ARTICLE | Article in Press

Epistasis detection and fitness valleys in SARS-CoV-2 evolution

I. V. Likhachev · I. M. Rouzine

Received: 20 May 2026 / Accepted: 3 November 2026
© The Author(s) 2026

Abstract
Probabilistic prognosis of virus evolution, vital for the design of effective vaccines and antiviral drugs, requires the knowledge of adaptive landscape including epistatic interactions. Although epistatic interactions can, in principle, be inferred from abundant sequencing data by various methods, it has been shown previously that genetic linkage between evolving sites obscures their signature and requires averaging over several independent populations. Here we probe the limits of epistasis detection based on pairwise correlations conditioned on the state of a third site on synthetic sequences evolved in a Monte Carlo algorithm with known epistatic pairs in two scenarios: an initially diverse population in the presence of recombination and mutation, with a uniform epistatic network, and an initially–uniform population in the presence of mutation and a fitness valley. In either scenario, the detection error decreases with the number of independent populations and increases with the sequence length. In the first scenario, the accuracy is enhanced by moderate recombination and is maximal, when epistasis magnitude approaches the point of full compensation. In the second scenario, simulation predicts the spontaneous crossing of a fitness valley and forming a new large strain. The method is then applied to several thousands of sequences of SARS-CoV-2. Results obtained under equal sampling from world regions are consistent with the existence of fitness valleys connecting groups of viral variants.

Keywords Epistasis · Epistasis detection · Genetic linkage · Respiratory virus · Virus evolution · Mathematical model

Introduction

Elucidating the connection between genotype and phenotype remains a challenging problem. Most traits in organisms are coded by several genomic sites, whose combined effect on the phenotype is not a sum of their effects, but also includes site-site interaction termed “epistasis”. Epistasis is responsible for a significant fraction of genetic inheritance[1,2]. In pathogens, epistasis facilitates the development of drug resistance and immune escape and impedes reversion of drug-resistant mutations[3–7]. Mutations in epitopes assisted by epistasis help a pathogen to escape the cytotoxic immune response in the host, when viral replication capacity lowered by an escape mutation

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



https://doi.org/10.1038/s41598-026-79846-8

Published online: 12 November 2026

is gradually restored later by compensatory mutations[8–11]. This process, termed fitness valley, is thought to be responsible for the high diversity and rapid HIV evolution in untreated patients[12].

Epistasis is also implicated in the evolution of variants of SARS-CoV-2[13]. Some mutations in neutralizing antibody-binding regions that help SARS-CoV-2 to escape immune recognition during transmission[14–18] negatively affect the function of receptor binding, and have to be compensated elsewhere[19]. The tradeoff between antibody escape and transmissibility has been inferred in immunocompromised patients[20]. The same is likely true for mutations conferring escape from innate responses to SARS-CoV-2[21,22].

While pairwise epistasis can be measured from the free energy of binding[23] and in culture[24], these results cannot be transferred directly to the biological levels of an organism or population, where epistasis is defined with respect to fitness, i.e. the reproduction number at the within-host or between-host levels. Therefore, a large number of approaches has been proposed to measure epistasis directly from genomic data[25–27]. A popular method of "quasi-linkage equilibrium" assumes that recombination is rapid compared to allele frequency changes, and that population is large, so that recombination and the natural selection with epistasis are the dominant evolutionary factors, while the effects of genetic linkage are relatively small[28–31]. However, none of these methods enable reliable measurement of epistasis in asexual populations or in sexual populations at close loci[32].

About a century ago, it was realized that the evolution of a population is strongly affected by the fact that the fates of alleles at different loci are linked unless separated by recombination[33]. Empirically, the effect of strong linkage between genomic sites has been discovered by Thomas Hunt Morgan who used the inter-dependent segregation of alleles in fruit flies to understand the role of chromosomes and sex in inheritance. He showed that some pairs of traits of fruit flies segregate in a correlated fashion, and some in independent fashion. He postulated that the first are close on the same chain of genes, so that recombination cannot split them. Later, the nature of this chain was shown to be DNA in chromosomes.

Since then, a large body of experiment and modeling demonstrated that linkage changes evolutionary properties drastically. In the absence of recombination, common phylogeny makes the sequences of each population relatively close to their most recent common ancestor, as compared to the other independently-evolving populations of the same origin. This is because the time to the ancestor of each population is more or less constant, while independent populations are diverging from each other in time[34]. Moreover, even very frequent recombination does not break up linkage effects in a sufficiently long genome[35]. As a result, any measure of co-variance produces strong noise of random sign. An attempt at detecting epistasis, whether by using covariance measures, such as Pearson coefficient, mutual entropy, universal footprint[36], or by tree-based methods[37,38] faces the same problem, the overwhelming genetic linkage effects.

The only way to resolve this issue and separate linkage (clonal) effects from epistasis is to average the haplotype frequencies over many independent populations evolving under similar conditions[34]. To decrease the required number of populations, it was proposed to use only sequences with a majority allele at a neighbor site of the measured site pair to interrupt a path along sites that creates a false-positive interaction[39]. This trick helps, as long as the system is close to the critical point of epistatic magnitude[39]. The study[39] provided a simulated demonstration and applied it to influenza sequence data to explain the emergence of the 2009 pandemic strain.

In the present work, this method is tested on simulated sequences in a broad range of parameters, for different scenarios of epistatic networks, and different initial states. The accuracy of epistasis detection against the background of genetic linkage (clonal) effects is tested on sets of simulated sequences with known epistasis[39]. The sequences are generated by Monte Carlo simulation including mutation, natural selection, and recombination. Then, the method is applied to real virus sequences from the adapting world population of SARS–CoV-2 to predict the strongest epistatic links.

These results are related to another important question: how did strain B.1.1.7 and other variants of concern of SARS-CoV-2 evolve, with so many mutations emerging at once? What are the evolutionary factors that brought it to existence? A computer simulation is used to demonstrate that the variant of concern could emerge due to

the spontaneous crossing of a fitness valley. Although the simulation does not constitute decisive proof, it offers evidence in favor of the hypothesis.

This approach differs from the alternative methods in the following respects. Firstly, it does not make unrealistic assumptions regarding the evolutionary factors at play (does not neglect linkage). Secondly, its accuracy is validated in a broad range of model parameters on simulated sequences, using a generic model of population genetics of the Wright Fisher type for different evolutionary scenarios: initial allele frequency $f_0 = 0.45$ and 0, in the presence of recombination and without, with periodic and star-like epistatic network. We demonstrate the accuracy on simulated sequences under conditions resembling the evolution of SARS CoV-2 with the focus on the emergence of variants of concern. Thirdly, the application to data is focused on well-balanced sampling between sequences related to major variants. The fourth feature is the use of the conditional correlator.

