Temporal epistasis inference from more than 3 500 000 SARS-CoV-2 genomic sequences

Hong-Li Zeng, Yue Liu, Vito Dichio, and Erik Aurell

Phys. Rev. E **106**, 044409 — Published 17 October 2022

DOI: 10.1103/PhysRevE.106.044409

Temporal epistasis inference from more than 3,500,000 SARS-CoV-2 Genomic Sequences

Hong-Li Zeng,[1,*] Yue Liu,[1] Vito Dichio,[2] and Erik Aurell[3,†]

[1] *School of Science, Nanjing University of Posts and Telecommunications, New Energy Technology Engineering Laboratory of Jiangsu Province, Nanjing, 210023, China*
[2] *Inria Paris, Aramis Project Team, Paris, France Institut du Cerveau, ICM, Inserm U 1127, CNRS UMR 7225, Sorbonne Universit, Paris, France*
[3] *Department of Computational Science and Technology, AlbaNova University Center, SE-106 91 Stockholm, Sweden*

(Dated: September 30, 2022)

We use Direct Coupling Analysis (DCA) to determine epistatic interactions between loci of variability of the SARS-CoV-2 virus, segmenting genomes by month of sampling. We use full-length, high-quality genomes from the GISAID repository up to October 2021, in total over 3,500,000 genomes. We find that DCA terms are more stable over time than correlations, but nevertheless change over time as mutations disappear from the global population or reach fixation. Correlations are enriched for phylogenetic effects, and in particularly statistical dependencies at short genomic distances, while DCA brings out links at longer genomic distance. We discuss the validity of a DCA analysis under these conditions in terms of a transient Quasi-Linkage Equilibrium state. We identify putative epistatic interaction mutations involving loci in Spike.

keywords: SARS-CoV-2 | Temporal Epistasis Inference | Genomic Data | Direct Coupling Analysis

I. INTRODUCTION

The global pandemic of the disease COVID-19 caused by coronavirus SARS-CoV-2 has led to more than 590 million confirmed cases and more than 6.4 million deaths [1]. Efforts to counter the epidemic have include extensive use of Non-Pharmaceutical Interventions (NPI) [2–6], and the development of more than ten widely used vaccines [1, 7, 8]. Although effective against severe disease manifestations, these have not stopped the ongoing spread of COVID19 which has likely become an endemic disease. While drugs such as dexamethasone which lower the fatality rate are now also in wide use, effective anti-viral drugs which would open an another frontline in the pandemic are so far lacking [9, 10].

Both vaccines and antiviral drugs are based on an understanding of the biology of the pathogen, its strengths and potential weaknesses [11–13]. The COVID pandemic is the first to have occurred after massive DNA sequencing became a commodity service. The number of SARS-CoV-2 genomes publicly available in data repositories is many orders of magnitude larger than ever seen in the past. While disparities in sampling and other sources of bias are serious issues [14], such large amounts of data should nevertheless be marshaled in support of the common good to the fullest extent possible. In this work we have relied on a full-length high-quality SARS-CoV-2 genome sequences from the GISAID repository [15] with sampling date up to October 2021: in all more than three and half million viral genomes. These were the virtually exact genetic blueprints of actual viruses infecting actual persons in more than one percent of the confirmed cases world-wide up to the cut-off date. That such quasi-real-time monitoring is at all possible and is a staggering achievement. We are likely only at the beginning of the process of understanding what information that can be unlocked from such vast yet extremely rich and precise data [16].

Among the most remarkable features of such datasets is the possibility to observe in real time the evolutionary process acting at a population level. In classical population genetics, evolution is driven by four main forces in many aspects analogous to mechanisms of statistical physics [17, 18]. Mutation is the change of a single genome due to a chance event and can be assimilated to thermal noise. Natural selection is the propensity of more fit individuals to have more offspring, and acts as an energy term. Recombination (sex) leads to offspring shared between two individuals and acts similarly to pairwise collisions: the earlier genomes are substituted by partly random new combinations and the distribution over genomes relaxes. All the above can be understood on the level of expected (mean) changes, which are exact descriptions in an infinite population. Genetic drift describes the random fluctuations in the numbers of gene variants in a finite population.

The focus on this work is on epistasis, synergistic or antagonistic contributions to fitness from allele variations at two or more loci. Long-standing theoretical arguments predict that the distribution of genotypes in a population directly reflect such multi-loci fitness terms when recombination is the dominant force of evolution [19–21]. Coronaviruses in general exhibit recombination due to their mode of RNA replication [22–25], and recombination has been observed between different strains of SARS-CoV-2 co-infecting the same human host [26–30]. Partly

* hlzeng@njupt.edu.cn
† eaurell@kth.se

conflicting reports have appeared in the literature as to the impact of recombination on the total SARS-CoV-2 population [31, 32]. In this context, recent theoretical advances have shown similar correspondences also when mutation is the dominant force of evolution, provided some recombination is present [33]. If both mutation and recombination are slower (weaker) processes than selection, there is on the other hand no simple relation between epistatic contributions to fitness and variability in the population [18, 34, 35].The approach taken here then does not apply. In this work we assume that this scenario does not pertain.

In an earlier contribution we inferred epistatic interactions from about 50,000 SARS-CoV-2 genome sequences deposited in GISAID until August 2020 [36]. A slightly later contribution from another group used about 130,000 sequences available until October 19, 2020, and reached largely consistent results [37]. In both analyses the mechanism behind linkage disequilibrium (LD) was separated as due to epistasis (the objective of this study) and LD due to phylogeny (a confounder). In [36] interactions imputed to phylogeny were separated out by a randomized null model procedure [38]. The authors of [37] on the other hand leveraged GISAID metadata (sample geographic position) and assignment of samples to clades. In this work we used almost two orders of magnitude more data than in those earlier studies. This necessitated a different approach, as will be described below. Additionally, we stratified genome sequences as to sampling date. We collected all sequences sampled in the same month since the beginning of the pandemic, and analyzed epistasis month-by-month. In contrast to the earlier analysis we find in the new larger data several mutations in Spike that appear epistatically linked to other mutations in Spike, and outside Spike. Among the highest-ranked such predictions we single out S:S112L (21897), recently associated to vaccine breakthrough infections [39].

The paper is organized as follows. In Section II we present the data used and how we prepared it for further analysis. In Section III we present the theoretical background and the key tools for the DCA analysis. In Section IV we present the results of the study and in Section V we discuss implications and future work. we present the theoretical background and the key tools for the DCA analysis. Technical details and supplementary data are presented in five appendices (A-E). Signs of epistasis in large-scale SARS-CoV-2 data were also investigated on smaller data sets and by other methods in [40, 41] and very recently in data from a wider family of coronaviruses in [42]. We comment on this latter important contribution in Discussion.

II. MATERIALS

A. Data collection

The input data are the genomic sequences of SARS-CoV-2 (high quality and full lengths) as stored in the GISAID public repository [15]. Each of these is a sequence of $\sim 30,000$ base pairs (bps), representing either a nucleotide `A,C,G,T`, or an unknown nucleotide `N`, or one out of a small number of other IUPAC symbols `KYF`, etc., which we will refer to as "minorities"; these represent different sets of the aforementioned nucleotides. Any site/position in the genomic sequence is called a *locus*.

Sequences were sorted by collection date – the typical delay with respect to their appearance on GISAID being > 2 weeks [43] – and grouped on a monthly basis until the end of October 2021. Considering the small number of sequences available for the first months after the outbreak of the pandemic in the 2020, data until the end of March 2020 are grouped together as one data set. In total, we hence have 20 datasets and 3,532,252 sequences. The number of collected sequences N_{seq} per month is shown in Fig. 6: it increases towards 2021, slightly decreases in the first half of 2021, then again greatly increases from July 2021 and finally drops down soon afterwards in September and October 2021.

B. Multiple-Sequence Alignment (MSA)

Multiple Sequence Alignments were constructed using the online tool `MAFFT` [44, 45]. The set of sequences pertaining to one month are aligned to the reference sequence "Wuhan-Hu-1" – GenBank accession number NC-045512 [46]. We note that this is a different procedure compared to what we did in [36], where a pre-aligned MSA was used to lighten the computational burden. The resulting MSAs are given as Supplementary Information (SI) Dataset S2, and are also available on the Github repository [47]. Each MSA is a matrix $\boldsymbol{\sigma} = \{\sigma_i^n | i = 1, ..., L, n = 1, ..., N_{seqs}\}$, where N_{seqs} represents the number of sequences for for a given MSA. This number varies from month to month as shown in Fig. 6. The L columns of the MSA stand for the genomic loci/sites [48, 49]. The total number of loci of the reference sequence is $L = 29,903$; other sequences have somewhat differing lengths. In the MSA gap sequences are inserted in each sequence to align the sequences. The length of sequences before alignment hence differs between the different sets and is different in distinct months. However, this length variation is modest in this data set. The sites between 256 and 29674 in the reference sequence are referred to as *coding region* since they code for the protein-coding genes in the SARS-CoV-2 genome. Inside this coding region there are several *Open Reading Frames* (ORFs) that are translated together into one or several proteins. The longest ORF (ORF1) comprises more than half of the SARS-CoV-2

genome, and is post-translationally divided into 12 proteins called *non-structural proteins*, from nsp1 to nsp12. The other SARS-CoV-2 proteins are mostly structural, and are translated each one from its own ORF. Some of them are named from their structural position such as Spike (S), Membrane (M) and Envelope (E), and some of them are named after their ORF such as ORF3a and ORF8. The two parts of the SARS-CoV-2 genome which are not in the coding region is in the *non-coding region*. Each entry σ_i^n of the MSA $\boldsymbol{\sigma}$ is one of the base pairs mentioned in Sec.(II A) or a new gap symbol "-" introduced for a nucleotide deletion or insertion in the alignment process.

C. MSA filtering

For the MSA filtering, we follow the methods already employed in [36]. As a first step, ambiguous minorities like `KYF`, etc., are converted into `N`, so that there are 6 states `-,N,A,C,G,T`, which we represent as 0,1,2,3,4 and 5 respectively. Subsequently, all the 20 MSAs are filtered. Each locus (column) in each of the 20 MSA is discarded if one of the following two condition is matched: (1) the frequency of a certain nucleotide along this locus is greater than a given value p (lack of variability). A locus is variable if at least a fraction $1-p$ carries other alleles than the major (most frequent) one.; (2) the sum of the frequencies of `A, C, G, T` at this position is less than 0.2 (non-significant). In Fig. 1 the number of survived loci L_s, normalized by N_{seqs}, for each MSA is shown for the threshold $p = 0.98$. Similar results with $p = 0.9$ and $p = 0.999$ are presented in Fig. 7.

