Community structure and temporal dynamics of SARS-CoV-2 epistatic network allows for early detection of emerging variants with altered phenotypes

Fatemeh Mohebbi[1], Alex Zelikovsky[1], Serghei Mangul[2], Gerardo Chowell[3], and Pavel Skums[1,*]

[1]Department of Computer Science, Georgia State University, 25 Park Place, Atlanta, GA, USA, 30303
[2]School of Pharmacy, University of Southern California, Los Angeles, CA, USA, 90089
[3]School of Public Health, Georgia State University, 140 Decatur St., Atlanta, GA, USA, 30303
[*]Corresponding author. *Email: pskums@gsu.edu*

Abstract

The rise of viral variants with altered phenotypes presents a significant public health challenge. In particular, the successive waves of COVID-19 have been driven by emerging variants of interest (VOIs) and variants of concern (VOCs), which are linked to modifications in phenotypic traits such as transmissibility, antibody resistance, and immune escape. Consequently, devising effective strategies to forecast emerging viral variants is critical for managing present and future epidemics. Although current evolutionary prediction tools mainly concentrate on single amino acid variants (SAVs) or isolated genomic changes, the observed history of VOCs and the extensive epistatic interactions within the SARS-CoV-2 genome suggest that predicting viral haplotypes, rather than individual mutations, is vital for efficient genomic surveillance. However, haplotype prediction is significantly more challenging problem, which precludes the use of traditional AI and Machine Learning approaches utilized in most mutation-based studies.

This study demonstrates that by examining the community structure of SARS-CoV-2 spike protein epistatic networks, it is feasible to efficiently detect or predict emerging haplotypes with altered transmissibility. These haplotypes can be linked to dense network communities, which become discernible significantly earlier than their associated viral variants reach noticeable prevalence levels. From these insights, we developed HELEN (Heralding Emerging Lineages in Epistatic Networks), a computational framework that identifies densely epistatically connected communities of SAV alleles and merges them into haplotypes using a combination of statistical inference, population genetics, and discrete optimization techniques. HELEN was validated by accurately identifying known SARS-CoV-2 VOCs and VOIs up to 10-12 months before they reached perceptible prevalence and were designated by the WHO. For example, our approach suggests that the spread of the Omicron haplotype or a closely related genomic variant could have been foreseen as early as the start of 2021, almost a year before its WHO designation. Moreover, HELEN offers greater scalability than phylogenetic lineage tracing methods, allowing for the analysis of millions of available SARS-CoV-2 genomes. Besides SARS-CoV-2, our methodology can be employed to detect emerging and circulating strains of any highly mutable pathogen with adequate genomic surveillance data.

Keywords: SARS-CoV-2; genomic surveillance; haplotype forecasting; epistasis; network community

1 Introduction

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 [1, 2, 3, 4].

The most evident example of 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 [5, 6, 7, 8], antibody resistance and immune escape [9, 10, 11, 12]. 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 [13, 14], many of which have been linked to enhanced fitness compared to preceding SARS-CoV-2 lineages [6, 7, 8, 15, 16, 13, 17].

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 [18] to detect spreading lineages 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 [19] 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 [20]. For viruses, the possibility of evolutionary predictions remains a topic of debate [21]. Nevertheless, studies attempting to address the SARS-CoV-2 evolutionary forecasting problem have emerged [3, 4, 22, 23, 24, 25]. 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 [3, 25].

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 [26, 27, 28, 4, 29, 30]. Using various methodologies, including phylogenetic analysis [27, 30], direct coupling analysis [4], and in vitro binding measurements [28], 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 [31, 32, 12]. Epistasis is closely linked to the complex structures of viral fitness landscapes [33, 4, 34, 26], 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. 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 [35].

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 [4].

This paper focuses on predicting haplotypes of SARS-CoV-2 using a novel approach based on analyzing dense communities of the epistatic network of the spike protein. 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 [36, 37, 38, 39, 40]. 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 [37, 41, 42, 43]. Although this approach is not directly transferable to haplotype prediction, it provided a foundation for this study.

2 Methods

2.1 Construction of epistatic networks

Given a multiple sequence alignment consisting of N genomes of length L, we define an epistatic network $\mathcal{G}$ as a graph with nodes representing SAVs are two nodes being connected by an edge 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 [42, 37]. 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 be described by the quasispecies model [44, 45] (or mutation-selection balance model in the classical population genetics terms [46]) in the following form:

$$E^t_{ij} = \sum_{k,l=0,1} f_{kl} q^U_{ki} q^V_{lj} E^{t-1}_{kl} \tag{1}$$

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.

$$q^U_{ij} < q^U_{ii}, q^V_{ij} < q^V_{ii}, i, j = 0, 1. \tag{2}$$

We use the model (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*

$$E^t_{11} \leq \frac{E^t_{01} \cdot E^t_{10}}{E^t_{00}} \tag{3}$$

Proof. The proof follows the same lines as the proof in [42]. Given that $f_{11} = 0$, we have

$$\begin{aligned} E^t_{00} \cdot E^t_{11} &= \left(\sum_{k,l=0,1} f_{kl} q^U_{k0} q^V_{l0} E^{t-1}_{kl} \right) \left(\sum_{k,l=0,1} f_{kl} q^U_{k1} q^V_{l1} E^{t-1}_{kl} \right) \\ = q^U_{00} q^V_{00} q^U_{01} q^V_{01} (f_{00} E^{t-1}_{00})^2 &+ q^U_{10} q^V_{00} q^U_{11} q^V_{01} (f_{10} E^{t-1}_{10})^2 + q^U_{00} q^V_{10} q^U_{01} q^V_{11} (f_{01} E^{t-1}_{01})^2 + \\ &+ (q^U_{00} q^V_{00} q^U_{01} q^V_{11} + q^U_{00} q^V_{10} q^U_{01} q^V_{01}) f_{00} f_{01} E^{t-1}_{00} E^{t-1}_{01} + \\ &+ (q^U_{00} q^V_{00} q^U_{11} q^V_{01} + q^U_{10} q^V_{00} q^U_{01} q^V_{01}) f_{00} f_{10} E^{t-1}_{00} E^{t-1}_{10} + \\ &+ (q^U_{00} q^V_{10} q^U_{11} q^V_{01} + q^U_{10} q^V_{00} q^U_{01} q^V_{11}) f_{01} f_{10} E^{t-1}_{01} E^{t-1}_{10} \end{aligned} \tag{4}$$