Results

Detection method

The method of epistasis detection[39] is based on a specific measure of allelic correlations (UFE) designed to separate epistatic magnitude from other model parameters, and isolate epistatic signature from clonal effects[36]. To average out random linkage effects, the balanced averaging of haplotype frequencies over independent populations is performed. In simulation, the detection error is averaged over Monte Carlo runs, which represents the statistical test necessary and sufficient in this context. Binarized sequences are used as the initial genomic data for testing the method. Parallel computing on more than 10,000 processors of GPU is employed to increase performance. To improve the detection accuracy near the point of full compensation, an additional correlator conditioned on an allele at a third locus is calculated. The details are described in *Methods.*

Accuracy test 1. The case of standing variation

To test the accuracy of the detection method, artificial genomic sequences are obtained from the simulation of the multi-site evolution of a genetically-diverse population by a Monte-Carlo algorithm with a set epistasis network (*Methods*). The input parameters are chosen in a range representative for several viruses (Table 1 in *Methods*). The predicted links are compared with the true links set in the algorithm. The error of detection is defined as the number of false-positive and false-negative interactions among predicted interactions divided by the number of true interactions. The error is calculated as a function of several parameters: genome length (total locus number), the number of independent populations used for averaging haplotype frequencies, the magnitude of epistasis, recombination probability, and the average crossover number (Fig. 1).

As expected, the accuracy of epistasis detection becomes higher when the number of independent populations used for averaging and the magnitude of epistasis increase. The accuracy decreases with the genome length. While modest recombination significantly improves the accuracy of detection by compensating linkage effects eclipsing epistasis, very frequent recombination counteracts associations of alleles caused by epistasis and so impairs detection (Fig. 1). Analogous procedure was carried out using Pearson coefficient and UFE[36] in the absence of recombination (*Supplement*, Figs. S2-S4).

Accuracy test 2: initially-monomorphous populations with a fitness valley

Pandemics of SARS-CoV-2 started from a single Wuhan strain. To describe this scenario, we carried out an additional simulation test with an initially-monomorphous virus, $f_0 = 0$. The epistatic network was changed to a single star comprising one deleterious allele and 10 rays with compensatory alleles. These mutations are deleterious separately, but the combination of the primary allele with the compensatory alleles is advantageous. Among

Table 1 Parameters of the model and data processing.

Notation	Value	Description
f_0	0 Figs. 2, 3 and 4 0.45 Fig. 1	Initial proportion of mutant alleles in the population
N	1000	Population size
μL	0.07 Fig. 1 0.1 Figs. 2, 3 and 4	Mutation rate per genome
t	30 Fig. 1 300 Figs. 2, 3 and 4	Simulation time in generations
s_0	0.1	Selection coefficient (the same for all sites)
r	0, 0.1, 0.5, or 1	Recombination probability
M	1,3, or 10	Number of crossovers
E	0.4,0.5,0.6,0.75 Fig. 1 0.4–4 Figs. 2, 3 and 4	Strength of epistatic interaction
run	6,20,50 or 200	Number of simulation runs used to average the data
L (seqlen)	20,40,100 or 200	Length of the simulated sequence
Sequence threshold	0.7	Maximum proportion of gaps and non-reads in one sequence
Gap threshold	0.5	Maximum proportion of gaps and non-reads at one site
Monomorphism threshold	0.8 or 0.95	Maximum degree of monomorphism per site

deleterious alleles, one is set to be the primary mutation of fitness valley and 10 are set to be compensatory mutations with the degrees of compensation ranging from 80 to 800%. The specific choice of parameters is a representative example, in which spontaneous crossing of the fitness valley could occur to resemble emergence of VOC. The value of selection coefficient is chosen to fit the timing of variant B.1.1.7

Initially, the population is monomorphous. Then a gradual accumulation of beneficial alleles occurs, which are assumed to take a 10% of loci (Fig. 2). Deleterious alleles also accumulate at the remaining 90% of loci due to the hitchhiking effect, i.e., genetic linkage with the beneficial alleles, although more slowly than beneficial alleles. Sooner or later, the primary mutation becomes abundant in a population due to the genetic hitchhiking effect, i.e., due to its linkage to advantageous alleles. This change makes the compensatory alleles advantageous, and they start to accumulate rapidly (Fig. 2). This events corresponds to crossing the fitness valley and the emergence of a new group of virus variants on its other side.

As a result of the virus passing the fitness valley, the mutational spectrum averaged over 6 populations at 150 generational steps has two maxima (Fig. 3). The sequences are divided between the two peaks with borders chosen at 6 (the local minimum) and 22 new alleles (upper cutoff). Then a bootstrapping procedure is conducted, in which sequences from each peak are oversampled many times with replacement and repetition. The resulting sequence sets are used for the epistasis detection method described above.

The detection error is calculated as the sum of false positive and false negative epistatic interactions out of 10 possible (Fig. 4), as a function of numbers of sequences sampled from peak 1 and 2, normalized to the sequence number of each peak. Thus, about a half of weaker interactions is not detected.

The number of false positive interactions remains zero, so that the total error is given by false negative interactions. Therefore, this method is highly accurate, but only in the conservative direction. Our results suggest the need to take either equal sampling from the two peaks or somewhat smaller sampling from the earlier peak (Peak I), as long as they are approximately equal in size, as in our example.

Results on epistasis detection in genomic data for SARS CoV-2 are consistent with fitness valleys

After testing the method on simulated sequences, we applied it to sequence data on SARS-CoV-2 for two different data sets, as follows.