III. METHODS

A. Static Quasi-Linkage-Equilibrium (QLE) phase

The Quasi-Linkage-Equilibrium (QLE) state of the distribution of genotypes in a population was found by M. Kimura in a study of the steady-state distribution over two bi-allelic loci evolving under selection, mutation and recombination in presence of both additive and epistatic contributions to fitness [19]. In that example the genotypes were hence AB, aB, Ab and bb where A (a) and B (b) are the major (minor) alleles at respectively the first and second locus. A global QLE state over many loci and its properties was reviewed and investigated in [20], for the case of all loci bi-allelic. The generalization to the case where some loci are multiallelic (more than two alleles per locus) can be found in [50]. The starting point is to model the evolutionary process as a high-dimensional differential equation

$$\dot{P}(\boldsymbol{\sigma}) = \mathscr{F}_{ev}(\boldsymbol{\sigma}) \ , \tag{1}$$

where P is a probability distribution in the space of all possible genomic sequences $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^{L}$ with $\sigma_i = -1, 1$ and $\mathscr{F}_{ev}$ encodes the evolutionary dynamics. This model is meant to capture all effects that act on each individual (genome) separately, and those that act on pairs of genomes. Effects of the first kind are *mutations* which on a bi-allelic genome are flip operations, and *natural selection*. The action of selection is to enhance the likelihood of survival of *more fit* individuals. Biological fitness is a complex and multi-faceted concept. In the present context we are only concerned with fitness that can be encoded in a *fitness function* which furthermore consist only of single-locus and pairwise terms:

$$F(\boldsymbol{\sigma}) = F_0 + \sum_i F_i(\sigma_i) + \sum_{i,j} F_{ij}(\sigma_i, \sigma_j) \tag{2}$$

In biological terminology the F_i are called *additive contributions to fitness* and the F_{ij} *epistatic contributions to fitness*, or simply *epistasis*. The third mechanism underlying QLE is *recombination*. This is the biological mechanism whereby two genomes combine to give a third, *i.e.* sex. In (1) they are represented by a term of the right-hand side which depends on the probability distribution at two different genomes $P(\boldsymbol{\sigma}_1)$ and $P(\boldsymbol{\sigma}_2)$. From a physical point of view, recombination is analogous to collision, and the term in (1) analogous to the collsion operator in the Boltzmann equation.

A static QLE state is then a stationary solution of (1). The covariance of alleles at each pair of loci is a small but non-zero quantity. In presence of pairwise epistasis $F_{ij} \neq 0$ and sufficiently high rate of recombination, the probability distribution $P(\boldsymbol{\sigma})$ reaches a steady-state distribution taking the Gibbs-Boltzmann form:

$$P(\sigma_1, ..., \sigma_L) = \frac{1}{Z} e^{-H(\sigma_1, ..., \sigma_L)}, \tag{3}$$

with

$$H(\sigma_1, ..., \sigma_L) = \sum_i h_i(\sigma_i) + \sum_{ij} J_{ij}(\sigma_i, \sigma_j) \ . \tag{4}$$

where for completely bi-allelic genomes $h_i(\sigma_i) = h_i \cdot \sigma_i$ and $J_{ij}(\sigma_i, \sigma_j) = J_{ij} \cdot \sigma_i \sigma_j$. The parameters h_i and J_{ij} describe the one-time stationary distribution of genomes in a population. As we will infer these parameters from data by the Direct Coupling Analysis (DCA, to be described below) we will refer to in particular the J_{ij} as *DCA terms*. As parameters of the probability distribution they depend on the parameters of the dynamics (1), the fixed point of which (in QLE) is of the form (3). In the (theoretically) simplest setting of recombination stronger than mutations stronger than selection, the relation between J_{ij} and F_{ij} does not involve mutations and has the form $J_{ij}(\sigma_i, \sigma_j) = \frac{1}{r c_{ij}} F_{ij}(\sigma_i, \sigma_j)$ where r is an overall recombination rate and c_{ij} is a measure of genomic distance between loci i and j [20]. This formula has been verified *in silico* [21] (in simulations). Generalizations in other parameter ranges have been derived, and have also been verified *in silico* [33]. It should be noted that distributions of the form (3) are not the only

stationary solutions of (1). In other parameter ranges qualitatively different distributions appear [34] and also non-stationary (but statistically static) solutions [35].

A non-zero correlation between alleles at different loci is called Linkage Disequilibrium (LD). In a probability distribution (3), LD can be identified with thermal spin-spin correlations $\langle\sigma_i\sigma_j\rangle_{th}$. When this holds the relation between correlations and epistasis is nevertheless indirect as it goes through the relation between correlations and parameters $J_{ij}(\sigma_i,\sigma_j)$ in probability distributions of this type; a problem variously called "parameter inference in models in exponential families" [51], or "Direct Coupling Analysis" [52] or "inverse Ising problem" [53, 54]. We note that LD as a concept is not limited to QLE; also other distributions than (3) can have non-zero correlations.

Genetic drift is finally the conventional term designating randomness in a dynamics of the type (1). In a formulation in terms of allele frequencies in a population, the effects of mutations, selection and recombination can be given in terms of their expected (mean) values. Genetic drift is typically inversely proportional to population size (N). In the limit when N tends to infinity it hence vanishes.

B. Transient Quasi-Linkage-Equilibrium (QLE) phase

A QLE state can prevail over a finite time in the sense that correlations and DCA terms J_{ij} in (4) inferred from a temporal snapshot of the population remain stable, while single-locus frequencies change. The mechanism behind such an effect is time-constant epistatic fitness parameters (F_{ij}) coexisting with genetic drift and/or time-changing additive fitness parameters (F_i). One scenario when this occurs is two weakly advantageous mutations at two different sites appear at about the same time in a population and then grow in frequency towards fixation. At the very beginning there is only one mutation present, and there is no variability on which epistatic effects can act. When both mutations are present but one is still at low prevalence, both correlation and DCA analysis will give non-zero but noisy output due to small sample size. In the other end, towards the end when one (or both) mutation are close to fixation both correlation and DCA analysis will give non-zero but noisy output due to small sample size. At the very end when there is only one mutation left, there is again no variability on which epistatic effects to act and correlation or DCA analysis applied to the data will again yield nothing.

In the intermediate region the equations satisfied by single-locus frequencies and two-locus joint frequencies in a finite population were derived in [20] (Eqs. 36 and 37) starting from the same equation (1) as above, for an Ising genome model (bi-allelic genome) and under a diffusion approximation. This approximation is valid when both allele frequencies at both loci are significant, i.e. none is close to zero or to one (fixation). The equations take the form

$$\dot{\chi}_i(t) = (1-\chi_i^2)F_i + \sum_{j\neq i}\chi_{ij}F_j - 2\mu\chi_i + \dot{\zeta}_i \qquad (5)$$

$$\dot{\chi}_{ij}(t) = \left[(1-\chi_i^2)(1-\chi_j^2)F_{ij} - rc_{ij}\right]\chi_{ij} + \dot{\zeta}_{ij} \qquad (6)$$

where $\chi_i = \langle\sigma_i\rangle$ and $\chi_{ij} = \langle\sigma_i\sigma_j\rangle - \chi_i\chi_j$ are signed frequencies and correlations in physical notation, F_i and F_{ij} are additive and epistatic fitness parameters, μ is mutation rate, r overall recombination rate, c_{ij} is a measure of closeness of loci i and j and $\dot{\zeta}_i$ and $\dot{\zeta}_{ij}$ genetic drift noise terms.

It is readily seen that (5) and (6) are qualitatively different. The first equation describes a process driven by noise and $(1-\chi_i^2)F_i - 2\mu\chi_i$, modulated, if there are non-zero correlations in the population, by $\sum_{j\neq i}\chi_{ij}F_j$. Depending on the sign of the net drift, it will hence tend to drive χ_i towards ±1 (fixation or elimination of the mutation). The second equation on the other hand has vanishing drift whenever the expression in the bracket vanishes. It can be checked that with the small field assumptions used in [20] , and stated in terms of the (in principle time-dependent) DCA terms, this vanishing of the bracket corresponds to the above noted (time-stationary) relation $J_{ij} = F_{ij}/rc_{ij}$, and that this is a stable equilibrium ([20], Eq. 25). There can thus be a transient QLE phase where single-locus frequencies may go up in a fluctuating manner for a fairly long time, while J_{ij} and two-mode frequencies remain steady because governed by a relaxation dynamics. An extension of the above to the case where the fastest process is mutations and not recombination can be found in [33].

C. Correlation Analysis and LD

Correlations are a measure of linkage disequilibrium (LD), *i.e.*, of non-random association between different alleles at different loci. For multi-allele distributions, statistical co-variance matrices are defined as

$$C_{ij}(a,b) = \left\langle \mathbf{1}_{\sigma_i,a}\mathbf{1}_{\sigma_j,b}\right\rangle - \left\langle \mathbf{1}_{\sigma_i,a}\right\rangle\left\langle \mathbf{1}_{\sigma_j,b}\right\rangle \qquad (7)$$

where $\mathbf{1}_{\sigma_i,a} = 1$ if $\sigma_i = a$ and zero otherwise, $\langle\cdot\rangle$ indicates the average over q different alleles per locus. As discussed above, in our representation of the GISAID data, $q = 6$. We compute overall correlation between site i and j as Frobenius norms of the statistical co-variance matrices (summation over the inner indices a, b)

$$C_{ij} = \sqrt{\sum_{a=1}^{q}\sum_{b=1}^{q} C_{ij}^2(a,b)}. \qquad (8)$$

D. plmDCA inference for epistasis between loci

Correlations differ from statistical dependency encoded in the J_{ij} through (3) - (4) because the distribution

may not be of the form (3) and because when it is, two loci i and j may be correlated even if their direct interaction J_{ij} is zero. This is possible if they both interact with a third locus k. Many techniques have been developed to infer the direct couplings in Eq. (3), see [54] and references therein. In this work we have used the Pseudo-Likelihood Maximization (plmDCA) method [50, 53, 55–58] to estimate the parameters J_{ij}. The basic idea of plmDCA is to substitute maximum-likelihood inference of parameters from the joint distribution (3) by the simpler one of estimating which parameters best match the conditional probabilities

$$P(\sigma_i|\boldsymbol{\sigma}_{\setminus i}) = \frac{\exp\left(h_i(\sigma_i) + \sum_{j\neq i} J_{ij}(\sigma_i,\sigma_j)\right)}{\sum_{\mathbf{q}} \exp\left(h_i(q) + \sum_{j\neq i} J_{ij}(q,\sigma_j)\right)}; \quad (9)$$

Here $\mathbf{q} = \{0,1,2,3,4,5\}$ are the possible states of σ_i in the dataset and $\boldsymbol{\sigma}_{\setminus i}$ stands for all the loci except the locus i. Assuming independent samples, the functions to optimize (one for each locus) are

$$\begin{aligned}\mathscr{PL}_i\left(h_i,\{J_{ij}\}_j\right) = \\ \frac{1}{N_{seqs}}\sum_s h_i\left(\sigma_i^{(s)}\right) + \frac{1}{N_{seqs}}\sum_s\sum_{j\neq i} J_{ij}(\sigma_i^{(s)},\sigma_j^{(s)}) \\ - \frac{1}{N_{seqs}}\sum_s \log\sum_{\mathbf{q}}\exp\left(h_i(q) + \sum_{j\neq i} J_{ij}(q,\sigma_j^{(s)})\right),\end{aligned} \quad (10)$$

where s labels the sequences (samples), from 1 to N_{seqs}. We use the asymmetric version of plmDCA [58] as implemented in [59] with l_2 regularization with penalty parameter $\lambda = 0.1$. The inferred DCA terms between loci i and j are scored by the Frobenius norm over the inner indices a, b as in (8), and as implemented in [59].