and

$$
\begin{aligned}
E^t_{01} \cdot E^t_{10} = \left(\sum_{k,l=0,1} f_{kl} q^U_{k0} q^V_{l1} E^{t-1}_{kl} \right) \left(\sum_{k,l=0,1} f_{kl} q^U_{k1} q^V_{l0} E^{t-1}_{kl} \right) \\
= q^U_{00} q^V_{01} q^U_{01} q^V_{00} (f_{00} E^{t-1}_{00})^2 + q^U_{10} q^V_{01} q^U_{11} q^V_{00} (f_{10} E^{t-1}_{10})^2 + q^U_{00} q^V_{11} q^U_{01} q^V_{10} (f_{01} E^{t-1}_{01})^2 + \\
+ (q^U_{00} q^V_{01} q^U_{01} q^V_{10} + q^U_{00} q^V_{11} q^U_{01} q^V_{00}) f_{00} f_{01} E^{t-1}_{00} E^{t-1}_{01} + \\
+ (q^U_{00} q^V_{01} q^U_{11} q^V_{00} + q^U_{10} q^V_{01} q^U_{01} q^V_{00}) f_{00} f_{10} E^{t-1}_{00} E^{t-1}_{10} + \\
+ (q^U_{00} q^V_{11} q^U_{11} q^V_{00} + q^U_{10} q^V_{01} q^U_{01} q^V_{10}) f_{01} f_{10} E^{t-1}_{01} E^{t-1}_{10}
\end{aligned} \tag{5}
$$

It is easy to see that the terms in (4) and (5) except for the last ones are equal. Thus we have

$$
\begin{aligned}
E^t_{01} \cdot E^t_{10} - E^t_{00} \cdot E^t_{11} = \\
= (q^U_{00} q^V_{11} q^U_{11} q^V_{00} + q^U_{10} q^V_{01} q^U_{01} q^V_{10} - q^U_{00} q^V_{10} q^U_{11} q^V_{01} - q^U_{10} q^V_{00} q^U_{01} q^V_{11}) f_{01} f_{10} E^{t-1}_{01} E^{t-1}_{10} = \\
= \left(1 - \frac{q^U_{01} q^U_{10}}{q^U_{00} q^U_{11}}\right) \left(1 - \frac{q^V_{01} q^V_{10}}{q^V_{00} q^V_{11}}\right) q^U_{00} q^V_{11} q^U_{11} q^V_{00} f_{01} f_{10} E^{t-1}_{01} E^{t-1}_{10} \geq 0,
\end{aligned} \tag{6}
$$

where the last inequality follows from (2). Thus, the inequality (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), 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

$$
p(X \geq O_{11} | f_{11} = 0) = 1 - F_X(O_{11} - 1) = 1 - \sum_{i=0}^{O_{11}-1} \binom{N}{i} p^i (1-p)^{N-i}, \tag{7}
$$

where F_X is the cumulative distribution function of the bionomial distribution. We assume that SAVs U_1 and V_1 are *linked* (i.e. adjacent in the epistatic network $\mathcal{G}$), when the probability (7) is low enough, i.e.

$$
p(X \geq O_{11} | f_{11} = 0) \leq \frac{\rho}{\binom{L}{2}}, \tag{8}
$$

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.

2.2 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 (8) to construct temporal epistatic 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.

$$d_{\mathcal{G}_t}(H) = \frac{|E(\mathcal{G}_t[H])|}{|H|} \tag{9}$$

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 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 the densities higher than that of H. However, SARS-CoV-2 temporal epistatic interaction networks are relatively sparse, and thus many sampled subgraphs will be a priori disconneced and, consequently, also sparse. As a result, such 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 more sophisticated randomized enumeration subgraph sampling scheme based on the network motif sampling algorithm introduced in [47]. This scheme uniformly samples only connected subgraphs and can be described as follows. Let us assume that all vertices of $\mathcal{G}_t$ are labelled 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 allows to avoid double-sampling [47]. 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. The entire sampling scheme is described by Algorithm 1.

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

$$p_{\mathcal{G}_t}(H) = \frac{|\{S_j \in \mathcal{S}^* : d_{\mathcal{G}_t}(S_j) \geq d_{\mathcal{G}_t}(H)|}{|\mathcal{S}^*|} \tag{10}$$

If, at some point, the subgraph induced by H is disconnected, we replace H with its largest connected component. For each analyzed spike epistatic 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.

2.3 Inference of viral haplotypes as dense communities in epistatic networks

We propose to infer viral haplotypes as dense communities of epistatic networks. Community detection is a well-established field of network science, with numerous algorithmic solutions proposed over the last two decades [48, 49, 50]. Typically (though not always), the collection of communities in a network is defined as a partition [51]. 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 epistatic networks, although containing

Algorithm 1 Sampling of connected k-subgraphs without forbidden pairs

Input: graphs $\mathcal{G}$, $\mathcal{F}$, an integer k and the sample size per vertex M.
for $v \in V(\mathcal{G})$ **do**
    $S \leftarrow \{v\}$; $W \leftarrow N_{\mathcal{G}}(v) \setminus (N_{\mathcal{F}}(v) \cup \{1, ..., v\})$; $X \leftarrow N_{\mathcal{G}}(v) \cup N_{\mathcal{F}}(v) \cup \{1, ..., v\}$
    global $M_v = 0$
    call `SampleSubgraph`(v,S,W,X)
end for

`SampleSubgraph`(S,W,X,v)
if $|S| = k$ **then**
    output S, $M_v \leftarrow M_v + 1$ and **return**
end if
while $W \neq \emptyset$ and $M_v \leq M$ **do**
    sample a random vertex $w \in W$ and set $W \leftarrow W \setminus \{w\}$
    $S' \leftarrow S \cup \{w\}$; $W' \leftarrow (W \cup (N_{\mathcal{G}}(v) \setminus X)) \setminus N_{\mathcal{F}}(w)$; $X' \leftarrow X \cup N_{\mathcal{G}}(w) \cup N_{\mathcal{F}}(w)$
    call `SampleSubgraph`(v,S',W',X')
end while