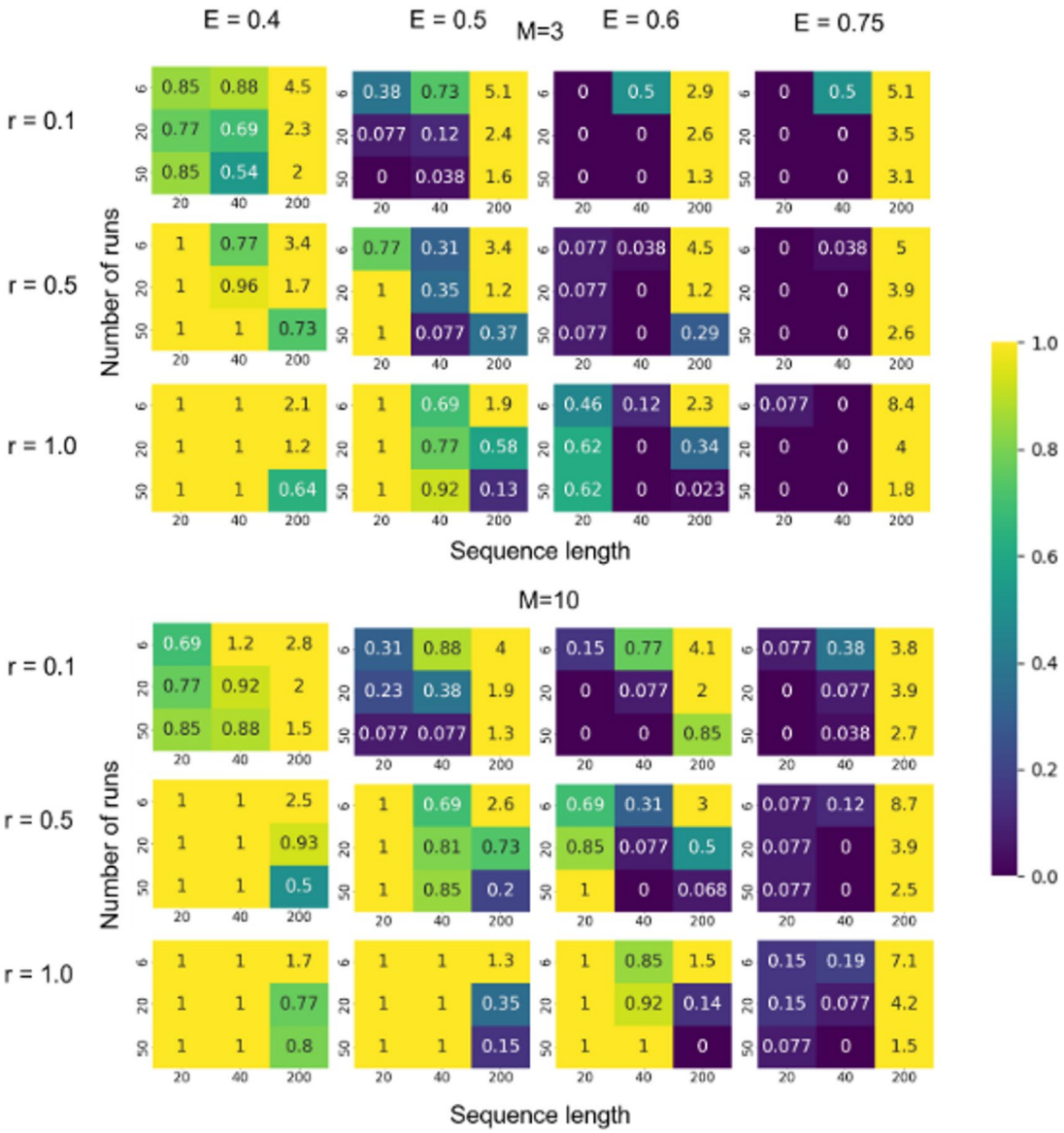


Fig. 1 The error of detection of epistatic interactions for the case of standing variation (Test1). The length of the sequence is plotted on the X–axis, and the number of runs is plotted on the Y-axis of each panel. The values inside the squares defined by color map on the right correspond to the sum of the false positive and false negative interactions divided by the number of true interactions. Detection is carried out using the universal footprint of epistasis (UFE)[36] and three-site haplotype method[39](*Methods*). All values greater than 1 are set to have the same red color as values equal to 1. Testing of the method was conducted on the sequences generated by Monte Carlo simulation (*Methods*) with variable parameters (shown): probability of recombination r, number of crossovers M, strength of the epistatic interaction E. The topology of the epistatic links (interactions) comprises isolated double arches, for which $E = 0.75$ is the threshold value of full compensation[36]. Fixed parameters: initial allele frequency $f_0 = 0.45$, population size $N = 1000$, mutation rate per genome $\mu L = 0.07$, simulation time in generations $t = 30$, selection coefficient $s = 0.1$. Results for $M = 1$ are shown in Figure S1. Variant without recombination is shown in Figure S2. Results using Pearson coefficient, with and without recombination, are in Figs. S3 and S4.

Dataset 1 includes 6000 sequences from 6 regions, 1000 from each: Africa, Asia, North America, Oceania, South America, for the period of one year, 09/24/2020 to 09/25/2021. When each sequence is compared to Wuhan variant, a mutation spectrum is obtained, where sequences segregate naturally into three groups, with gaps between them (Fig. 5).

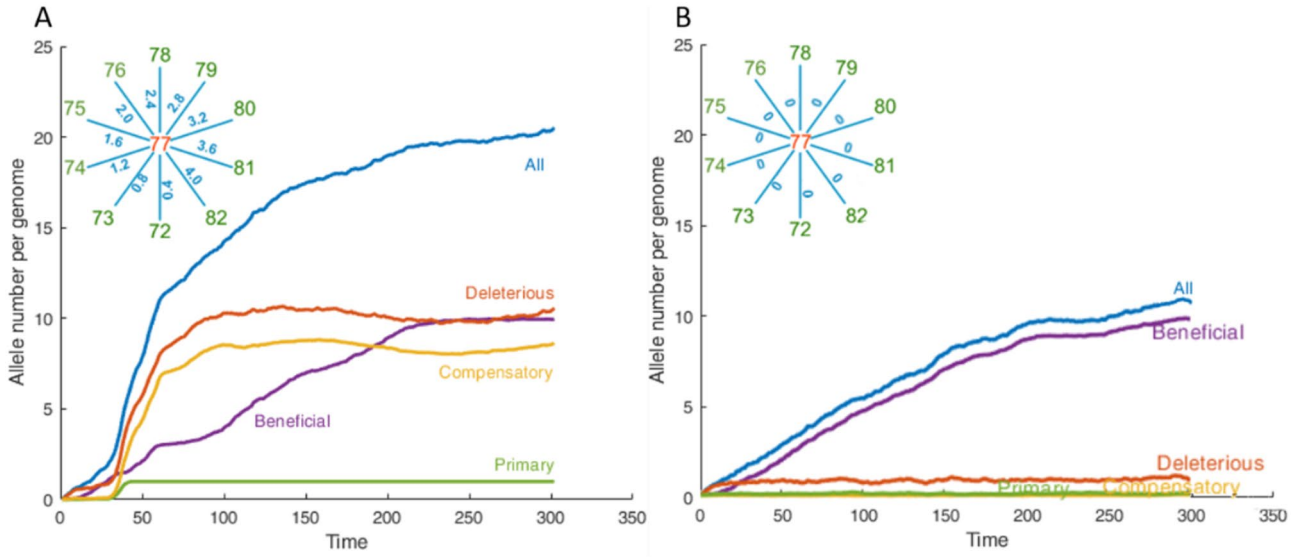


Fig. 2 Monte Carlo simulation of the evolutionary dynamics of a population with a fitness valley. X-axis: Time in generations. Y axis: Average numbers of the alleles of different classes (shown) per genome. Different curves correspond to alleles at different types of sites (shown). (**A**) The case of fitness valley created by a star-like epistatic network. Degree of compensation is shown by blue numbers. (**B**) Negative control experiment without epistasis. Parameters: initial allele frequency $f_0 = 0$, population size $N = 1000$, mutation rate per genome $\mu L = 0.1$,10% of the alleles are beneficial, 90% are deleterious with selection coefficient $s = 0.1$.