Inference of epistatic interactions as in (2) would then require knowledge of additional model parameters, in particular the overall recombination rate r[20, 21, 33, 50]. This has not been attempted here; we have used DCA terms as proxies for epistatic fitness terms.

E. Removal of phylogenetic confounders

Statistical dependency between allele distributions at two loci can arise both from epistatic contributions in QLE, and from inheritance, for example when two unrelated mutations appeared by chance at the same time in a very fit individual which spread in the same geographic area (phylogenetic effect). The global distribution of genotypes then does not have to be of the Boltzmann form (3), but can instead reflect mixtures of clones [34].

All data from which one wishes to infer epistasis from LD to some extent contain such a combination of the intrinsically epistatic effects, and of phylogeny. In particular, when recombination acts approximately in the same manner along a genome, LD due to phylogeny dominates between pairs of loci that are close, while LD due to epistasis can dominate between pairs of loci that are distant. In earlier studies on whole-genome data from bacterial pathogens, a distance cut-off was therefore employed [60, 61] as well as in previous work on SARS-CoV-2 data [36, 37]. The effect of phylogenetic correlations in DCA-based contact prediction in proteins was recently investigated in [38].

In the current work we have leveraged the well-documented growth of large clones in the global SARS-CoV-2 population. In particular, we have ascribed large scores J_{ij} between pairs of loci to phylogenetic effect when i or j is included in one of three Variants of Concern (VoC) 'alpha' [62], 'beta' [63, 64] or 'delta' [65]. The corresponding tables of mutations and time evolution of mutation frequencies were recently reported by us in [66].

F. Fraction of residual couplings over rank

Let us define the fraction of residual couplings R over n as follows. We start by ranking all possible couplings by their score C_{ij} or J_{ij} computed as in Eq.(8). Within the nth highest ranked couplings, a number k of them is removed when it is likely not to be due to an underlying epistatic effect. We include as possible confounders to be removed when at least one of the terminals, *i.e.*, i or j in C_{ij} / J_{ij} is located in non-coding regions; the loci lie too close ($|i-j| \leq 5$ bps); at least one of the terminals, i or j of C_{ij} or J_{ij}, lie in one of the VoCs.

The fraction of residual couplings is then defined as

$$R(n) = \frac{n-k}{n}. \quad (11)$$

We compute this quantity for J_{ij}s and C_{ij}s respectively. As shown in Fig. 2, the curve for residual couplings defined by plmDCA lies well above the curve defined by correlations for small values of rank n. This strongly suggests that the leading couplings J_{ij} are more likely to capture epistatic effects than leading correlations C_{ij}, as a higher number of the latter is removed according to the above criteria for the same n. We will comment on this in Sec.(IV B).

IV. RESULTS

A. The genome-wide variability (GWV) of SARS-CoV-2 changes in time

We here define the genome-wide variability (GWV) for a genome as the number of loci that shows variability *i.e.*, the number of loci that survive filtering L_s, normalized by N_{seqs} for each MSA. The threshold p used for filtering subjects to $(1-p)\cdot N_{seqs} \gg 1$, indicates the expected number of all minor alleles is greater than one. We define

GWV as

$$\mathrm{GWV} = \frac{L_s}{N_{seqs}} \tag{12}$$

Fig. 1 shows GWV for threshold $p = 0.98$. The GWV increased in the beginning of the pandemic until May of 2020, and then decreased with light fluctuations. In the same time period the number of sequences increased tenfold (Fig. 6), thus the GWV per month has decreased. Similar results hold for other choices of p equals 90% and 99, 9% are discussed further in Appendix B.

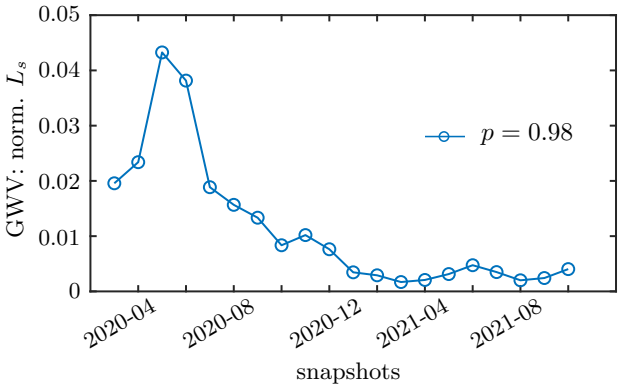


FIG. 1. GWV, the number of loci L_s surviving the filtering process normalized by N_{seqs} for each MSA with threshold $p = 0.98$ until the end of October 2021. At the beginning of the pandemic, GWV was increasing, however with the rapidly increasing number of available sequences on GISAID, the normalized L_s later decreased over month.

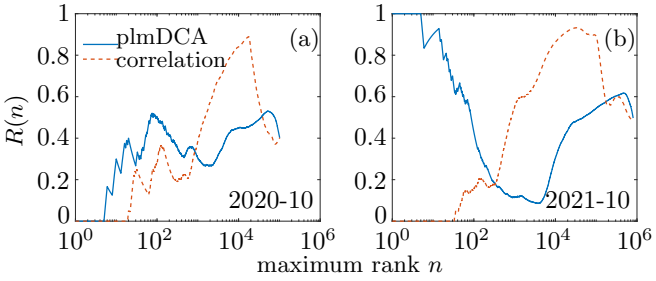


FIG. 2. Examples of fraction of retained couplings $R(n)$ eq. (11) as function of maximum rank considered n for October 2020 (a) and 2021 (b). Blue solid lines show plmDCA J_{ij}, red dashed lines show correlations C_{ij}. Couplings with i, j satisfying the conditions mentioned in Sec.(III F) have been removed. Panel (a) shows one exceptional case for the highest 5 ranked plmDCA scores and the highest 20 ranked correlations have been removed. In this case the distinction between correlation analysis and DCA is not of a qualitative nature. Panel (b) shows the typical case where none of the highest 4 ranked plmDCA scores have been removed while the highest 30 ranked correlations have all been removed. In this case the two curves are qualitatively different.

B. Leading correlations can mostly be explained by the growth of focused SARS-CoV-2 VoCs

A simple way to visualize if ranked effects are due to one out of several factors is to plot the contribution of the factor of interest as function of rank. A standard procedure in DCA analysis of tables of homologous protein sequences is indeed top-k plots illustrating the fraction of k highest rank predictions which correspond to spatially proximate residue pairs. For instance, if we want to assess the effect of the rise of variants in the computed couplings, we can proceed as described in Sec.(III F) by ranking them by magnitude, taking out those related to the variants and plotting $R(n)$ as in eq. (11). This is done in Fig. 2 for one representative and one exceptional month (in fact, the only exceptional month, see Fig. 8 in Appendix C). In both plots the fractions of highest ranked correlations C_{ij} and plmDCA terms J_{ij} are presented, with part of C_{ij} or J_{ij} being removed when i or j matches the removal conditions listed in (Sec.III F). The representative month (October 2021, Fig. 2(b)) shows that the leading correlations can mostly be explained by variations in VoCs alpha, beta and delta. For the exceptional month (October 2020, Fig. 2(a)) the separation between correlations and DCA terms is not clear.

The essence of the argument is that for the same n, leading DCA terms contain much fewer pairs where one or both terminals appear in the VoCs. Lists of leading DCA terms are hence, compared to correlations, enriched for epistatic interactions. Analogous (and similar) results are shown for the other months in Appendix C.

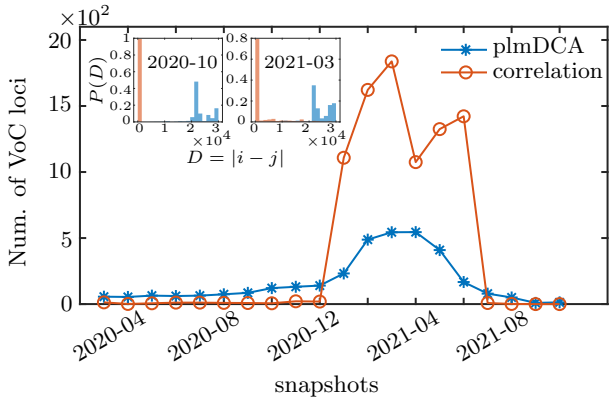


FIG. 3. Main panel: number of loci located in the VoCs from top 2000 J_{ij} (blue squares) and C_{ij} (orange dots) respectively over each month. Links with i, j satisfying the 1st and 2nd removal conditions are discarded. Links with one or both terminals in VoCs are however retained. Correlations provide much more VoC loci than DCAs during December 2020 to July 2021. This is the time period when one or more of the VoC dominated the set of genomes on GISAID on a monthly basis.Insert: the probability of distance $D = |i - j|$ in top 2000 C_{ij} and J_{ij} for October 2020 and March 2021 respectively. Here links that meet any of the removal conditions are removed. Correlations are enriched in links with short D both in and out of the time period dominated by the VoCs.

To check the other possible confounding effects, the number of loci that appears in the focused VoCs are counted in the top 2000 C_{ij} and J_{ij} over month, as presented in the main panel of Fig. 3. It shows that the leading correlations containing more VoC loci comparing with DCAs during the prevalent period of the focused VoCs. To check the possible epistatical or phylogenetic effect, the distance distribution $P(D)$ between locus i and j are computed. Two examples of October 2020 and March 2021 (out of and in the taken over period of the focused VoCs) are presented in the inner panels of Fig. 3 respectively. In both cases, correlations tend to single out links between closely spaced loci. This explains the big jump between C_{ij} and J_{ij} during the end of 2020 up to the middle of 2021.

C. Inferred epistasis has both invariant and variant aspects

One novelty of the analysis presented in this work is that the dataset is much larger than in previous contributions [36, 37]. For this reason we grouped data by month of sampling time. As we will see this leads to new effects. A second novelty is that phylogenetic confounders have been eliminated by excluding inferred links where one or both loci appear in the VoCs of SARS-CoV-2.

Fig. 4 displays the rank in logarithm scale of leading residual epistatic interactions J_{ij} (solid lines) and correlations C_{ij} (dashed lines) with same i and j as a function of sampling time. A subset of top 200 J_{ij} with is or js excluded or included in the focused VoCs are provided in Fig. 4(a) and (b) respectively. Their counterpart C_{ij} are shown in dashed lines. One feature that stands out on these two sub-graph is that as long as they appear in the data, both types of ranks appear fairly stable, but the ranks of correlations is far lower. The ranks of the DCA terms fluctuate (roughly) in the interval 1-20 while that of the corresponding correlations vary between 10,000 and 200,000. Similarly, a subset of top 2000 C_{ij} and their corresponding J_{ij} with i or j located out of or inside the focused VoCs are displayed in Fig. 4(c) and (d) respectively. Here the J_{ij}s last longer than the corresponding C_{ij}s over time.

Furthermore Fig. 4 shows that none of the interactions appear for the entire period, but only in some time window. Outside this window, the frequency of the major allele of one or both loci in a pair rises above the threshold p and is hence discarded because of lack of variability. As a consequence, the pair hence disappears from the analysis. We can therefore at best have a *transient* QLE phase, as defined in Sec.(III B).