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 [37]. 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 [52] to find the densest subgraphs of epistatic interaction 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:

$$\sum_{ij \in E(\mathcal{G}_t)} y_{ij} \to \max \tag{11}$$

$$y_{ij} \leq x_i, \quad y_{ij} \leq x_j, \quad ij \in E(\mathcal{G}_t) \tag{12}$$

$$\sum_{i \in V(\mathcal{G}_t)} x_i \leq 1 \tag{13}$$

$$x_i, y_{ij} \geq 0, \quad i \in V(\mathcal{G}_t), ij \in E(\mathcal{G}_t) \tag{14}$$

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 (11)-(14) and the maximum subgraph density of $\mathcal{G}_t$ coincide [52]; 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 (11)-(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. The approach to producing 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

[53, 54]. Therefore, for each value of k, we use the following Integer Linear Programming formulation:

$$\frac{1}{k} \sum_{ij \in E(\mathcal{G}_t)} y_{ij} \to \max \tag{15}$$

$$y_{ij} \leq x_i, \quad y_{ij} \leq x_j, \quad ij \in E(\mathcal{G}_t) \tag{16}$$

$$\sum_{i \in V(\mathcal{G}_t)} x_i = k \tag{17}$$

$$\sum_{i \in V(\mathcal{G}_t) \setminus S} x_i \geq 1, \quad S \in \bigcup_{k'=k+1}^{k_{\max}} \mathcal{S}_{k'} \tag{18}$$

$$x_i, y_{ij} \in \{0, 1\}, \quad i \in V(\mathcal{G}_t), ij \in E(\mathcal{G}_t) \tag{19}$$

The problems (11)-(14) and (15)-(19) are solved using Gurobi [55]; 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 $\hat{\mathcal{S}}t = St, 1, ..., S_{t,|\hat{\mathcal{S}}t|}$ 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 1 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 $|S_{t,i} \cap S_{t,j}| \geq \min\{|S_{t,i}|, |S_{t,j}|\} - \lambda$ (by default $\lambda = 1$).

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 [56], 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 [57]. 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 [57] and whose existence is guaranteed by Menger's theorem [58]). 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 [59]; second, cluster the obtained embedded graph by a spectral clustering algorithm [60] 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 correspond 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 [61].

3.2) Find the densest subgraph $H_{t,i}$ of $\mathcal{G}_{t,i}[C_i]$ using the LP formulation (11)-(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 2.1-2.3 is called HELEN (Heralding Emerging Lineages in Epistatic Networks).

3 Results

3.1 Data

Genomic data and associated metadata analyzed in this study was obtained from GISAID [19]. 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 [62]. We extracted the spike protein alignment from the whole genome multiple sequence alignment, replacing ambiguous characters with gaps, and 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 [63]: 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 epistatically linked pairs of SAVs and dense communities in epistatic networks is affected by the number of sequences. Thus we focused on data from countries with the

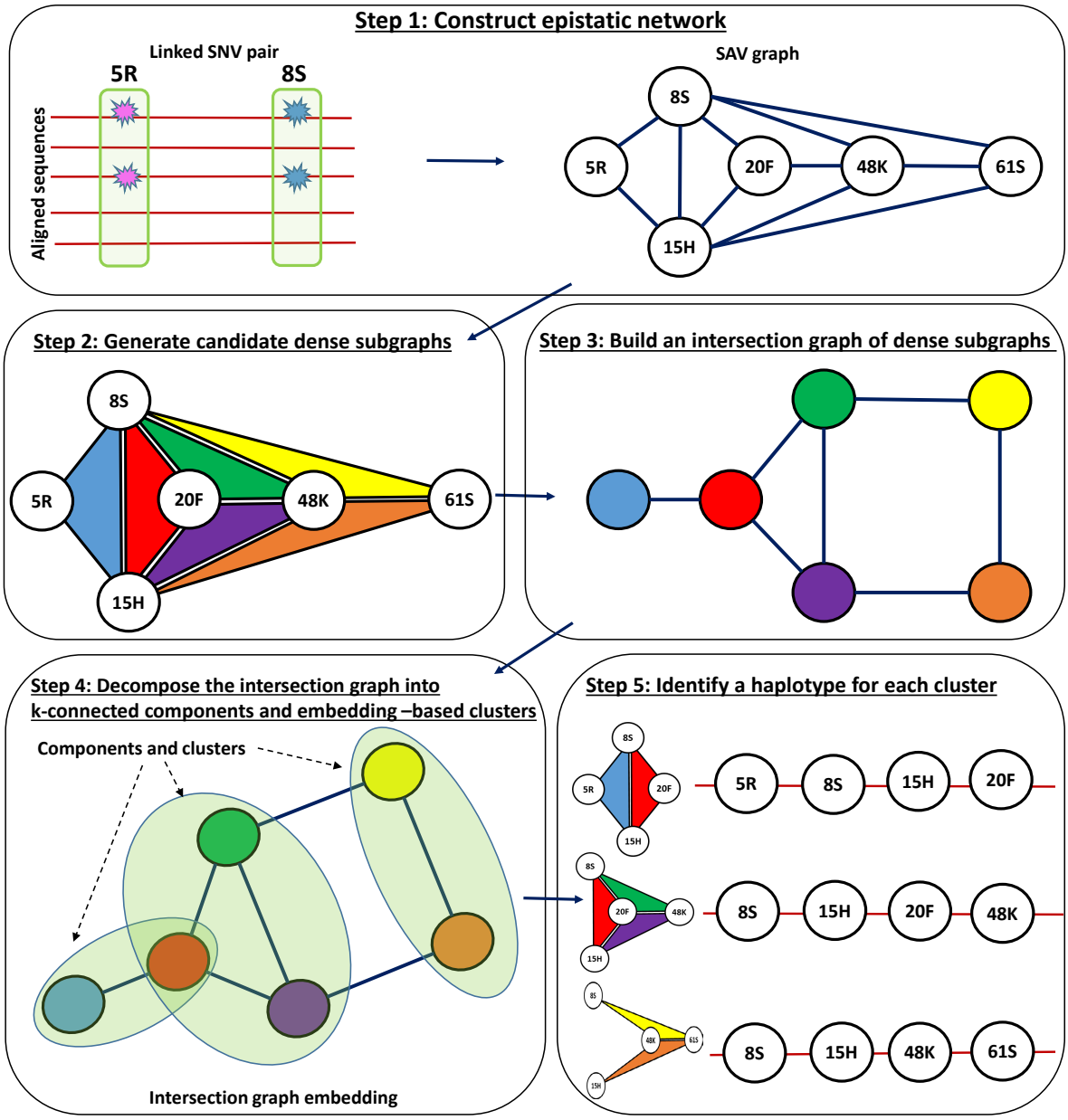

Figure 1: HELEN (Heralding Emerging Lineages in Epistatic Networks): a computational framework for inference of viral variants as dense communities in epistatic networks