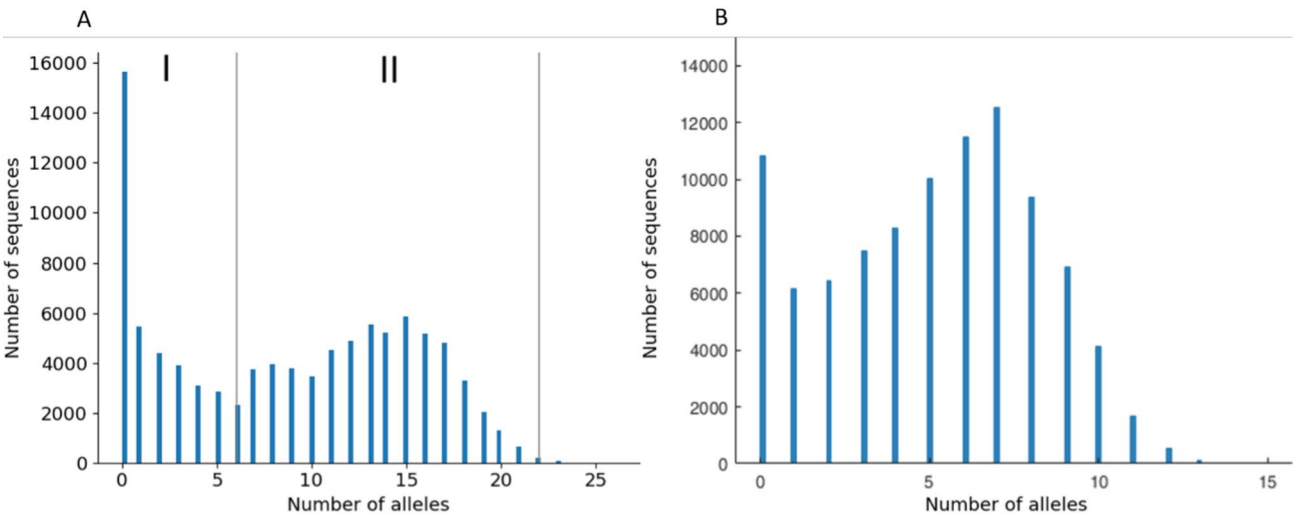


Fig. 3 Mutational spectrum for six simulated populations, initial state is monomorphous. Y–axis: number of sequences with a number of replacements shown in X–axis. (**A**) Epistasis is present. (**B**) Epistasis is absent. Parameters as in Fig. 2. Six Monte Carlo runs at timepoints multiple of 10 until time 150 are used for sampling.

Figure 6 shows the inferred epistatic network from groups I and II. To obtain this diagram groups I and II were oversampled 10 times with replacement. As it was shown in Fig. 4 equal sampling is essential for obtaining reliable results. This procedure was also repeated for sequences mixed from groups I and III, where groups I and III were resampled 10 and 20 times respectively since group III almost two times smaller than group I (Fig. S5).

The detected network (Fig. 6) is consistent with the presence of fitness valleys, although does not prove their existence directly, presumably, due to the absence of weaker links not captured by the method. The method is

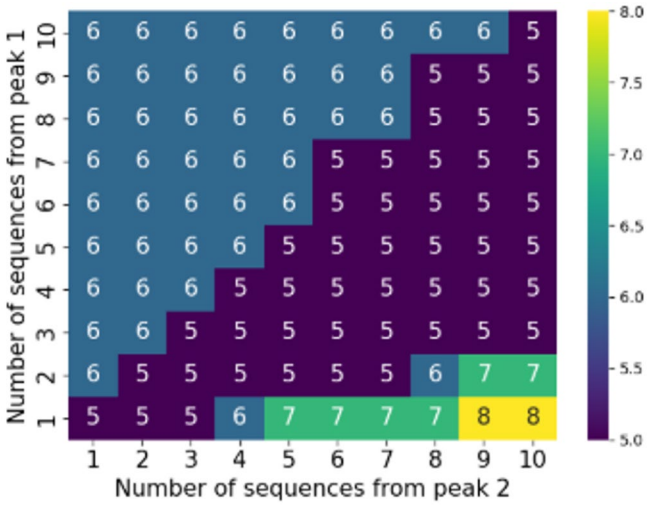


Fig. 4 Error of epistatic detection for six simulated populations that are initially monomorphous. Axes X and Y show the rescaled numbers of sequences sampled from parts I and II of the dataset in Fig. 3, respectively, relative to the total number of sequences in each part. Numbers in cells show the number of false-negative epistatic interactions out of 10 possible. The measured number of false–positives is always zero. Parameters are as in Fig. 2. Minimum sample size is approximately 1000 good sequences per one population, minimal population number is six.

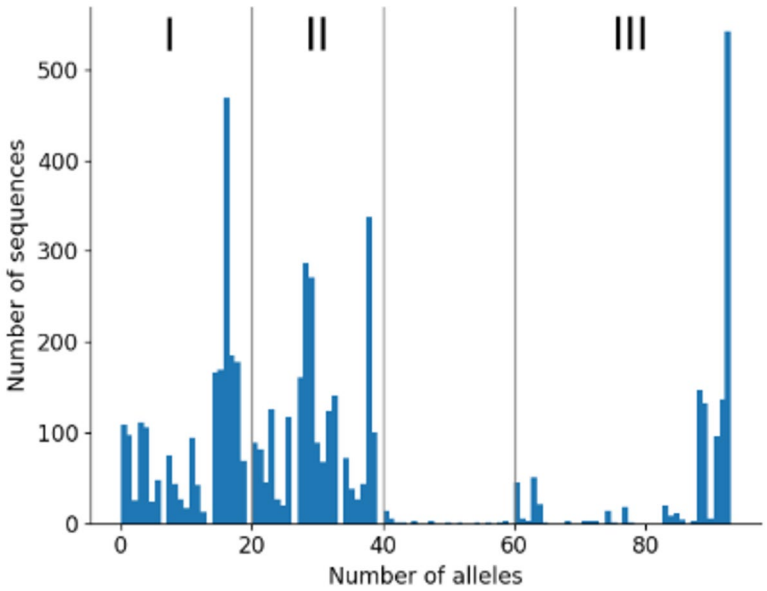


Fig. 5 Mutational spectrum for Dataset 1. Y–axis: number of sequences with the number of differences shown in X–axis from the consensus sequence, defined as Wuhan variant. Gray lines at 20, 440 and 60 divide the three groups of sequences numbered by the Roman numerals.

very conservative, and its accuracy at small epistatic magnitudes is low (Figs. 1 and 4 and the legend to Fig. 4). Hence the epistastic interactions in Figs. 6 are the strongest in possible fitness valleys.