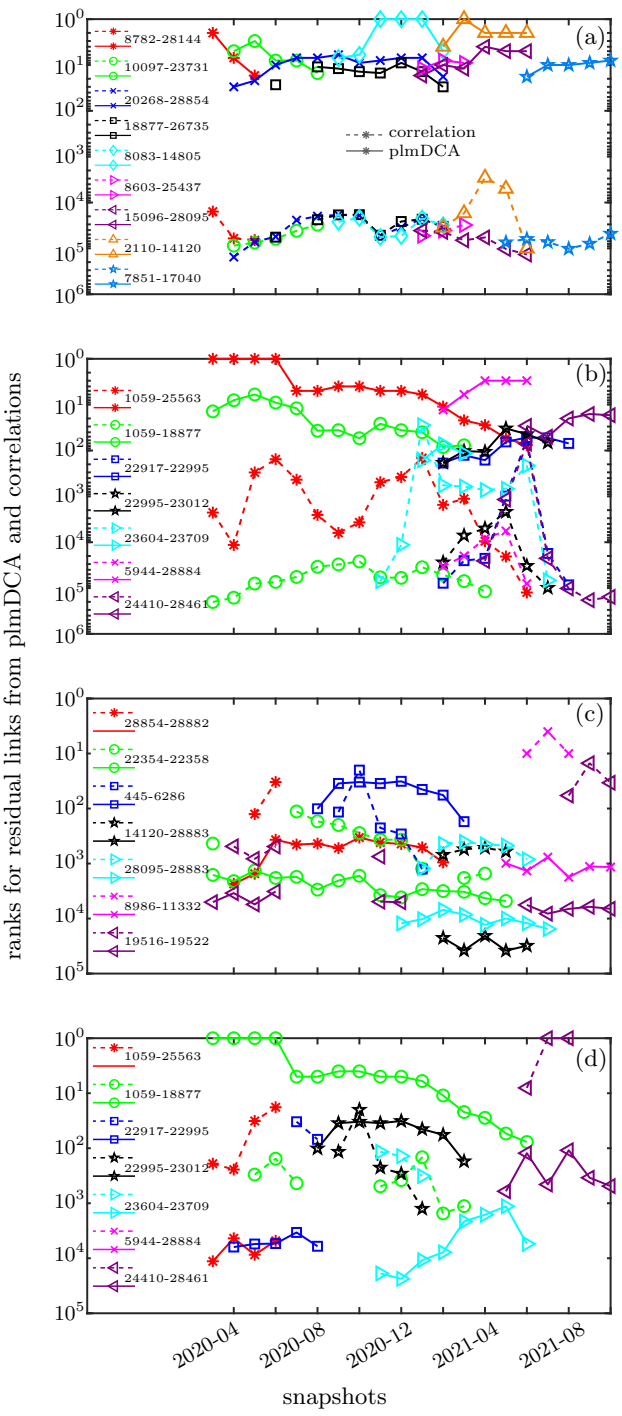


FIG. 4. Ranks for a subset of residual epistatic interactions per month inferred by plmDCA (solid) J_{ij} and correlations (dashed) C_{ij}. The pairs with i or j matching condition 1 and 2 described in Sec. (III F) have been discarded. (a) and (b) display the ranks of a subset of top-200 J_{ij} and their counterparts C_{ij} with i, j excluded and included in the focused VoCs respectively.J_{ij} extend more or less the same period with their counterparts but with much higher rank. Similarly, the ranks of a subset of top-2000 C_{ij} and the corresponding J_{ij} with i or j located out of or inside the focused VoCs are presented in (c) and (d) respectively. J_{ij} extend much longer than C_{ij} over time.The data for this figure is given in Datasets 2 and 3 of Appendix E.

D. A subset of mutations of Variant of Concern omicron has non-trivial dynamics

We here also find it of interest to describe the dynamics of the individual mutations listed for the more recent VoC omicron [67]. We note that this variant, dominant in the world-wide population today, rose to prominence after the time interval on which the present study is based.

In earlier work we showed that a subset of the mutations in alpha, beta and delta have different dynamics than would be expected from a clone growing de novo [66]. In Fig. 5 we show that the same holds for omicron. The red-dot trajectory shows the frequency of a nucleotide substitution at position 21846, which corresponds to S:T95I in Spike, this being listed as one of the defining mutations of omicron [67]. The formula S:T95I is to be read as a point mutation at the locus 21846 which lies in the codon coding for the 95th amino acid of the Spike protein (S). The mutation changes the amino acid from that of the reference sequence T (threonine) to that of the mutant I (isoleucine). This mutation in the N-terminal domain of S1 subunit of the Spike protein rose quickly in frequency from May-21 to June-21 in GISAID database, and has since been found in about half of the samples. Its prevalence therefore cannot be explained by the rise of omicron, which appeared later. Indeed, T95I was among the most common in the variant B.1.526 (Iota) which spread mainly in USA in late 2020 and early 2021 [68], and has been observed in strains classified as VoC delta circulating in France [69], and in UK and Germany [70]. It is therefore an example of a mutation which though classified as a defining mutation of one strain of the virus, in fact is more widely spread, and can be found also in other strains. The other curves in Fig. 5 with a dynamics different from omicron can mostly be explained as also belonging to VoCs alpha and delta.

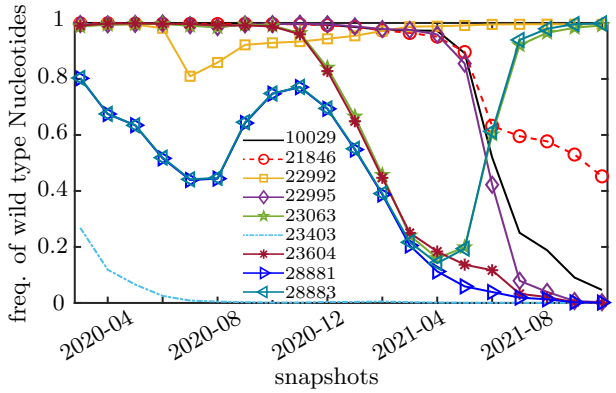


FIG. 5. Wild-type nucleotide frequencies of the loci of the omicron variant with heavy fluctuations listed in [67]. The wild type is defined here as reference sequence "Wuhan-Hu-1", GISAID access ID: EPI_ISL_402125; nucleotides are numbered based on their locus in the latter. The growth of the omicron variant happens later than the time-window displayed. For most of the omicron loci the corresponding wild-type frequency remains close to 1 (thin lines in figure). Loci showing large fluctuations are plotted with bold lines. 23403 is the Spike mutation D614G which rose early in the pandemic while 23063 and 23604 are S:N501Y and S:P681H, which appeared in alpha. 22995 is S:T478K which appeared in delta. The remaining mutations are 21846 (dashed marked red line) discussed in text, 28881 (N:R203M) which is listed in delta but follows another dynamic [66], 10029 which follows a trajectory characteristic of delta without being listed for it, and 28883 which does the same for alpha.

E. DCA detects epistatic interactions between loci in Spike, and between Spike and other genes

Let us focus here on the couplings involving sites on the SARS-CoV-2 that code for the Spike protein, crucial for the virus to bind to target cells. In an earlier study using data up to August 2020, only one large plmDCA score involving a locus in Spike was detected, at genomic position 23403 [36] (Table 1). This G→A substitution was deemed to be due to a phylogenetic effect as detected by a randomization procedure [36] (Table 2), and therefore not retained as a predicted epistatic link in that study. In fact, this substitution corresponds to the well-known mutation Spike D614G which grew in frequency in the early phase of the pandemic.

In the present (larger) month-by-month data we find several persistent plmDCA couplings with terminals in Spike. Some of them are related to the variants alpha, beta, delta and also the more recent omicron [67] (in the sense above), some others are not. Table IV E gives for the months August-October 2021 the couplings whose both terminals are in a Spike coding region while Table IV E lists those for which only one of the extrema is in a Spike coding region; corresponding data for all months is mentioned in Appendix E Dataset 4.

The largest inferred Spike-Spike interaction in all three months, not related to any mutation appearing in alpha, beta, delta or omicron, is between two synonymous mutations $D574D$ and $D1259D$. Most of the other pairs in Table IV E either have one terminal listed in delta, or are somewhat close along the genome, within 35 bp, or involve a synonymous mutation. This includes the pair $(21846, 24208)$ appearing in October, where the first locus is the mutation S:T95I, part of the definition of omicron and discussed above, while the second locus is S:I882I.

The first prediction appearing in Table IV E, in all three months, is the pair $(17236, 24208)$ where the first locus is in gene nsp13 and the second is a synonymous mutation in Spike. This is fact the largest effect detected by the DCA analysis in all three months (rank 1 in Table IV E).

The second prediction in Table IV E, ranked respectively 14, 13 and 10, is the pair $(7851, 21846)$ where the first is the mutation A1711V in nsp3. In Table IV E the variation at locus 21846 (S:T95I) appears also together with nsp2:K81N in all three months, to-

gether with ORF8:P38P and ORF7a:V71I in August, and together with nsp6:A2V, N:S327L, nsp13:I334V, nsp12:S837S, ORF3a:E239, N:Q9L and ORF7a:G38G in October. The mutation nsp3:1711V was a defining mutation of Variant of Interest labelled N.9 discovered in Brazil in 2020 [71]. The mutation nsp2:K81N has been detected in variants of VoC delta circulating in Russia [72].

The third prediction in Table IV E, ranked respectively 20, 16 and 17, is the pair (28461, 24410) where the first is the mutation G63D in N, and the second is N950D in Spike. N:G63D is a defining mutation of VoC delta while S:N950D is a reverse mutation of delta defining mutation D950N, identified as such in a recent study [70].

The next two predictions which appear in all three months and which do not involve any of the above or any variants both involve locus 21897 (S:S112L) with partners respectively 26107 (ORF3a:E239Q) at ranks 52, 57 and 60, and 27507 (ORF7a:G38G) at ranks 57, 71 and 74. Spike mutation S112L was recently associated to vaccination breakthrough infections in New York City [39]. That study also identified that genes ORF3a (56%) and ORF8 (67%) had higher numbers of sites with enriched mutations in breakthrough sequences. ORF3a mutation E239Q is located at the protein C-terminal, and appears in the sub-variant of VoC delta variously labelled AY.25 and B.1.617.2.25; it has no annotation in UniProt.

V. DISCUSSION

In this work we have applied the Direct Coupling Analysis (DCA) methodology [52, 54, 57, 58, 73, 74] to identify putative epistatic interactions between pairs of loci in the SARS-CoV-2 virus. We have described the rationale for such an approach based on the Quasi-Linkage Equilibrium (QLE) mechanism of Kimura [19, 20], which we have recently combined with DCA in *in silico* validation [21, 33, 50]. As part of the world-wide effort to combat the COVID19 epidemic an unprecedented number of genomes of the disease agent have been obtained and released through open repositories. In this study we have thus been able to use more than three and a half million full-length high-quality SARS-CoV-2 genomes from GISAID deposited until October 2021 [15]. Such very large, quasi-exact and easily accessible data resources will very likely be the norm in future pandemics. Methods to turn them into actionable information in new ways are therefore of high relevance. Except for the more that order of magnitude larger data size, the main methodological novelty in this study has been to separate genomes as to sampling date (by month). We have hence been able to carry out a *temporal* epistasis inference, to the best of our knowledge for the first time.