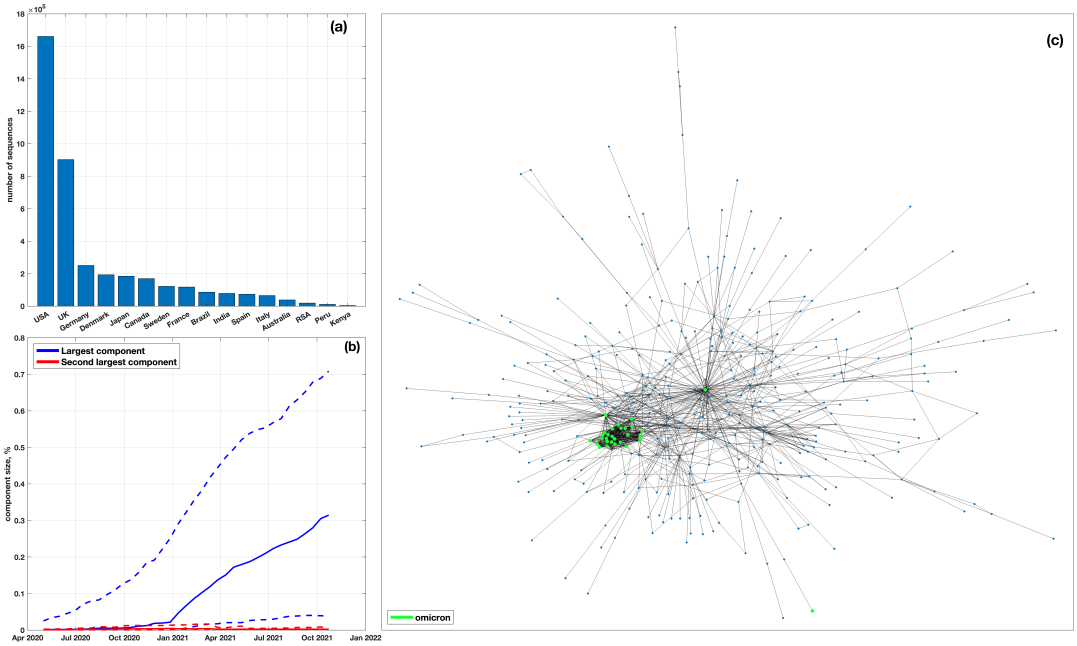

Figure 2: (a) Numbers of analyzed spike amino acid sequences per country. (b) Relative sizes of the largest and second largest connected components of epistatic interaction networks over time. Solid and dashed lines depict median and maximum/minimum values over 16 countries at each time point, respectively. (c) An example of a giant component of an epistatic 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.

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: 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. Fig. 2a 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 The structure of S-gene epistatic networks

We utilized the method outlined in Subsection 2.1 to construct epistatic 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 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 (Fig. 2b). Most of these smaller components consisted of isolated vertices.

Epistatic networks of the S-gene have a tendency 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 epistatic networks inferred in this study have a sufficiently rich community structure [64, 65] that can be analyzed and utilized to evaluate and forecast the SARS-CoV-2 evolutionary dynamics.

3.3 Dense communities in S-gene epistatic networks as indicators of variant emergence

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

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

(H2) conversely, dense communities within temporal epistatic 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 epistatic 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% ($b = prev$, similar benchmark was used in [3]) [1]. The value of $FD^b(h)$ can be positive or negative, thus indicating early or late prediction, respectively.

It's worth noting that the presence of a viral variant as a dense community within an epistatic network 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 epistatically 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 epistatic links.

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

3.3.1 VOCs/VOIs as communities in epistatic networks

To validate the hypotheses H1) and 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 2.2. 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 Benjamini-Hochberg procedure [66]. The resulting time series of adjusted p-values are illustrated in Fig. 3A and Supplemental Figs. S1-S10.

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 [67] 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 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 epistatic network's giant component. In our analysis, 52% of VOCs/VOIs, analyzed separately for different countries, became significantly dense at some moment of time. This percentage increased to 76% when only considering VOCs. Moreover, these variants were identified as significantly dense at low cumulative frequencies (median value $\mu = 4.2 \cdot 10^{-4}$, Fig. 3d) and low prevalences ($\mu = 1.4 \cdot 10^{-3}$, Fig. 3e).

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 (Fig. 3a); the same applies to FD^{des} with the exception of the theta variant (Fig. 3b).