For sequences sampled from groups II and III (Fig. 5), no significant epistatic links are found, indicating the absence of a direct passage between them. The results we obtain are consistent with the existence of several fitness valleys connecting large groups of virus variants.

We repeated the same procedure for Dataset 2 comprising 20,000 random sequences from the NCBI Virus database for the same period of time as Dataset 1, from 09/24/2020 to 09/25/2021, but without geographic classification nor equal sampling from continents. Results differ strongly from those for Dataset 1, as follows. The number of interactions is smaller, and they are detected only for groups I and II, but not any other pairs of the four groups (Fig S6). Thus, equal sampling from world regions is critical for detection.

Biological function of the predicted interacting sites

Mutations E484K and N501Y are well-known mutations of the receptor-binding domain of the coronavirus in Spike. The first is responsible for the viral escape from the immune response and is capable of reducing the

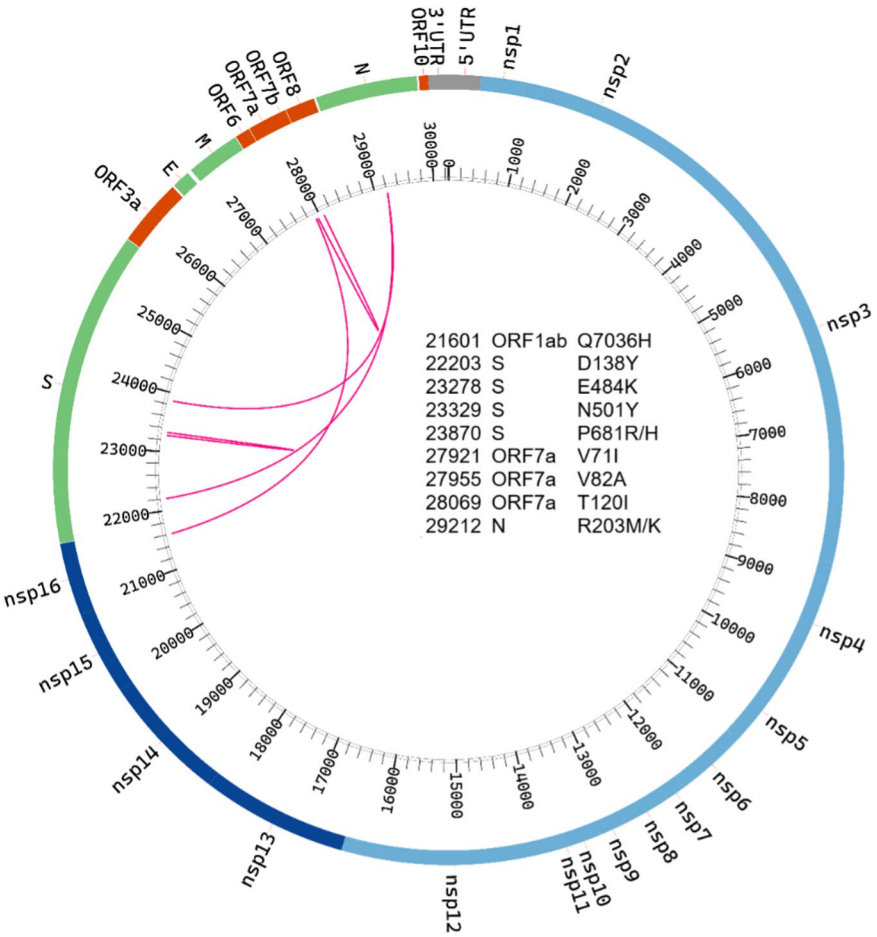


Fig. 6 Epistatic interactions for SARS-CoV-2 genome predicted for Dataset 1 for sequence groups I and II (Fig. 5). To obtain the circular diagram using software Circos[40], each group was oversampled 10 times with replacement. The interacting pairs are shown.

affinity of antibodies produced to earlier strains by 1–2 orders of magnitude, while the second improves the binding of coronavirus to the human ACE2 receptor on cell surface and increases viral replication in upper respiratory tract cells[41].

Mutations P681H and P681R, most often found in Alpha (B.1.1.7) and Delta (B.1.617.2) variants of the virus, respectively, are located in the furin cleavage region of Spike. They both confer a change in the amino acids from nonpolar hydrophobic to positively charge. They affect tropism (repertoire of infectable cells) by increasing the cleavage of S1/S2 proteins in human airway epithelial cells[42].

Mutation D138Y in Spike is believed to affect the protein's secondary structure thus reducing antibody -binding efficiency[43]. The R203K mutation in the N protein was shown to increase transmission and virulence of SARS-COV-2[44], while the R203M mutation, which creates a start codon for a new open reading frame, was reported to suppress interferon induction, thereby increasing viral infectivity[45].

V82A and T120I in ORF7a are most frequently found in variant Delta of the coronavirus[46]. They are located within the di-lysine motif of ORF7a, which is important for the localization of viral proteins in endoplasmic reticulum. Existing data on mutation Q7036H in protein ORF1ab and mutation V71I in ORF7a do not permit to determine their functions definitively; however, their location within viral non-structural proteins hint at their interaction with the immune system of the host.

Discussion

Detection and measurement of epistasis from genetic data represents a challenge. In our work, we have upgraded a linkage-resistant method of detection and applied it to sequences from SARS-CoV-2. Results are consistent with the existence of fitness valleys that provide passages between large groups of virus variants.

As mentioned before, a fitness valley comprises a transient decrease in fitness due to initial mutation(s) forced by changes in the environment, which is followed by compensatory mutations. The fitness valley effect was previously observed or inferred for various viruses, including HIV[47,48] and influenza[37–39], and studied using mathematical models[8,39]. For example, a primary mutation can help a virus to escape the immune response, adapt to a species or a tissue, or expand the pH range in which it can replicate. Mutations conferring these benefits to a virus, but reducing its overall fitness, are followed by compensatory mutations that elevate its fitness above the initial level.

The loss of fitness due to escape mutations, and the partial loss of recognition by the immune system, together determine the order, the number, and the evolution rates of escape mutations in HIV infection[49,50]. Compensatory mutations prevent the reversion of escape mutations after the decay of CD8 T cell clones against escaped epitopes. Bacteria developing antibiotic resistance also pass through such fitness valleys[5]. A population-level simulation study proposed this effect in conjunction with the evolution in immunocompromised patients infected with SARS-CoV-2[51]. Another modeling work proposed crossing a deep fitness valley using recombination[52].