Our main finding is that the leading terms identified by DCA and those counterparts by correlations are fairly stable over time, while the ranks of correlations are much lower and with larger fluctuations. Furthermore, the leading correlations appear in the lists as leading for shorter time periods compared to DCA terms. This observation is an argument in favor of the global SARS-CoV population exhibiting characteristics of QLE, as would be expected from the substantial rate of recombination characteristic of coronaviruses [22] and the sometimes high rate of circulating infections in the human population world-wide. The finding however comes with a caveat: DCA analysis (and correlation analysis) is necessarily based on observed variability which disappears if an allele at a locus is lost. This is indeed what we find. The stability of DCA terms therefore only pertain for the time window when the mutations at both terminals appear in a significant proportion of the samples. Few of the epistatic interactions found in two earlier studies [36, 37] are therefore in fact found in the later data, as one or both of the corresponding mutations have either since been lost or reached fixation.

We refer to the resulting setting as *temporal epistasis inference*. In earlier theoretical work we identified the possibility of retrieving epistatic parameters from pairwise variations in a population even though single-locus frequencies vary greatly [59]. In this work we have found that such an effect appears in data, and is reflected in the epistasis prediction pipeline through the appearance and/or disappearance of predicted pairs. The biological relevance is that epistasis can be detected in a transient phase, and then used as input to further analysis at a later time, when variations at one or both terminals will have disappeared, and epistasis can no longer be detected from the sequences present in the population. We further remark that in the data at hand (SARS-CoV-2 sequences collected in the COVID-19 pandemic) evolutionary parameters are themselves most likely changing with time. The most immediate effect is the changed fitness landscape (to the virus) after large-scale vaccination (of the human hosts). We have in this work not tried to estimate such effects.

The main success story of DCA applied to biological data has been to predict spatial residue-residue contacts in proteins [48]. In that important application accuracy of predictions can be assessed by comparing to distance data in resolved protein structures. Spatial proximity is the main mechanism behind and a relevant proxy for epistasis within one gene (one protein). It is a general feature of DCA that the accuracy is generally highest for the largest predictions, typically visualized through plots of the True Prediction Rate of k'th largest predictions ($TPR(k)$)[48]. On the global genome scale labelled test data of the same kind is not available, and evaluation will necessarily be in terms of potential biological or medical relevance, compared to literature, or other data.

In the bacterial domain, in an earlier study based on around 3,000 full-length genomes of the bacterial pathogen *Streptococcus pneumoniae* we hence found as main terms epistatic interactions between loci in the PBP family of proteins central to antibiotic resistance in the pneumococcus [60]; analogous results have also

August 2021					September 2021					October 2021				
rank	locus 1	AA-m.	locus 2	AA-m.	rank	locus 1	AA-m.	locus 2	AA-m.	rank	locus 1	AA-m.	locus 2	AA-m.
7	23284	D574D	25339	D1259D	7	23284	D574D	25339	D1259D	9	23284	D574D	25339	D1259D
16	21987	G142D	24410	D950N	15	21987	G142D	24410	D950N	11	21995	T145H	22227	A222V
67	22093	M177I	22104	G181V	45	21995	T145H	22227	A222V	15	21987	G142D	24410	D950N
70	22917	R452L	22995	K478T						135	21846	T95I	24208	I882I
71	22082	P174S	22093	M177I										
74	22081	Q173H	22093	M177I										
190	22082	P174S	22104	G181V										
195	22081	Q173H	22104	G181V										

TABLE I. Largest DCA terms with both terminals in Spike coding region, August-October 2021. Top-200 couplings computed as plmDCA scores are considered. For each of them in the three months displayed, there's the indication of the rank, the two loci involved and the corresponding amino acid (AA) mutations. Light gray color indicates that this mutation is found in delta variant. Dark gray color indicates that this mutation is found in omicron variant. Couplings with one or both terminals colored in light gray are attributed to a phylogenetic effect. The single pair with one terminal colored in dark gray is not attributed to a phylogenetic effect, the growth of omicron being later than October 2021. Omicron mutations used here are taken from [67] on page 18, deletions not considered.

August 2021					September 2021					October 2021				
rank	Partner		Spike		rank	Partner		Spike		rank	Partner		Spike	
	locus	AA-m.	locus	AA-m.		locus	AA-m.	locus	AA-m.		locus	AA-m.	locus	AA-m.
1	17236	nsp13:I334V	24208	I882I	1	17236	nsp13:I334V	24208	I882I	1	17236	nsp13:I334V	24208	I882I
14	7851	nsp3:A1711V	21846	T95I	13	7851	nsp3:A1711V	21846	T95I	10	7851	nsp3:A1711V	21846	T95I
20	28461	N:G63D	24410	D950N	16	28461	N: D63G	24410	D950N	17	28461	N:D63G	24410	D950N
27	1048	nsp2:K81N	21846	T95I	36	1048	nsp2:K81N	21846	T95I	20	25614	ORF3a:S74S	21995	T145H
52	26107	ORF3a:E239Q	21897	S112L	52	25614	ORF3a:S74S	21995	T145H	21	25614	ORF3a: S74S	22227	A222V
57	27507	ORF7a:G38G	21897	S112L	57	26107	ORF3a:E239Q	21897	S112L	30	1048	nsp2:K81N	21846	T95I
62	18086	nsp14:T16I	22792	I410I	58	25614	ORF3a:S74S	22227	A222V	51	10977	nsp6:A2V	21846	T95I
76	27291	ORF6:D30D	24208	I882I	71	27507	ORF7a:G38G	21897	S112L	56	27291	ORF6:D30D	24208	I882I
79	1729	nsp2:V308V	22792	I410I	82	27291	ORF6:G30G	24208	I882I	60	26107	ORF3a:E239Q	21897	S112L
151	28007	ORF8:P38P	21846	T95I	83	11514	nsp6:T181I	22227	A222V	63	29253	N:S327L	21846	T95I
168	27604	ORF7a:V71I	21846	T95I	128	17236	nsp13:I334V	21846	T95I	64	18744	nsp14:T235T	24130	N856N
174	17236	nsp13:I334V	21846	T95I	151	18744	nsp14:T235T	24130	N856N	74	27507	ORF7a:G38G	21897	S112L
197	11514	nsp6:T181I	22227	A222V	190	5584	nsp3:T955T	22227	A222V	80	17236	nsp13:I334V	21846	T95I
					195	13019	nsp9:L112L	22227	A222V	124	15952	nsp12:S837S	21846	T95I
										153	26107	ORF3a:E239	21846	T95I
										163	28299	N:Q9L	21846	T95I
										190	27507	ORF7a:G38G	21846	T95I
										194	11562	nsp6:C197F	21897	S112L
										197	11514	nsp6:T181I	22227	A222V

TABLE II. Largest DCA terms with only one terminal in Spike coding region, August-October 2021. Top-200 couplings computed as plmDCA scores are considered. For each of them in the three months displayed, theres the indication of the rank, the locus in the Spike coding region and corresponding amino acid (AA) mutation, the locus in the partner coding region and corresponding amino acid (AA) mutation. Light gray color indicates that this mutation is found in delta variant. Dark gray color indicates that this mutation is found in omicron variant. Pairs with one or both terminals colored in light gray are attributed to a phylogenetic effect, while the several pairs with one terminal colored in dark gray are not, the growth of omicron being later than October 2021. Omicron mutations used here are taken from [67] on page 18, deletions are not considered.

been found for the gonococcus [61]. Recent results use on one hand over $60,000$ *Escherichia coli* genomes, and on the other hand a set of closely related other bacterial genomes, leading to testable predictions on amino acid variability [75].

In the viral domain DCA methods have been applied to the genes coding for the envelope of HIV in a well-known series of papers [76–81]; more have led to experimental tests [82] promising for anti-viral drug and vaccine development. The same group has also extended the analysis to polio [83]. On the global genome level a recent contribution used whole-genome sequences of a set of coronaviruses to predict mutability using DCA methods which were then assessed by the use of the same GISAID data base as we have used here [42]. Leveraging a more variable set of genomes is an alternative and possibly more robust avenue to obtain biologically viable predictions; the issue however merits further investigations.

We have here limited ourselves to a discussion of the top-200 predictions per month that are also stable in rank over the last three months of data (August-October 2021), and which involve loci in Spike. We find several

DCA terms associated to Variants of Concern delta and omicron, which we in the case of delta attribute to a phylogenetic effect. On methods to remove phylogeny as a confounder of DCA we refer to [38, 84], and as described in our earlier contribution [36]. Recently, an alternative way to infer and tease out epistasis from the linkage effect has been proposed in [85, 86] based on the average over multiple independently-evolving populations. The Wright-Fisher model used there containing no recombination which is distinct with the assumption from QLE. It would be worth comparing both methods for the epistasis inference from the available SARS-CoV-2 genomic sequences.

The most prominent of the mutations in omicron is S:T95I at genomic position 21846. Although a defining mutation for this VoC, it was actually found in approximately half of the genomes collected world-wide in the time period August-October 2021. The inferred epistatic interactions between S:T95I and loci in other genes are hence examples of interactions that were detectable in data up to the end of 2021, but which is not detectable anymore as the omicron variant has taken over fully.

Our results of potential biological and medical relevance are given in Table IV E for epistatic interactions between two loci out of which at least one in Spike. We surmise that the most interesting of those are two epistatic interactions involving Spike mutation S112L, recently shown to be associated to vaccination breakthrough infections [39]. One of its interaction partners is mutation ORF3a:E239Q where ORF3a is a cation channel protein unique to the coronavirus family [87] and known to be involved in inflammation of lung tissue and severe disease outcomes [88–90]. In the earlier study [36] several other mutations in ORF3a appeared prominently; in this study a new one does so together with a mutation in Spike.

ACKNOWLEDGMENTS

We thank Dr Edwin Rodríguez Horta, Profs Martin Weigt and Roberto Mulet for numerous discussions. EA thanks Kaisa Thorell and Rickard NordRén for useful suggestions. The work of HLZ was sponsored by NSFC 11705097, NY221101. YL was supported by KYCX21-0696. The work of EA was supported by the Swedish Research Council grant 2020-04980.

Appendix A: The number of sequences sampled per month has increased during the pandemic

Fig. 6 shows the number of whole-genome high-quality SARS-CoV-2 sequences deposited in GISAID and stratified by month. With some irregularity this number has grown exponentially since the summer of 2020, and was towards the end of the studied period around half a million SARS-CoV-2 genomes per month.

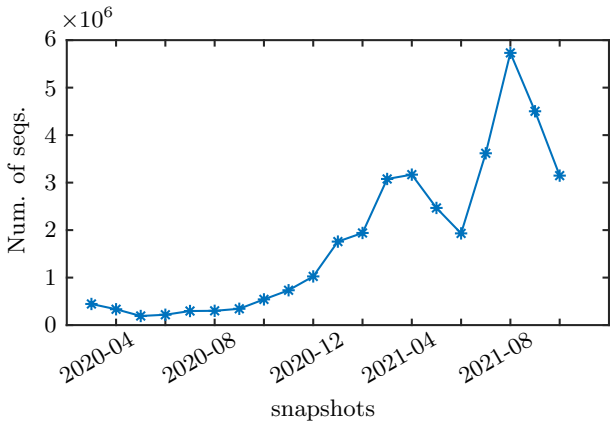


FIG. 6. Number of complete and high quality SARS-CoV-2 sequences deposited in the GISAID repository per month until the end of October 2021, stratified by month of sampling time. The accession IDs of the samples from GISAID used in this work are given in Appendix E Datasets 1.