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

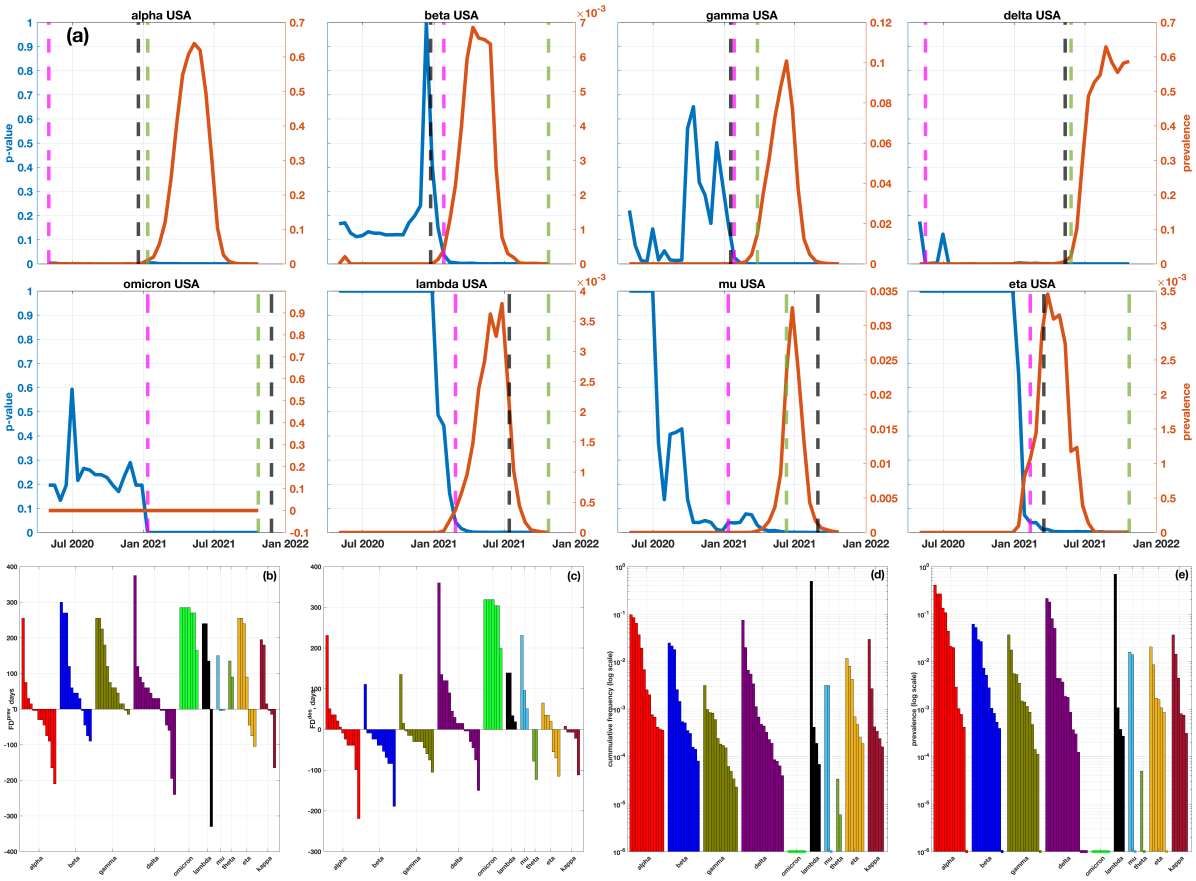

Figure 3: **Density-based adjusted p-values of VOCs/VOIs.** (a) p-values (blue) and prevalences (red) of 8 VOCs and VOIs in the USA epistatic networks (refer to the Supplement for plots of all VOCs/VOIs across all countries). Black, green, and magenta lines represent the times of VOC designation, achieving 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.

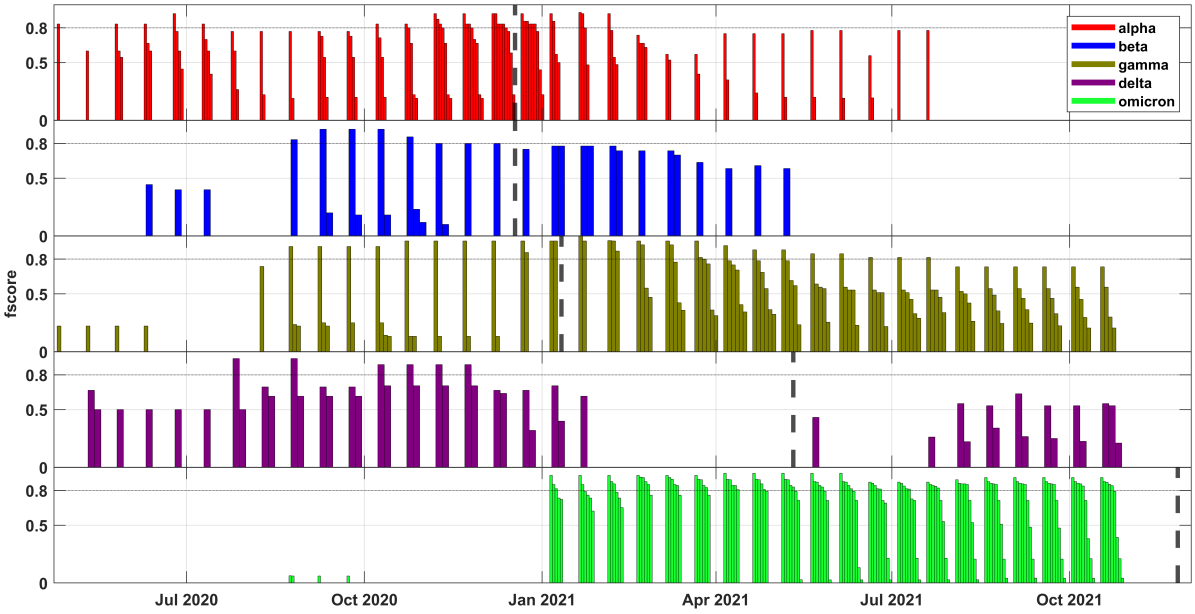

Figure 4: **Comparison between VOCs and densest subnetworks of temporal epistatic networks** (results for individual countries are shown in Fig. S11-S13). Each bar plot depicts the comparison results for a particular VOC; at each time point, bars correspond to the densest subgraphs from different countries 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.

3.3.2 Inference of viral variants as dense communities in epistatic networks

The most straightforward way to partially assess the validity of hypotheses H2) and H3) is to retrieve the densest subnetworks of epistatic 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 2.3). We used the f-score as a metric for detection accuracy, which in our context is defined as:

$$R_{t,i} = \frac{|C_t \cap V_i|}{|V_i|}, \quad P_{t,i} = \frac{|C_t \cap V_i|}{|C_t|}, \quad F_{t,i} = 2\frac{R_{t,i} \cdot P_{t,i}}{R_{t,i} + P_{t,i}} \tag{20}$$

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 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 (Fig. 4 and Fig. S14). Furthermore, these communities emerged when the corresponding VOC haplotypes had low cumulative frequencies (median value $\mu_f = 3.8 \cdot 10^{-4}$, Fig. S14c) and low prevalences (median value $\mu_p = 7.8 \cdot 10^{-4}$, Fig. S14d). 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 $FD^{\text{prev}} = 60$, Fig. 4 and Fig. S11-S14). 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) and H3), a more advanced algorithmic approach is essential for a comprehensive forecasting frame-

work, 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 epistatic 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 epistatic network communities (Subsection 2.3). Briefly, the algorithm generates a pool of distinct dense subnetworks of varied sizes, partitions them into clusters, and assemble 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 (Fig. 5). 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 $\mu_f = 4.8 \cdot 10^{-4}$ (Fig. 5d) and the median variant prevalence was $\mu_d = 2.2 \cdot 10^{-3}$, Fig. 5e).