The fates of alleles at different loci are linked due to common phylogeny unless separated by recombination[33]. For this reason, one-locus models usually do not apply in real populations, and multi-locus models must be used instead[53,54]. These models, based on traveling-wave technique, have been successfully applied to rapidly-evolving viruses, such as HIV[55] and influenza[56–58]. Linkage effects include clonal interference[33,59], background selection, genetic hitchhiking[60], enhanced accumulation of deleterious alleles (Muller's ratchet)[61,62], and the increase of genetic drift at one locus due to selection at another[63]. Multi-site adaptation has been analyzed analytically in various asexual models, with the result that linkage effects strongly decrease the adaptation rate compared to the limit of independently-evolving sites[62,64–66]. Importantly, genetic linkage also creates random associations between pairs of mutations occurring on the same branch of the ancestral tree, which is a serious obstacle to the detection of epistatic interactions in genetic data[32]. In a single asexual population, stochastic linkage completely overshadows the epistatic footprint, except in a narrow range of times and parameters[34]. The same limitation exists for the tree-based methods of detection[37,38].

For testing the accuracy, we used a model with directional selection. In fact, SARS-CoV–2 evolves in antibody–binding regions due to collective immunity of population, which is a case of the selection for change (Red Queen effect). Nevertheless, as it has been shown by several groups[56–58], the multi-locus epidemiological model of respiratory viruses is effectively reduced to a Wright-Fisher model with constant directional selection. The selection coefficient is expressed in terms of basic reproduction ratio in a naïve population and the breadth of antibody neutralization, calculated in these works using different approximations.

The most popular method for searching for epistasis signature in genomic data is the "quasi linkage equilibrium" approach[28,30], which assumes frequent recombination compared to changes in allelic frequency and a large population size. In this limit, genetic linkage effects are negligible compared to epistasis and recombination. In the present work, we do not rely on QLE approximation, so that the method works for finite population size and for modest recombination or even in its absence. The linkage effects are mostly cancelled out by averaging over well-balanced independent populations and the use of 3-site haplotypes. Such cancellation was not carried out in[68], hence the abundance of false negative links (Fig. 7C) compared to our results (Fig. 7A).

To address the problem of false-negative bonds, another team[67] compared the sequences of SARS CoV-2 not between themselves, but with other coronaviruses. They used "direct coupling analysis", which postulates a statistical distribution of sequences (usually, of a Gaussian form) including individual effects of alleles and their pairwise interaction. A physicist would recognize here the Gibbs distribution for the Onsager model of

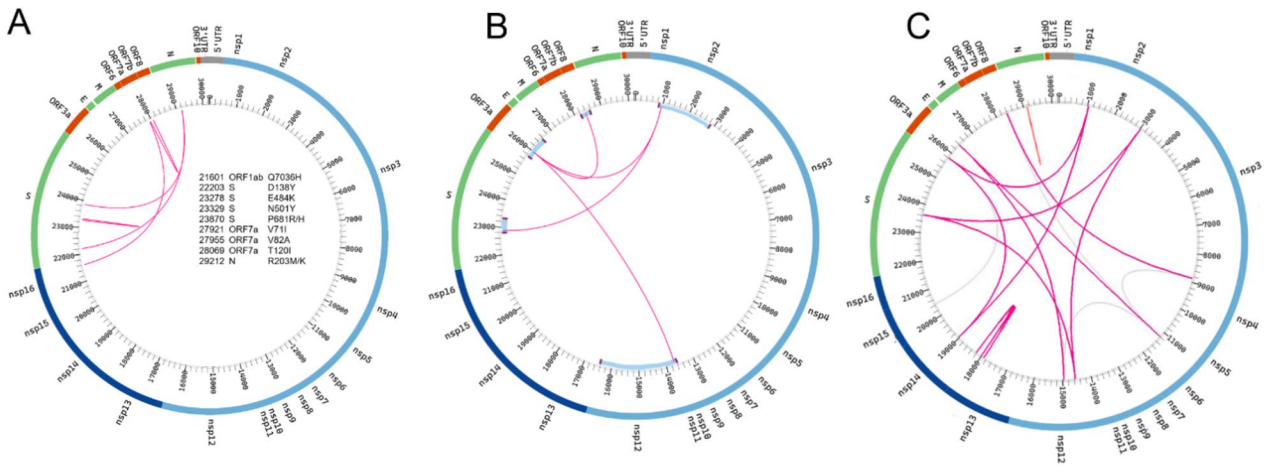


Fig. 7 Epistatic interactions for SARS-CoV-2 genome: comparison of the present results with two previous studies. (**A**) Results of the present work for Dataset 1 from Fig. 6 obtained from mixing sequence groups I and II in Fig. 5. (**B**, **C**) Results from studies[67] and[68], respectively. Blue stripes in panel B represent protein domains reported to interact with each other in[67].

ferromagnetism. In population genetics, this Gaussian form has a justification in the limit of QLE[28,30]. Parameters of the Gaussian are fit to sequence data using various approximations[67]. DCA also includes down-weighing sequences with over 80% identity, which is a phenomenological trick to deal with the effects of random linkage. As result, the number of false-positive epistatic bonds due to genetic linkage is sharply decreased (Fig. 7B). The problem here is that the nature of such correction is not derived from a model of population genetics. QLE does not work in SARS CoV-2 evolution, because linkage effects, negligible in QLE limit, are quite strong.

Note that the use of tree-based methods does not automatically compensate for strong linkage effects. This is because, in the absence of sufficiently strong recombination, all sequences in one single population are very similar to their MRCA, and MRCAs for different independent populations are diverging from each other in genetic space. This process was demonstrated to be responsible for strong linkage effects[34]. To exclude the effects of linkage, one group used synonymous sites as markers of linkage[37,38].

It is worth noting that SARS-CoV-2 sequences represent a meta-population with many connected islands, not a single well-mixed population, nor completely independent populations. The approximation used in our work here is that the average over different geographic locations allows to obtain, at least, a partial average over the ensemble of independent populations. It is unclear to which extent the lack of the full ensemble average affects the estimate, but at least, we are able to compensate the residual linkage errors within that partial ensemble, something that has not been done before.