Appendix B: GWV with other filtering thresholds

We computed the frequencies of nucleotides along each locus/column in each MSA matrix. If the frequency of any of the nucleotides is larger than the given value of p, this locus will be excluded in the following epistasis analysis. To complement Fig. 1 in the main text, we show plots for the normalized number of survived loci L_s by the number of sequences N_{seqs} in each MSA with different values of p here in Fig. 7. The upper panel is for $p = 0.9$ while the bottom one for $p = 0.999$ respectively. They show similar patterns with $p = 0.98$ in the main text.

Appendix C: Fractions of residual couplings

This appendix shows the fraction of residual (epistatic) couplings for plmDCA and correlation analysis as a function of the top-k links considered, as shown in Fig. 8. Data for the months Oct 2020 and Oct 2021 are also shown in Fig. 2 of the main text. For the highest ranks, plmDCA gives a greater fraction of true epistatic predictions with respect to correlation analysis. Couplings are removed if one or both i, j meets/meet the removal conditions as described in Sec.(III F).

Appendix D: Circos plots with different filtering values for each month

For each monthly dataset, three different p values are employed for filtering the loci. If the percentage of a same major nucleotide along a column is larger than the given value (0.93, 0.95 and 0.98), then the column is discarded in the following DCA analysis.

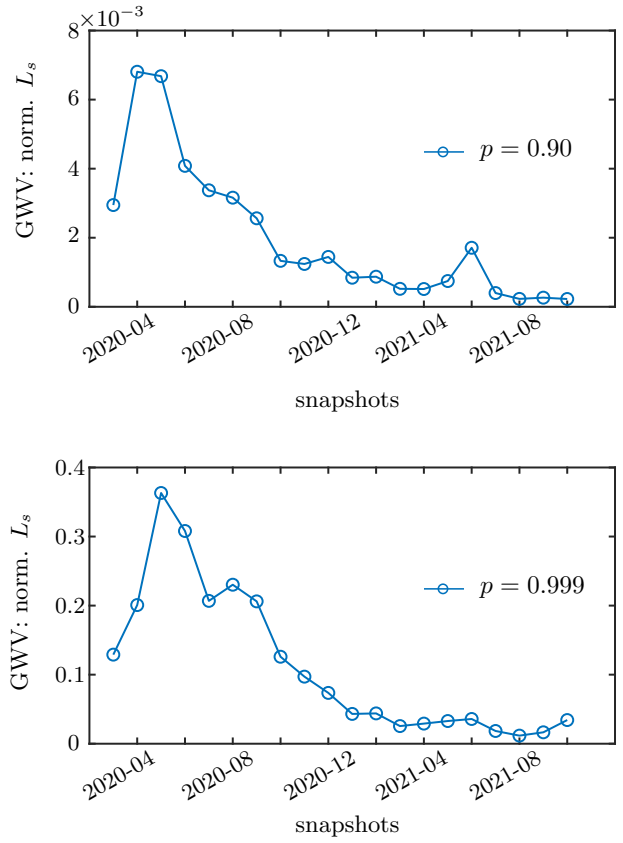


FIG. 7. Normalized survived loci L_s by the total number of sequences N_{seqs} per month with $p = 0.9$ (upper) and $p = 0.999$ (bottom). In our analysis, loci (MSA columns) where any of the nucleotides is found with a frequency greater than p is excluded.

With plmDCA analysis, each pair of retained loci gets a score, which is related to the the epistasis between them. The pairwise epistatic links can be sorted by ranking the scores. Here, we plot the top-200 epistasis with the "Circos" software [91]. Only those located in the coding region are shown. The short links *i.e.*, those with distances between two terminals less than 4bps, are not included. Colored links are for those within top 50 ranks. Red ones for short links while blue ones for the long links. The greys are those within rank 51 to 200.

Appendix E: Data resources

All datasets listed below are available on github [47].

Dataset1 Accession_IDs.xlsx
The Accession IDs for the genomic sequences we used in the analysis. The prefix of each sequence "EPI_ISL_" is excluded to decrease the file size. This dataset is cut into two separate files further to satisfy the limitation of file size on Github, names as "Dataset1-1-Mar-2020-May-2021-Accession_IDs.xlsx" and "Dataset1-2-Jun2021-Oct-2021-Accession_IDs.xlsx" respectively on Github.

Dataset2 p0.98_plm_Top200_No_3variants.xlsx
This dataset contains selected links in top 200 plmDCA epistasic couplings, as ranked by their score. The plmDCA links shown in Fig. 4(a) and (b) in the main text are based on this dataset. Here, the links located in the non-coding region and with close locus (≤ 5) and any loci included in alpha, beta, delta are excluded.

Dataset3 p0.98_Top_2000_CA_No_variants.xlsx
This dataset lists the sorted correlation scores which correspond to the dashed correlations in the middle and bottom Fig. 4(c) and (d). Similarly to its plm counterpart, links in coding region and the distance between loci is larger than 5bps are considered. No variant is included.

Dataset4 links_with_Spike_locus_or_loci_ranks.xlsx
The epistasis provided by plmDCA and correlation analysis are included in this dataset for each month. Only those within top-200s, for which the distance between two terminals is > 5 loci and whose both terminals located in the coding region are listed in the dataset. The genomic positions provided in Table 1 and 2 in the main text are based on this dataset.

Dataset5 protein_aa_mut_for_links_in_Dataset4.xlsx
We provide the links within top 200s plmDCA scores that containing Spike terminals, for each month. The short links with loci located within 5 bps are discarded. Here, we also annotate the genes to which loci in the Dataset4 belong to and the corresponding amino acid mutations. The annotated genes and corresponding amino acid mutations in Table IV E and IV E in the main body of the manuscript are based on this dataset.

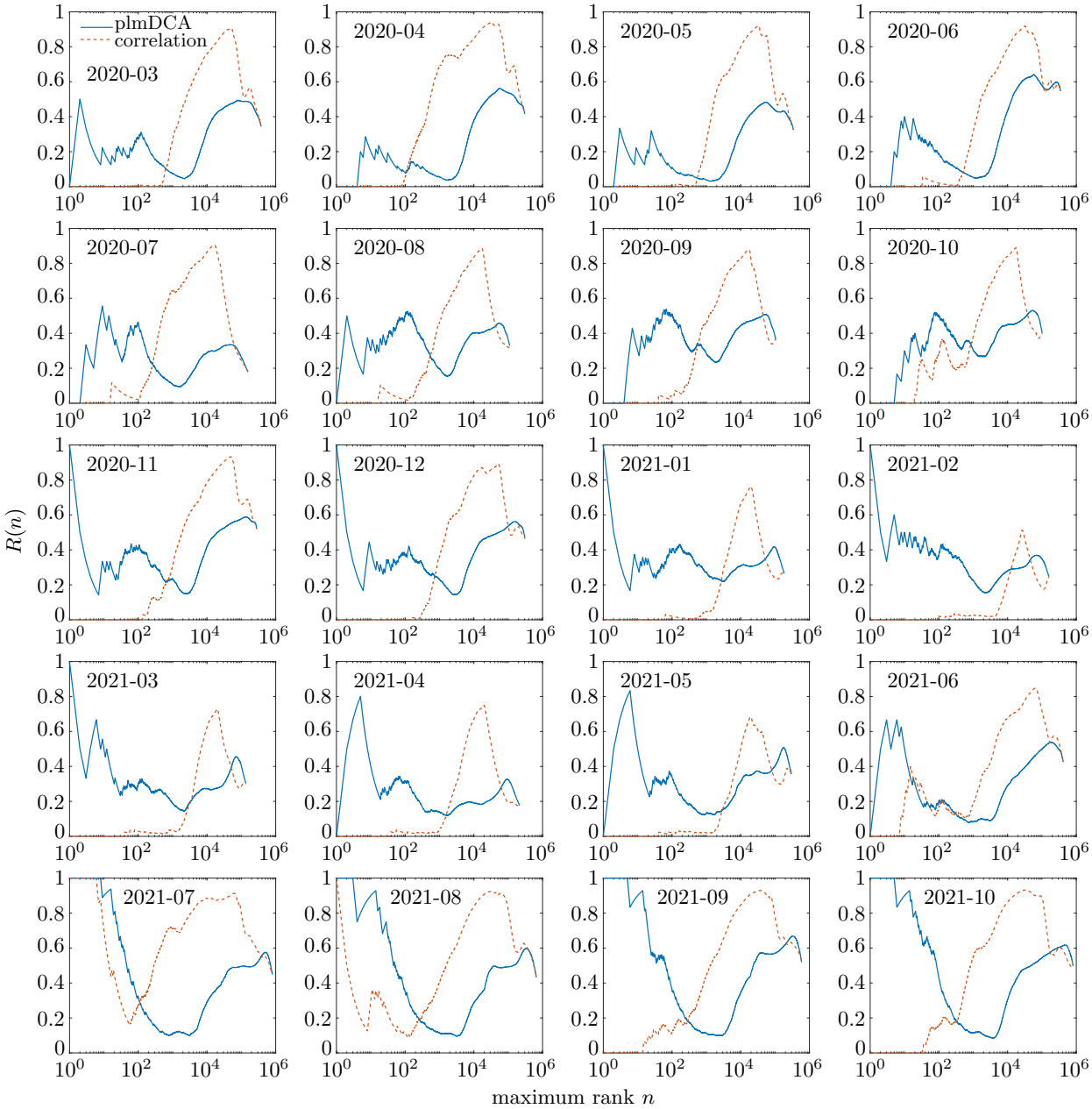
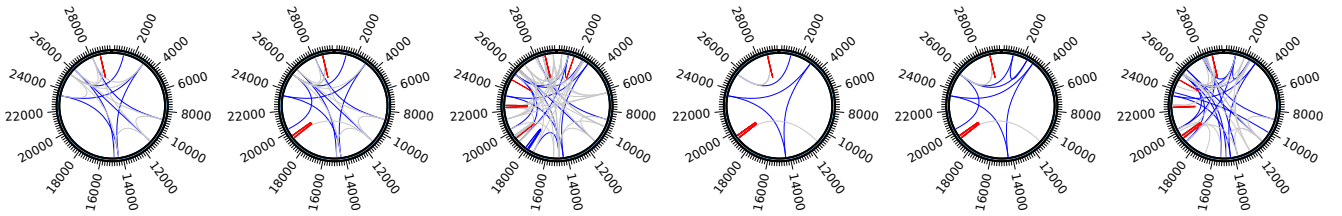
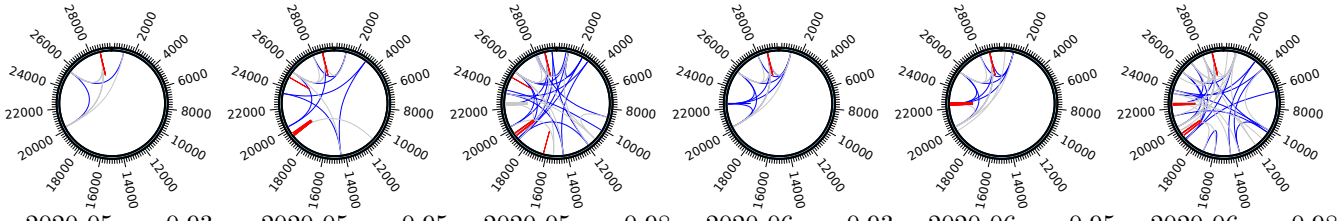
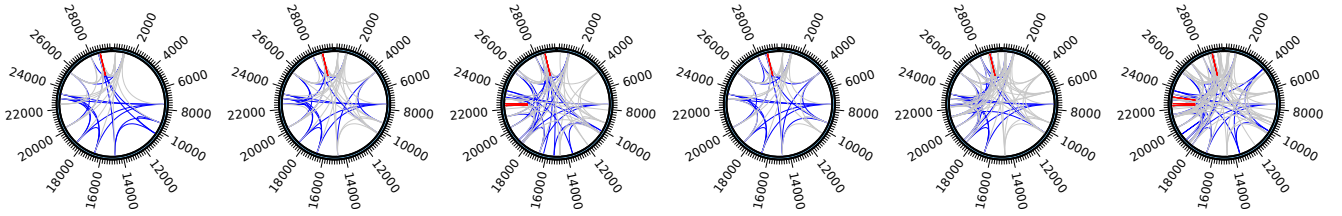
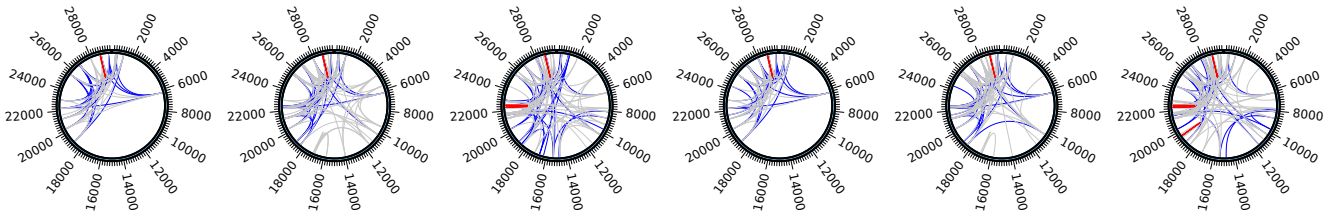