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 (Fig. 5b,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 (Fig. S15-S20).

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 [3]. A variant

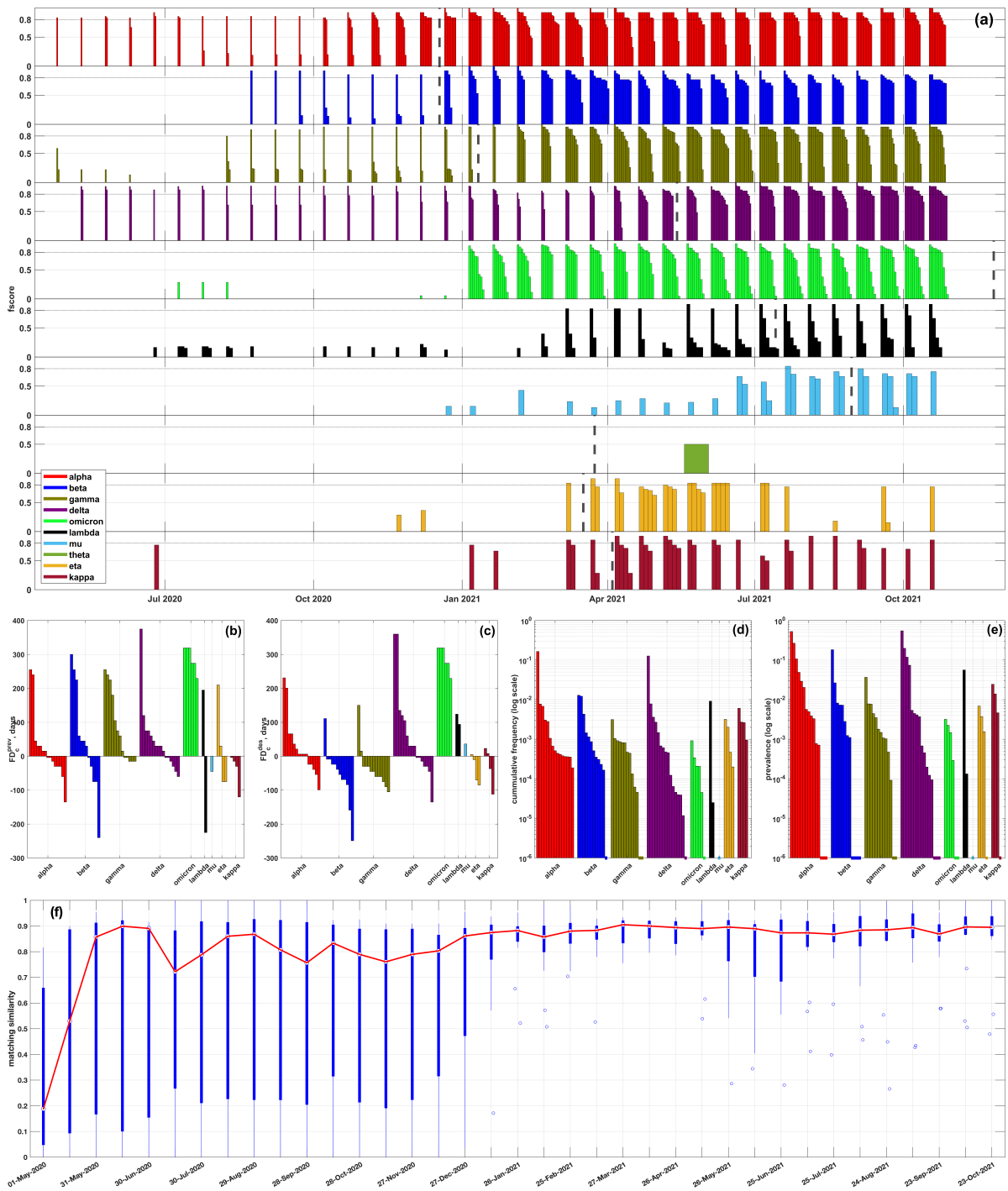

Figure 5: **(a) Summary of comparison between VOCs/VOIs and inferred haplotypes** (results for individual countries are shown on Fig. S15-S20). Each bar plot depicts the comparison results for a particular VOC/VOI; at each time point, bars correspond to inferred haplotypes from different countries closest 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 cummulative frequencies or prevalences 0. **(f) 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.

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:

$$A_{I \to S} = \sum_{i \in I} \sigma_i \max_{s \in S} f_{i,s} \tag{21}$$

A similar measure, in the reverse form of a *matching error*, was used, e.g., in [37].

The summary statistics for matching similarity at each time point across different countries is summarized in Fig. 5f. 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 epistatic networks in different countries, whereas the precision increases as more epistatically linked SAVs are identified.

3.4 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 epistatic 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.

4 Discussion.

This study explores the hypothesis that viral variants with higher transmissibility can be associated with dense communities in epistatic 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 epistatic 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 [42, 37] 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 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. 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.

The utilized epistasis model is another limitation of this study as it only considers the epistatic interactions between SAV pairs (*"second-order epistasis"*). Although combinations of mutations can have more complex and non-linear fitness effects involving higher orders of epistasis, this model is justifiable for several 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 [3] 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.

Finally, 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.

5 Data and materials availability

Code developed and used in this study, as well as generated secondary data, are available at `https://github.com/compbel/HELEN` (DOI: 10.5281/zenodo.7768736).

6 Acknowledgements

PS was supported by NSF grants CCF-2047828 and IIS-2212508.

We acknowledge all researchers and laboratories who contributed to GISAID genomic data used in this study. The full GISAID acknowledgement list can be found at the tool repository `https://github.com/compbel/HELEN`

References

[1] Michael Lässig, Ville Mustonen, and Aleksandra M Walczak. Predicting evolution. *Nature ecology & evolution*, 1(3):1–9, 2017.