One might ask whether the inferred mutations in Fig. 2A are simply a driving mutation with hitchhiker mutations present in the invading strain due, for example, migration between different continents. As easy to understand, in this case, our method would give zero signal instead of the network inferred (Fig. 2A). To make this fact obvious, consider an extreme example, where the population represents a mix of the uniform old and uniform new strain. We focus on the part of the genome where the two strain differ, for example, 101,100.. and 010011.., respectively. These differences can be either due to driving or hitchhiking mutations. Note that, for any pair of these sites, only two haplotypes out of the four are present. For example, for the first and second sites, we have haplotypes 10 and 01. Hence, when calculating correlation UFE, these pairs are excluded from analysis due to the cutoff condition on the allelic frequency. Thus, by design, our method does not measure the difference between strains, but only a specific type of association between the fluctuations of alleles that is caused by epistasis.

To conclude, the present study confirms the existence of fundamental limitations on epistasis detection due to genetic linkage. Within this limitation, a conditional correlation method combined with balanced sampling from

quasi-independent populations and testing the accuracy in simulation allows the testable detection of strong epistatic links.

Methods

Generating sequences for method testing

Evolution of a haploid asexual population of N binary sequences is simulated by a Monte Carlo code[36,69], as follows. In an individual genome, each locus (site, nucleotide position, amino acid position) numbered $i = 1, 2, \ldots, L$ is occupied by one of two alleles, either the wild-type allele, denoted $K_i = 0$, or the mutant allele, $K_i = 1$.We use a discrete generation scheme in the absence of generation overlap (Wright-Fisher model). The evolutionary factors included in the model are random mutation with rate μL per genome, constant directional selection, random genetic drift due to random sampling of progeny and recombination with rate r. Selection includes an epistatic network with a set strength and topology. The logarithm of the average progeny number of a genome, W, depends on sequence $[K_i]$ as given by[36]

$$W[K_i] = -\sum_{i=1}^{L} s_i K_i + \sum_{i<j}^{L} s_{ij} K_i K_j \tag{1}$$

$$s_{ij} = E_{ij}(s_i + s_j) T_{ij} \tag{2}$$

where $T_{ij} = 0$ or 1 is the binary matrix that shows epistatically-interacting pairs. Here the selection coefficients s_i and s_j denote the individual fitness costs of two deleterious mutations, E_{ij} is the compensation degree at sites i and j..

In our simulation example in Fig. 1, we consider a haploid population with the initial frequency of deleterious alleles $f_0 = 0.45$, mutation rate per genome $\mu L = 0.07$, fixed selection coefficient $s_i \equiv s = 0.1$, population size $N = 1000$ individuals, genome length $L = 40$ sites, and epistatic network is double arches.

In our simulation example in Figs. 2, 3 and 4, we consider a haploid population with the initial frequency of deleterious alleles $f_0 = 0$, mutation rate per genome $\mu L = 0.1$, selection coefficients with a fixed absolute value $s_i \equiv s = 0.1$, where 10% of potential mutations are advantageous and 90% are deleterious, population size $N = 1000$ individuals and genome length $L = 100$ sites, epistatic network – star with deleterious mutation in the middle and 10 compensatory mutations as rays, degree of compensation gradually increases from 80% to 800%. This choice of parameters is a representative example, in which the spontaneous crossing of the fitness valley resembles the emergence of VOC. Variation of the initial allelic frequency has a small effect on the dynamics. The value of selection coefficient is chosen to fit the timing of variant B.1.1.7[70].

The Monte-Carlo simulation code is written in MATLAB™ and deposited at https://github.com/FireFrog17/Epistasis-detection-and-fitness-valleys-in-SARS-CoV-2-evolution-Code.

Processing genomic data of SARS-CoV-2

In Dataset 1, 6000 sequences sampled from 6 regions (Africa, Asia, North America, Oceania, South America) in the period from 09/24/2020 to 09/25/2021 were downloaded from database NCBI Virus. In Dataset 2, we used 20,000 random sequences from the NCBI Virus database for the same period of time as in Dataset 1.

The sequences were not stratified by sampling date for two reasons, as follows. Firstly, the necessary size of statistics would be missing. Also, previous experience with hemagglutinin data[39] demonstrates that combining sequences from different dates is an advantage, because one compares earlier and later strains and thus finds a fitness valley connecting them. In contrast, estimation of selection coefficients, indeed, requires classification by time[71].

Downloaded data were processed by a custom-made sequence-processing program *ProcessGenomes.py*. Multiple alignment was performed in MAFT online, using the "Fast calculation of full-length MSA of closely related viral genomes"[72]. From the total number of downloaded and aligned sequences, sequences in which the percentage of symbols "–" (gap) or "N" (nonread) is greater than *Sequence threshold* are removed. From the remaining sequences, sites in which the percentage of gaps or letters N in across sequences is higher than *Gap threshold* are removed. Then, sites with a polymorphism degree (Hamming distance) lower than *Monomorphism threshold* are removed. We put the monomorphism threshold at two values (Table 1) and make sure that the choice does not affect final results, i.e., results are in a plateau in this parameter. The sequence data are binarized: the variant similar to consensus defined as Wuhan variant is declared 0, the others are 1, with the resulting population matrix K_{ni}.

After binarization, the sequence data of virus undergo an additional processing stage, as follows. Based on the number of differences from sample consensus, a mutational spectrum is constructed using a custom-made sequence processing program *ProcessGenomes.py* (Figs. 3 and 5). Several groups of sequences are identified from this diagram. The new matrix K_{ni} is formed by random sampling from two chosen groups, with the possibility of repetition. The dependence of the predicted number of epistatic links on the sample size for simulated data is shown in Figs. 1 and 4.

Epistasis detection and measurement

Based on matrix K_{ni} obtained as described above, the universal footprint of epistasis[36] denoted UFE_{ij} is calculated for each possible pair of sites i, j as $UFE_{ij} = 1 - \log\left(\frac{f_{11}}{f_{00}}\right)/\log\left(\frac{f_{01}f_{10}}{f_{00}^2}\right)$. Then, the conditional coefficient UFE_{ij0} is recalculated for each locus pair[39], as follows. Only sequences where at least one of the neighboring loci of the pair has allele 0 are used. The minimum of all UFE over such neighbors is denoted UFE_{ij0}. Because the array of data used is large, these coefficients are calculated using parallel computing on the video card with more than 1000 parallel processors with the help of CUDA language. Pairs of loci, where the ratio of UFE_{ij} to UFE_{ij0} is equal to or greater than 0.5, are considered "true pairs", while the remaining pairs are excluded. Results are displayed in the form of circular diagrams (Figs. 6 and S5).

Error calculation

The estimated epistatic pairs are compared with the matrix of true pairs T_{ij} set in the Monte Carlo simulation code that generates genomic sequences for testing the accuracy. The relative error of the method is calculated as the sum of the number of false positive and false negative pairs for an initially–uniform population (Fig. 4). For the case of standing variation, the sum is divided by the total number of true pairs (Fig. 1).