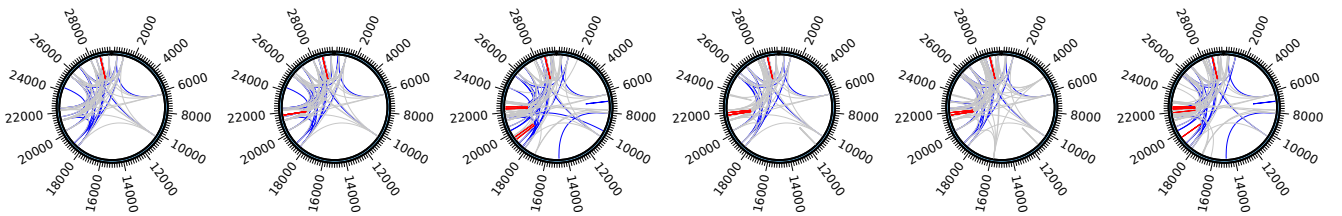
FIG. 8. Fraction of the residual epistatic couplings over the top-k considered. Couplings between close sites (< 6 bps), those with extrema in non-coding regions and those related to VoCs are excluded. Blue line for plmDCA while red for correlation analysis.

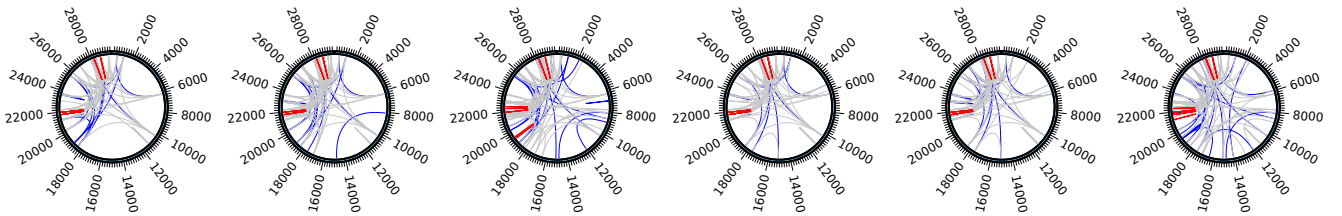


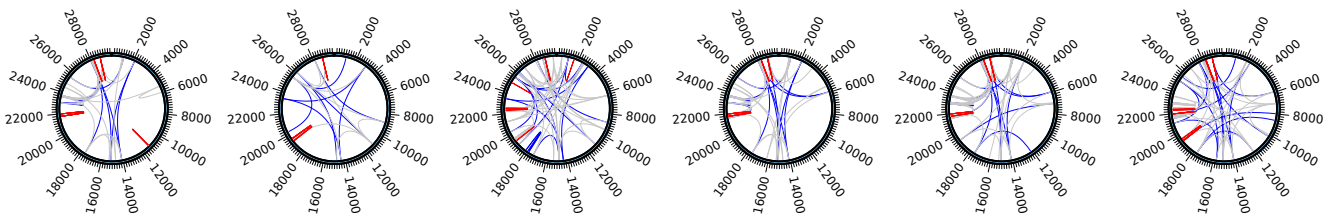












(Fig.9 To be continued)

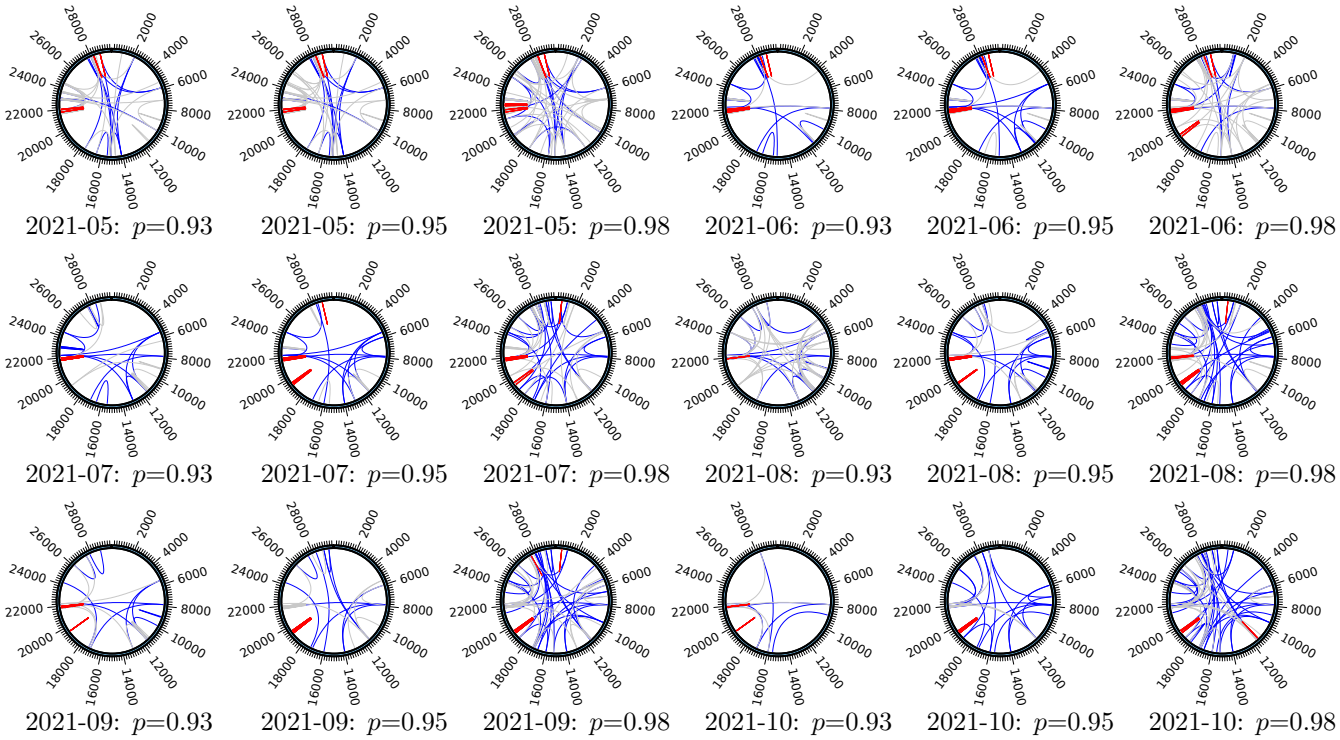


FIG. 9. Circos plots for epistatic links located in the coding region, top 200s ranks. The p values for each month are selected as 93%, 95% and 98% respectively.

