P. Sagulenko, V. Puller, and R. A. Neher. Treetime: Maximum-likelihood phylodynamic analysis. *Virus evolution*, 4(1):vex042, 2018.

160. Z. R. Sailer and M. J. Harms. Molecular ensembles make evolution unpredictable. *Proceedings of the National Academy of Sciences*, 114(45):11938–11943, 2017.

161. D. Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.

162. P. Sashittal and M. El-Kebir. Sharptni: counting and sampling parsimonious transmission networks under a weak bottleneck. *bioRxiv*, page 842237, 2019.

163. P. Sashittal and M. El-Kebir. Sampling and summarizing transmission trees with multi-strain infections. *Bioinformatics*, 36(Supplement_1):i362–i370, 2020.

164. T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18:6097–6100, 1990.

165. A. Schrijver. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer Science & Business Media, 2003.

166. D. Seifert, F. Di Giallonardo, K. J. Metzner, H. F. Günthard, and N. Beerenwinkel. A framework for inferring fitness landscapes of patient-derived viruses using quasispecies theory. *Genetics*, 199(1):191–203, 2015.

167. Y. Shu and J. McCauley. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13), 2017.

168. P. Skums and L. Bunimovich. Graph fractal dimension and the structure of fractal networks. *Journal of Complex Networks*, 8(4):cnaa037, 2020.

169. P. Skums, D. S. Campo, Z. Dimitrova, G. Vaughan, D. T. Lau, and Y. Khudyakov. Numerical detection, measuring and analysis of differential interferon resistance for individual hcv intrahost variants and its influence on the therapy response. *In silico biology*, 11(5-6):263–269, 2011.

170. P. Skums, A. Zelikovsky, R. Singh, W. Gussler, Z. Dimitrova, S. Knyazev, I. Mandric, S. Ramachandran, D. Campo, D. Jha, et al. Quentin: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics*, 34(1):163–170, 2017.

171. S. Sledzieski, C. Zhang, I. Mandoiu, and M. S. Bansal. Treefix-tp: Phylogenetic error-correction for infectious disease transmission network inference. *bioRxiv*, page 813931, 2019.

172. T. Stadler, D. Kühnert, S. Bonhoeffer, and A. J. Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). *Proceedings of the*

National Academy of Sciences, 110(1):228–233, 2013.

173. A. Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

174. T. N. Starr, L. K. Picton, and J. W. Thornton. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, 549(7672):409–413, 2017.

175. T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, et al. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *Cell*, 182(5): 1295–1310, 2020.

176. A. G. Suryaprasad, J. Z. White, F. Xu, B.-A. Eichler, J. Hamilton, A. Patel, S. B. Hamdounia, D. R. Church, K. Barton, C. Fisher, et al. Emerging epidemic of hepatitis c virus infections among young nonurban persons who inject drugs in the united states, 2006–2012. *Clinical infectious diseases*, 59(10):1411–1419, 2014.

177. K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular Biology and Evolution*, 10:512–526, 1993.

178. J. Tang, O. Toovey, K. Harvey, and D. Hui. Introduction of the south african sars-cov-2 variant 501y.v2 into the uk. *The Journal of Infection*, 82:e8–e10, 2021.

179. The World Health Organization. Tracking SARS-CoV-2 variants, 2022. URL `https://www.who.int/activities/tracking-SARS-CoV-2-variants`.

180. R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via

the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

181. V. d. F. Vieira, C. R. Xavier, and A. G. Evsukoff. A comparative study of overlapping community detection methods from the perspective of the structural properties. *Applied Network Science*, 5(1):1–42, 2020.

182. L. Villandre, D. A. Stephens, A. Labbe, H. F. Günthard, R. Kouyos, T. Stadler, S. H. C. Study, et al. Assessment of overlap of phylogenetic transmission clusters and communities in simple sexual contact networks: applications to hiv-1. *PloS one*, 11(2):e0148459, 2016.

183. H. Vöhringer, M. Sinnott, R. Amato, I. Martincorena, D. Kwiatkowski, J. C. Barrett, and M. Gerstung. Lineage-specific growth of sars-cov-2 b. 1.1. 7 during the english national lockdown, 2020.

184. E. Volz, V. Hill, J. T. McCrone, A. Price, D. Jorgensen, Á. O'Toole, J. Southgate, R. Johnson, B. Jackson, F. F. Nascimento, et al. Evaluating the effects of sars-cov-2 spike mutation d614g on transmissibility and pathogenicity. *Cell*, 184(1):64–75, 2021.

185. E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, Á. O'Toole, et al. Assessing transmissibility of sars-cov-2 lineage b. 1.1. 7 in england. *Nature*, 593:266–269, 2021.

186. P. Wang, M. S. Nair, L. Liu, S. Iketani, Y. Luo, Y. Guo, M. Wang, J. Yu, B. Zhang, P. D. Kwong, et al. Antibody resistance of sars-cov-2 variants b. 1.351 and b. 1.1. 7. *Nature*, 593 (7857):130–135, 2021.

187. C. Wei, K.-J. Shan, W. Wang, S. Zhang, Q. Huan, and W. Qian. Evidence for a mouse origin

of the sars-cov-2 omicron variant. *Journal of genetics and genomics*, 48(12):1111–1121, 2021.