[2] Pelin B Icer Baykal, James Lara, Yury Khudyakov, Alex Zelikovsky, and Pavel Skums. Quantitative differences between intra-host hcv populations from persons with recently established and persistent infections. *Virus evolution*, 7(1):veaa103, 2021.

[3] M Cyrus Maher, Istvan Bartha, Steven Weaver, Julia Di Iulio, Elena Ferri, Leah Soriaga, Florian A Lempp, Brian L Hie, Bryan Bryson, Bonnie Berger, et al. Predicting the mutational drivers of future sars-cov-2 variants of concern. *Science translational medicine*, 14(633):eabk3445, 2022.

[4] Juan Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, and Martin Weigt. Epistatic models predict mutable sites in sars-cov-2 proteins and epitopes. *Proceedings of the National Academy of Sciences*, 119(4):e2113118119, 2022.

[5] Nicholas G Davies, Sam Abbott, Rosanna C Barnard, Christopher I Jarvis, Adam J Kucharski, James D Munday, Carl AB Pearson, Timothy W Russell, Damien C Tully, Alex D Washburne, et al. Estimated transmissibility and impact of sars-cov-2 lineage b. 1.1. 7 in england. *Science*, 372(6538):eabg3055, 2021.

[6] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra 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.

[7] Markus Hoffmann, Hannah Kleine-Weber, and Stefan Pöhlmann. A multibasic cleavage site in the spike protein of sars-cov-2 is essential for infection of human lung cells. *Molecular Cell*, 2020.

[8] Lizhou Zhang, Cody B Jackson, Huihui Mou, Amrita Ojha, Erumbi S Rangarajan, Tina Izard, Michael Farzan, and Hyeryun Choe. The d614g mutation in the sars-cov-2 spike protein reduces s1 shedding and increases infectivity. *BioRxiv*, 2020.

[9] Pengfei Wang, Manoj S Nair, Lihong Liu, Sho Iketani, Yang Luo, Yicheng Guo, Maple Wang, Jian Yu, Baoshan Zhang, Peter 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.

[10] Markus Hoffmann, Prerna Arora, Rüdiger Groß, Alina Seidel, Bojan F Hörnich, Alexander S Hahn, Nadine Krüger, Luise Graichen, Heike Hofmann-Winkler, Amy Kempf, et al. Sars-cov-2 variants b. 1.351 and p. 1 escape from neutralizing antibodies. *Cell*, 184(9):2384–2393, 2021.

[11] Wilfredo F Garcia-Beltran, Evan C Lam, Kerri St Denis, Adam D Nitido, Zeidy H Garcia, Blake M Hauser, Jared Feldman, Maia N Pavlovic, David J Gregory, Mark C Poznansky, et al. Multiple sars-cov-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell*, 184(9):2372–2383, 2021.

[12] Delphine Planas, David Veyer, Artem Baidaliuk, Isabelle Staropoli, Florence Guivel-Benhassine, Maaran Michael Rajah, Cyril Planchais, Françoise Porrot, Nicolas Robillard, Julien Puech, et al. Reduced sensitivity of sars-cov-2 variant delta to antibody neutralization. *Nature*, 596(7871):276–280, 2021.

[13] Andrew Rambaut, Nick Loman, Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli, Tom Connor, Tom Peacock, David L Robertson, and Erik Volz. Preliminary genomic characterisation of an emergent sars-cov-2 lineage in the uk defined by a novel set of spike mutations. `https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563`. Accessed: 2020-12-25.

[14] Laxmi Kirola. Genetic emergence of b. 1.617. 2 in covid-19. *New Microbes and New Infections*, 43:100929, 2021.

[15] Erik Volz, Verity Hill, John T McCrone, Anna Price, David Jorgensen, Áine O'Toole, Joel Southgate, Robert Johnson, Ben Jackson, Fabricia F Nascimento, et al. Evaluating the effects of sars-cov-2 spike mutation d614g on transmissibility and pathogenicity. *Cell*, 184(1):64–75, 2021.

[16] Yang Liu, Jianying Liu, Bryan A Johnson, Hongjie Xia, Zhiqiang Ku, Craig Schindewolf, Steven G Widen, Zhiqiang An, Scott C Weaver, Vineet D Menachery, et al. Delta spike p681r mutation enhances sars-cov-2 fitness over alpha variant. *bioRxiv*, 2021.

[17] Kathy Leung, Marcus HM Shum, Gabriel M Leung, Tommy TY Lam, and Joseph T Wu. Early empirical assessment of the n501y mutant strains of sars-cov-2 in the united kingdom, october to november 2020. *medRxiv*, 2020.

[18] Sergey Knyazev, Karishma Chhugani, Varuni Sarwal, Ram Ayyala, Harman Singh, Smruthi Karthikeyan, Dhrithi Deshpande, Pelin Icer Baykal, Zoia Comarova, Angela Lu, et al. Unlocking capacities of genomics for the covid-19 response and future pandemics. *Nature Methods*, 19(4):374–380, 2022.

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

[20] Stephen Jay Gould. *Wonderful life: the Burgess Shale and the nature of history*. WW Norton & Company, 1990.

[21] Pierre Barrat-Charlaix, John Huddleston, Trevor Bedford, and Richard A Neher. Limited predictability of amino acid substitutions in seasonal influenza viruses. *Molecular Biology and Evolution*, 38(7):2767–2777, 2021.

[22] Chen Bai, Junlin Wang, Geng Chen, Honghui Zhang, Ke An, Peiyi Xu, Yang Du, Richard D Ye, Arjun Saha, Aoxuan 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.

[23] Syed Faraz Ahmed, Ahmed A Quadeer, and Matthew 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.

[24] Mark Yarmarkovich, John M Warrington, Alvin Farrel, and John M Maris. Identification of sars-cov-2 vaccine epitopes predicted to induce long-term population-scale immunity. *Cell Reports Medicine*, 1(3):100036, 2020.

[25] Fritz Obermeyer, Martin Jankowiak, Nikolaos Barkas, Stephen F Schaffner, Jesse D Pyle, Leonid Yurkovetskiy, Matteo Bosso, Daniel J Park, Mehrtash Babadi, Bronwyn L MacInnis, et al. Analysis of 6.4 million sars-cov-2 genomes identifies mutations associated with fitness. *Science*, 376(6599):1327–1332, 2022.