[1] World Health Organization, Coronavirus disease (covid-19) pandemic (2022), accessed August 19, 2022.
[2] S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, *et al.*, Nature **584**, 257 (2020).
[3] H. Salje, C. T. Kiem, N. Lefrancq, N. Courtejoie, P. Bosetti, J. Paireau, *et al.*, Science **369**, 208 (2020).
[4] M. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, K. Brennan, P. David, *et al.*, Science **368**, 493 (2020).
[5] G. N. Wong, Z. J. Weiner, A. V. Tkachenko, A. Elbanna, S. Maslov, and N. Goldenfeld, Phys. Rev. X **10**, 041033 (2020).
[6] N. Perra, Physics Reports **913**, 1 (2021).
[7] T. T. Le, J. Cramer, R. Chen, and S. Mayhew, Nat. Rev. Drug. Discov. **19**, 667 (2020).
[8] F. Amanatand F. Krammer, Immunity **52**, 583 (2020).
[9] D. Gordon, G. Jang, M. Bouhaddou, and *et al.*, Nature **16**, 026002 (2020).
[10] A. Frediansyah, R. Tiwari, K. Sharun, K. Dhama, and H. Harapan, Clin. Epidemiology Glob. Health **9**, 90 (2021).
[11] L. Tse, R. Meganck, R. Graham, and R. Baric, Front. Microbiol. **11**, 658 (2020).
[12] F. K. Yoshimoto, Protein J **39**, 198 (2020).
[13] E. Hartenian, D. Nandakumar, A. Lari, M. Ly, J. Tucker, and B. Glaunsinger, J. Biol. Chem. **295**, 12910 (2020).
[14] M. A. Martin, D. VanInsberghe, and K. Koelle, Science **371**, 466 (2021).
[15] Y. Shuand J. McCauley, Eurosurveillance **22**, 1 (2017).
[16] J. Phelan, W. Deelder, D. Ward, S. Campino, M. L. Hibberd, and T. G. lark, Controlling the SARS-CoV-2 outbreak, insights from large scale whole genome sequences generated across the world, biorxiv (2020).
[17] R. A. Blytheand A. J. McKane, J. Stat. Mech.: Theory Exp. **2007** (07), P07018.
[18] R. A. Neherand B. I. Shraiman, Proc. Natl. Acad. Sci. **106**, 6866 (2009).
[19] M. Kimura, Genetics **52**, 875 (1965).
[20] R. A. Neherand B. I. Shraiman, Rev. Mod. Phys. **83**, 1283 (2011).
[21] H.-L. Zengand E. Aurell, Phys. Rev. E **101**, 052409 (2020).
[22] M. M. Laiand D. Cavanagh, Adv. Virus Res. **48**, 1100 (1997).
[23] R. L. Grahamand R. S. Baric, J. Virol. **84**, 3134 (2010).
[24] F. Robson, K. S. Khan, T. K. Le, C. Paris, S. Demirbag, P. Barfuss, *et al.*, Mol. Cell **79**, 710 (2020).
[25] J. Gribble, A. J. Pruijssers, M. L. Agostini, J. Anderson-Daniels, J. D. Chappell, X. Lu, *et al.*, PLoS Pathog. **17**, e1009226 (2021).
[26] V. Avanzato, J. Matson, S. Seifert, R. Pryce, B. Williamson, S. L. Anzick, *et al.*, Cell **183**, 1901 (2020).
[27] J. H. Baang, C. Smith, C. Mirabelli, A. Valesano, D. Manthei, M. Bachman, *et al.*, J. Infect. Dis. **223**, 23 (2021).
[28] B. Choi, M. C. Choudhary, J. Regan, J. A. Sparks, R. F. Padera, X. Qiu, *et al.*, N. Engl. J. Med. **383**, 2291 (2020).
[29] M. Hensley, W. Bain, J. Jacobs, S. Nambulli, U. Parikh, A. Cillo, *et al.*, Clin. Infect. Dis. **28**, ciab072 (2021).
[30] S. Kemp, D. Collier, R. Datir, I. Ferreira, S. Gayed, A. Jahun, *et al.*, Nature **592**, 277 (2021).
[31] B. Jackson, M. Boni, M. Bull, A. Colleran, R. Colquhoun, A. Darby, *et al.*, Cell **184**, 5179 (2021).
[32] D. VanInsberghe, A. S. Neish, A. C. Lowen, and K. Koelle, Virus Evol. **7**, veab059 (2021).
[33] H.-L. Zeng, E. Mauri, V. Dichio, S. Cocco, R. Monasson, and E. Aurell, J. Stat. Mech. Theory Exp. **2021**, 083501 (2021).
[34] R. A. Neher, M. Vucelja, M. Mezard, and B. I. Shraiman, J. Stat. Mech. Theory Exp. **2013**, P01008 (2013).
[35] V. Dichio, H.-L. Zeng, and E. Aurell, Statistical genetics and direct coupling analysis in and out of quasi-linkage equilibrium (2021), arXiv:2105.01428 [q-bio.PE].
[36] H.-L. Zeng, V. Dichio, E. Rodríguez Horta, K. Thorell, and E. Aurell, Proc. Natl. Acad. Sci. **117**, 31519 (2020).
[37] E. Cresswell-Clayand V. Periwal, Math. Biosci. **341**, 108678 (2021).
[38] E. R. Hortaand M. Weigt, PLoS Comput. Biol. **17**, e1008957 (2021).
[39] R. Duerr, D. Dimartino, C. Marier, P. Zappile, S. Levine, F. François, *et al.*, medRxiv 10.1101/2021.12.07.21267431 (2021).
[40] N. Rochman, Y. Wolf, G. Faure, P. Mutz, F. Zhang, and E. Koonin, Proc. Natl. Acad. Sci. **118**, e2104241118 (2021).
[41] N. Rochman, G. Faure, Y. Wolf, P. Freddolino, F. Zhang, E. Koonin, and M. Diamond, mBio **13**, e00135 (2022).
[42] J. Rodriguez-Rivas, G. Croce, M. Muscat, and M. Weigt, Proc. Natl. Acad. Sci. **119**, e2113118119 (2022).
[43] K. Kalia, G. Saberwal, and G. Sharma, Nat. Biotechnol. **39**, 1058 (2021).
[44] K. Katoh, J. Rozewicki, and K. D. Yamada, Briefings in Bioinformatics **20**, 1160 (2017), `https://mafft.cbrc.jp/alignment/server/`.
[45] S. Kuraku, C. M. Zmasek, O. Nishimura, and K. Katoh, Nucleic Acids Res. **41**, W22 (2013).
[46] J. Chen, B. Malone, E. Llewellyn, M. Grasso, P. Shelton, P. Olinares, *et al.*, Cell **182**, 1560 (2020).
[47] H.-L. Zeng, hlzeng/Filtered_MSA_SARS_CoV_2, Github (2020), `https://github.com/hlzeng/Filtered_MSA_SARS_CoV_2`.
[48] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Rep. Prog. Phys. **81**, 032601 (2018).
[49] R. Horta, P. Barrat-Charlaix, and M. Weigt, Entropy **21**, 1 (2019).
[50] C.-Y. Gao, F. Cecconi, A. Vulpiani, H.-J. Zhou, and E. Aurell, Phys. Biol. **16**, 026002 (2019).
[51] M. J. Wainwrightand M. I. Jordan, Found. Trends Mach. Learn. **1**, 1 (2008).
[52] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. **106**, 67 (2009).
[53] E. Aurelland M. Ekeberg, Phys. Rev. Lett. **108**, 090201 (2012).
[54] H. C. Nguyen, R. Zecchina, and J. Berg, Adv. Phys. **66**, 197 (2017).
[55] J. Besag, J. R. Stat. Soc. Ser. D. Stat. **24**, 179 (1975).
[56] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, Ann. Stat. **38**, 1287 (2010).
[57] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, Phys. Rev. E **87**, 012707 (2013).
[58] M. Ekeberg, T. Hartonen, and E. Aurell, J. Comput. Phys. **276**, 341 (2014).
[59] C.-Y. Gao, gaochenyi/cc-plm, Github (2018), `http://`

github.com/gaochenyi/CC-PLM.
[60] M. Skwark, N. Croucher, S. Puranen, C. Chewapreecha, M. Pesonen, Y. Y. Xu, *et al.*, PLos Genet. **13**, e1006508 (2017).
[61] B. Schubert, R. Maddamsetti, J. Nyman, M. R. Farhat, and D. S. Marks, Nat. Microbiol. **4**, 328 (2019).
[62] M. Chand, S. Hopkins, G. Dabrera, C. Achison, W. Barclay, N. Ferguson, *et al.*, SARS-CoV-2 variants of concern and variants under investigation in England Technical briefing 29, Public Health England (2020).
[63] M. Chand, S. Hopkins, G. Dabrera, C. Achison, W. Barclay, N. Ferguson, *et al.*, Investigation of SARS-CoV-2 variants of concern in England, Public Health England (2021).
[64] H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, *et al.*, Nature **592**, 438443 (2021).
[65] Tracking of variants (2021), April 26, 2021. Retrieved 20 August 2021.
[66] H.-L. Zeng, Y. Liu, V. Dichio, K. Thorell, R. Nordn, and E. Aurell, Mutation frequency time series reveal complex mixtures of clones in the world-wide sars-cov-2 viral population (2021), arXiv:2109.02962 [q-bio.PE].
[67] M. Chand, S. Hopkins, G. Dabrera, C. Achison, W. Barclay, N. Ferguson, *et al.*, Investigation of novel SARS-CoV-2 Variant of Concern 202112/01, UK Health Security Agency (2021).
[68] A. West, J. Wertheim, J. Wang, T. Vasylyeva, J. Havens, M. Chowdhury, *et al.*, Nat. Commun. **12**, 4886 (2021).
[69] L. Verdurme, G. Danesh, S. Trombert-Paolantoni, M. Sofonea, V. Noel, V. Foulongne, *et al.*, medRxiv 10.1101/2021.09.13.21263371 (2021).
[70] E. K. Rono, bioRxiv 10.1101/2021.10.08.463334 (2021).
[71] P. Resende, T. Gräf, A. C. Paixão, L. Appolinario, R. Lopes, A. Mendona, *et al.*, Viruses **13**, 724 (2021).
[72] G. Klink, K. Safina, E. Nabieva, N. Shvyrev, S. Garushyants, E. Alekseeva, *et al.*, Virus Evol. **8**, 10.1093/ve/veac017 (2022), veac017.
[73] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc.Natl. Acad. Sci. **108**, E1293 (2011).
[74] T. Hopf, L. Colwell, R. Sheridan, B. Rost, C. Sander, and D. Marks, Cell **149**, 1607 (2012).
[75] L. Vigué, G. Croce, M. Petitjean, E. Ruppé, O. Tenaillon, and M. Weigt, bioRxiv 10.1101/2022.01.21.477185 (2022).
[76] A. Ferguson, J. K. Mann, S. Omarjee, T. Ndungu, B. Walker, and A. K. Chakraborty, Immunity **38**, 606 (2013).
[77] K. Shekhar, C. F. Ruberman, A. L. Ferguson, J. P. Barton, M. Kardar, and A. K. Chakraborty, Phys. Rev. E **88**, 062705 (2013).
[78] T. C. Butler, J. P. Barton, M. Kardar, and A. K. Chakraborty, Phys. Rev. E **93**, 022412 (2016).
[79] A. K. Chakrabortyand J. P. Barton, Rep. Prog. Phys. **80**, 032601 (2017).
[80] R. H. Louie, K. J. Kaczorowski, J. P. Barton, A. K. Chakraborty, and M. R. McKay, Proc. Natl. Acad. Sci. **115**, E564 (2018).
[81] J. P. Barton, E. Rajkoomar, J. K. Mann, D. K. Murakowski, M. Toyoda, M. Mahiti, P. Mwimanzi, T. Ueno, A. K. Chakraborty, and T. Ndungu, Virus evolution **5**, vez029 (2019).
[82] D. K. Murakowski, J. P. Barton, L. Peter, A. Chandrashekar, E. Bondzie, A. Gao, D. H. Barouch, and A. K. Chakraborty, Proc. Natl. Acad. Sci. **118**, e2022496118 (2021).
[83] A. A. Quadeer, J. P. Barton, A. K. Chakraborty, and M. R. McKay, Nature communications **11**, 1 (2020).
[84] E. R. Horta, A. Lage-Castellanos, M. Weigt, and P. Barrat-Charlaix, Journal of Statistical Mechanics: Theory and Experiment **2021**, 073501 (2021).
[85] G. Pedruzziand I. M. Rouzine, PLOS ONE **14**, 1 (2019).
[86] G. Pedruzziand I. M. Rouzine, PLOS Pathogens **17**, 1 (2021).
[87] D. Kern, B. Sorum, S. Mali, C. Hoel, S. Sridharan, J. P. Remis, *et al.*, Nat. Struct. Mol. Biol. **28**, 573 (2021).
[88] W. Lu, B.-J. Zheng, K. Xu, W. Schwarz, L. Du, C. Wong, *et al.*, Proc. Natl. Acad. Sci. **103**, 12540 (2006).
[89] K.-L. Siu, K.-S. Yuen, C. C. no Rodriguez, Z.-W. Ye, M.-L. Yeung, S.-Y. Fung, *et al.*, FASEB J. **33**, 8865 (2019).
[90] Y. Ren, T. Shu, D. Wu, J. Mu, C. Wang, M. Huang, *et al.*, Cell. Mol. Immunol. **17**, 881 (2020).
[91] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, *et al.*, Genome Res. **19**, 1639 (2009).