188. D. M. Weinreich, Y. Lan, C. S. Wylie, and R. B. Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Current opinion in genetics & development*, 23(6):700–707, 2013.

189. S. Wernicke. Efficient detection of network motifs. *IEEE/ACM transactions on computational biology and bioinformatics*, 3(4):347–359, 2006.

190. J. O. Wertheim, A. J. Leigh Brown, N. L. Hepler, S. R. Mehta, D. D. Richman, D. M. Smith, and S. L. Kosakovsky Pond. The global transmission network of hiv-1. *The Journal of infectious diseases*, 209(2):304–313, 2014.

191. J. O. Wertheim, S. L. Kosakovsky Pond, L. A. Forgione, S. R. Mehta, B. Murrell, S. Shah, D. M. Smith, K. Scheffler, and L. V. Torian. Social and genetic networks of hiv-1 transmission in new york city. *PLoS pathogens*, 13(1):e1006000, 2017.

192. A. West, J. Wertheim, J. Wang, T. Vasylyeva, J. Havens, M. Chowdhury, E. Gonzalez, C. Fang, S. Di Lonardo, S. Hughes, J. Rakeman, H. Lee, C. Barnes, P. Gnanapragasam, Z. Yang, C. Gaebler, M. Caskey, M. Nussenzweig, B. Keeffe, JR, and P.J. Detection and characterization of the sars-cov-2 lineage b.1.526, 2021. doi: 10.1101/2021.02.14.431043.

193. D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

194. C. O. Wilke. Quasispecies theory in the context of population genetics. *BMC evolutionary biology*, 5(1):1, 2005.

195. C. J. Worby, P. D. O'Neill, T. Kypraios, J. V. Robotham, D. De Angelis, E. J. Cartwright, S. J. Peacock, and B. S. Cooper. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *The annals of applied statistics*, 10(1):395, 2016.

196. C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, C. Fraser, T. M. P. C. STOP-HCV Consortium, and T. B. Collaboration. Phyloscanner: inferring transmission from within-and between-host pathogen genetic diversity. *Molecular biology and evolution*, 35(3):719–733, 2017.

197. P. Yadav, G. Sapkal, P. Abraham, R. Ella, G. Deshpande, D. Patil, D. Nyayanit, N. Gupta, R. Sahay, A. Shete, S. Panda, B. Bhargava, and V. Mohan. Neutralization of variant under investigation B.1.617 with sera of BBv152 vaccinees. *Clin Infect Dis*, ciab411, 2021. doi: 10.1093/cid/ciab411.

198. Z. Yang. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus a. *Journal of molecular evolution*, 51(5):423–432, 2000.

199. M. Yarmarkovich, J. M. Warrington, A. Farrel, and J. M. Maris. Identification of sars-cov-2 vaccine epitopes predicted to induce long-term population-scale immunity. *Cell Reports Medicine*, 1(3):100036, 2020.

200. R. J. Ypma, W. M. van Ballegooijen, and J. Wallinga. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, 195(3):1055–1062, 2013.

201. J. Zahradník, S. Marciano, M. Shemesh, E. Zoler, D. Harari, J. Chiaravalli, B. Meyer, Y. Rudich, C. Li, I. Marton, et al. Sars-cov-2 variant prediction and antiviral drug design are enabled by rbd in vitro evolution. *Nature microbiology*, 6(9):1188–1198, 2021.

202. H.-L. Zeng, V. Dichio, E. R. Horta, K. Thorell, and E. Aurell. Global analysis of more than 50,000 sars-cov-2 genomes reveals epistasis between eight viral genes. *Proceedings of the National Academy of Sciences*, 117(49):31519–31526, 2020.

203. L. Zhang, C. B. Jackson, H. Mou, A. Ojha, E. S. Rangarajan, T. Izard, M. Farzan, and H. Choe. The d614g mutation in the sars-cov-2 spike protein reduces s1 shedding and increases infectivity. *BioRxiv*, 2020.

204. W. Zhang, B. D. Davis, S. S. Chen, J. M. S. Martinez, J. T. Plummer, and E. Vail. Emergence of a novel sars-cov-2 variant in southern california. *Jama*, 325:1324–1326, 2021.

205. Y. Zhang, C. Wymant, O. Laeyendecker, M. K. Grabowski, M. Hall, S. Hudelson, E. Piwowar-Manning, M. McCauley, T. Gamble, M. C. Hosseinipour, et al. Evaluation of phylogenetic methods for inferring the direction of human immunodeficiency virus (hiv) transmission: Hiv prevention trials network (hptn) 052. *Clinical Infectious Diseases*, 2020.

206. J. E. Zibbell, K. Iqbal, R. Patel, A. Suryaprasad, K. Sanders, L. Moore-Moravian, J. Serrecchia, S. Blankenship, J. Ward, and D. Holtzman. Increases in hepatitis c virus infection related to injection drug use among persons aged ¡ 30 years-kentucky, tennessee, virginia, and west virginia, 2006-2012. *MMWR. Morbidity and mortality weekly report*, 64(17):453–458, 2015.

207. N. Zucman, F. Uhel, D. Descamps, D. Roux, and J. Ricard. Severe reinfection with south african sars-cov-2 variant 501y.v2: A case report. *Clinical Infectious Diseases*, ciab129, 2021. doi: 10.1093/cid/ciab129.