[26] Hong-Li Zeng, Vito Dichio, Edwin Rodríguez Horta, Kaisa Thorell, and Erik 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.

[27] Nash D Rochman, Yuri I Wolf, Guilhem Faure, Pascal Mutz, Feng Zhang, and Eugene V Koonin. Ongoing global and regional adaptive evolution of sars-cov-2. *Proceedings of the National Academy of Sciences*, 118(29), 2021.

[28] Jiří Zahradník, Shir Marciano, Maya Shemesh, Eyal Zoler, Daniel Harari, Jeanne Chiaravalli, Björn Meyer, Yinon Rudich, Chunlin Li, Ira 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.

[29] Nash D Rochman, Guilhem Faure, Yuri I Wolf, Peter L Freddolino, Feng Zhang, and Eugene 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.

[30] Alexey Dmitrievich Neverov, Gennady Fedonin, Anfisa Popova, Daria Bykova, and Georgii Bazykin. Coordinated evolution at amino acid sites of sars-cov-2 spike. *Elife*, 12:e82516, 2023.

[31] Fedaa Ali, Amal Kasry, and Muhamed 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.

[32] Binquan Luan, Haoran Wang, and Tien 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.

[33] Niko Beerenwinkel, Lior Pachter, and Bernd Sturmfels. Epistasis and shapes of fitness landscapes. *Statistica Sinica*, pages 1317–1342, 2007.

[34] Alief Moulana, Thomas Dupic, Angela M Phillips, Jeffrey Chang, Serafina Nieves, Anne A Roffler, Allison J Greaney, Tyler N Starr, Jesse D Bloom, and Michael M Desai. Compensatory epistasis maintains ace2 affinity in sars-cov-2 omicron ba. 1. *Nature Communications*, 13(1):7011, 2022.

[35] World Health 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.

[36] Sergey Knyazev, Lauren Hughes, Pavel Skums, and Alexander Zelikovsky. Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. *Briefings in Bioinformatics*, 2020.

[37] Sergey Knyazev, Viachaslau Tsyvina, Anupama Shankar, Andrew Melnyk, Alexander Artyomenko, Tatiana Malygina, Yuri B Porozov, Ellsworth M Campbell, William M Switzer, Pavel 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.

[38] Yunxi Liu, Joshua Kearney, Medhat Mahmoud, Bryce Kille, Fritz J Sedlazeck, and Todd J Treangen. Rescuing low frequency variants within intra-host viral populations directly from oxford nanopore sequencing data. *Nature communications*, 13(1):1–9, 2022.

[39] Dehan Cai and Yanni Sun. Reconstructing viral haplotypes using long reads. *Bioinformatics*, 38(8):2127–2134, 2022.

[40] Xiaoli Jiao, Hiromi Imamichi, Brad T Sherman, Rishub Nahar, Robin L Dewar, H Clifford Lane, Tomozumi Imamichi, and Weizhong Chang. Quasiseq: profiling viral quasispecies via self-tuning spectral clustering with pacbio long sequencing reads. *Bioinformatics*, 2022.

[41] Alexander Artyomenko, Nicholas Mancuso, Alex Zelikovsky, Pavel Skums, and Ion Măndoiu. kgem: An em-based algorithm for local reconstruction of viral quasispecies. In *2013 IEEE 3rd International Conference on Computational Advances in Bio and medical Sciences (ICCABS)*, pages 1–1. IEEE, 2013.

[42] Alexander Artyomenko, Nicholas C Wu, Serghei Mangul, Eleazar Eskin, Ren Sun, and Alex 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.

[43] Andrew Melnyk, Fatemeh Mohebbi, Sergey Knyazev, Bikram Sahoo, Roya Hosseini, Pavel Skums, Alex Zelikovsky, and Murray Patterson. From alpha to zeta: Identifying variants and subtypes of sars-cov-2 via clustering. *Journal of Computational Biology*, 28(11):1113–1129, 2021.

[44] Martin A Nowak. *Evolutionary dynamics: exploring the equations of life*. Harvard university press, 2006.

[45] Manfred Eigen, John McCaskill, and Peter Schuster. Molecular quasi-species. *The Journal of Physical Chemistry*, 92(24):6881–6891, 1988.

[46] Claus O Wilke. Quasispecies theory in the context of population genetics. *BMC evolutionary biology*, 5(1):1–8, 2005.

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

[48] Hocine Cherifi, Gergely Palla, Boleslaw K Szymanski, and Xiaoyan Lu. On community structure in complex networks: challenges and opportunities. *Applied Network Science*, 4(1):1–35, 2019.

[49] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical review E*, 74(1):016110, 2006.

[50] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[51] Vinícius da Fonseca Vieira, Carolina Ribeiro Xavier, and Alexandre Gonçalves Evsukoff. A comparative study of overlapping community detection methods from the perspective of the structural properties. *Applied Network Science*, 5(1):1–42, 2020.

[52] Moses 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.

[53] Uriel Feige, David Peleg, and Guy Kortsarz. The dense k-subgraph problem. *Algorithmica*, 29(3):410–421, 2001.

[54] Yuichi Asahiro, Refael Hassin, and Kazuo Iwama. Complexity of finding dense subgraphs. *Discrete Applied Mathematics*, 121(1-3):15–26, 2002.

[55] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2022.

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

[57] Shimon Even and R Endre Tarjan. Network flow and testing graph connectivity. *SIAM journal on computing*, 4(4):507–518, 1975.

[58] Douglas Brent West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

[59] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[60] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[61] Robert Tibshirani, Guenther Walther, and Trevor 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.

[62] Áine O'Toole, Oliver G Pybus, Michael E Abram, Elizabeth J Kelly, and Andrew Rambaut. Pango lineage designation and assignment using sars-cov-2 spike gene nucleotide sequences. *BMC genomics*, 23(1):1–13, 2022.

[63] The World Health Organization. Tracking SARS-CoV-2 variants, 2022.

[64] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *nature*, 435(7043):814–818, 2005.

[65] Pavel Skums and Leonid Bunimovich. Graph fractal dimension and the structure of fractal networks. *Journal of Complex Networks*, 8(4):cnaa037, 2020.

[66] Yoav Benjamini and Yosef 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.

[67] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.