Assumptions and approximations

The method implies the weak selection limit, defined as $\max(1/N,\ \mu) \ll s \ll 1$, and is appropriate in the travelling wave regime, when many mutations at different loci are being fixed gradually and compete with each other for the space in population (clonal interference). The traveling wave regime takes place at $N\mu L \gg 1$ and on time scales longer than $1/s$..

Next, the fluctuations of mutation rate along the genome, which are typically strong for RNA viruses, do not matter much in the multi-site evolution regime, where the mutation rate enters all the observables in the argument of a large logarithm[53].

Recombination, although observed for example for Omicron variants, is relatively rare in SARS-CoV-2, because it requires coinfection of an individual with two genetically remote variants, as estimated before for another virus[69,73]. Multiplicity of infection is incorporated into an effective recombination rate measurable

directly from genomic data[69,74]. The effect of recombination on detection accuracy was tested an initially-diverse population and found to be modest (Fig. 1).

Also, here we consider the population level evolution only, because SARS–CoV-2 is acute and represented by a single variant in 99,5% patients, similar to influenza. In principle, the 0.5% chronic patients could seed new variants, in which case inter-scale evolutionary conflicts are possible[75].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-79846-8.

Acknowledgements Acknowledgements. We express our gratitude to Richard Neher for stimulating comments. We thank A.A. Makashov and S.V. Lagunov for technical help and discussion. We are grateful to D. A. Gorshkova, A. I. Pikhulya, E. I. Stroganova, and E. A. Vigovskaya for technical help.

Author contributions I.V.L.: Conducted research, wrote codes, analysed simulation results, curated and analysed genomic data, prepared figures, wrote MS (first draft)I.M.R.: Designed research, secured funding, scientific supervision, administration, wrote MS (editing).

Funding The study was funded by Russian Science Foundation, grant 24-24-00529.

Data availability Availability of data and software. Codes are available at https://github.com/FireFrog17/Epistasis-detection-and-fitness-valleys-in-SARS-CoV-2-evolution-Code.

Declarations

Competing interests The authors declare no competing interests.

Ethical statement No ethical statement is required.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Carlborg, O., Jacobsson, L., Ahgren, P., Siegel, P. & Andersson, L. Epistasis and the release of genetic variation during long-term selection. *Nat. Genet.* **38**, 418–420. https://doi.org/10.1038/ng1761 (2006).
2. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U S A*. **109**, 1193–1198. https://doi.org/10.1073/pnas.1119675109 (2012).
3. Gonzalez-Ortega, E., Ballana, E., Badia, R., Clotet, B. & Este, J. A. Compensatory mutations rescue the virus replicative capacity of VIRIP-resistant HIV-1. *Antiviral Res.* **92**, 479–483. https://doi.org/10.1016/j.antiviral.2011.10.010 (2011).
4. Handel, A., Regoes, R. R. & Antia, R. The role of compensatory mutations in the emergence of drug resistance. *PLoS Comput. Biol.* **2**, e137. https://doi.org/10.1371/journal.pcbi.0020137 (2006).
5. Levin, B. R., Perrot, V. & Walker, N. Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. *Genetics* **154**, 985–997 (2000).
6. Nijhuis, M. et al. Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy. *AIDS* **13**, 2349–2359 (1999).
7. Piana, S., Carloni, P. & Rothlisberger, U. Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations. *Protein Sci.* **11**, 2393–2402. https://doi.org/10.1110/ps.0206702 (2002).

8. Weissman, D. B., Desai, M. M., Fisher, D. S. & Feldman, M. W. The rate at which asexual populations cross fitness valleys. *Theor. Popul. Biol.* **75**, 286–300. https://doi.org/10.1016/j.tpb.2009.02.006 (2009).
9. Liu, D. et al. Preexisting compensatory amino acids compromise fitness costs of a HIV-1 T cell escape mutation. *Retrovirology* **11**, 101. https://doi.org/10.1186/s12977-014-0101-0 (2014).
10. Crawford, H. et al. Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant HLA-B*5703-restricted Gag epitope in chronic human immunodeficiency virus type 1 infection. *J. Virol.* **81**, 8346–8351. https://doi.org/10.1128/JVI.00465-07 (2007).
11. Song, H. et al. Impact of immune escape mutations on HIV-1 fitness in the context of the cognate transmitted/founder genome. *Retrovirology* **9**, 89. https://doi.org/10.1186/1742-4690-9-89 (2012).
12. Rouzine, I. M. & Coffin, J. M. Search for the mechanism of genetic variation in the pro gene of human immunodeficiency virus. *J. Virol.* **73**, 8167–8178. https://doi.org/10.1128/JVI.73.10.8167-8178.1999 (1999).
13. Rouzine, I. M. Evolutionary Mechanisms of the Emergence of the Variants of Concern of SARS-CoV-2. *Viruses* **17** https://doi.org/10.3390/v17020197 (2025).
14. Carabelli, A. M. et al. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat. Rev. Microbiol.* **21**, 162–177. https://doi.org/10.1038/s41579-022-00841-7 (2023).
15. Shrestha, L. B., Foster, C., Rawlinson, W., Tedla, N. & Bull, R. A. Evolution of the SARS-CoV-2 omicron variants BA.1 to BA.5: Implications for immune escape and transmission. *Rev. Med. Virol.* **32**, e2381. https://doi.org/10.1002/rmv.2381 (2022).
16. Cobey, S., Larremore, D. B., Grad, Y. H. & Lipsitch, M. Concerns about SARS-CoV-2 evolution should not hold back efforts to expand vaccination. *Nat. Rev. Immunol.* **21**, 330–335. https://doi.org/10.1038/s41577-021-00544-9 (2021).
17. Yewdell, J. W. Antigenic drift: Understanding COVID-19. *Immunity* **54**, 2681–2687. https://doi.org/10.1016/j.immuni.2021.11.016 (2021).
18. Rouzine, I. M. & Rozhnova, G. Evolutionary implications of SARS-CoV-2 vaccination for the future design of vaccination strategies. *Commun. Med. (Lond)*. **3**, 86. https://doi.org/10.1038/s43856-023-00320-x (2023).
19. Moulana, A. et al. Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA.1. *Nat. Commun.* **13**, 7011. https://doi.org/10.1038/s41467-022-34506-z (2022).
20. Harari, S. et al. Drivers of adaptive evolution during chronic SARS-CoV-2 infections. *Nat. Med.* **28**, 1501–1508. https://doi.org/10.1038/s41591-022-01882-4 (2022).
21. Reuschl, A. K. et al. Evolution of enhanced innate immune suppression by SARS-CoV-2 Omicron subvariants. *Nat. Microbiol.* **9**, 451–463. https://doi.org/10.1038/s41564-023-01588-4 (2024).
22. Thorne, L. G. et al. Evolution of enhanced innate immune evasion by SARS-CoV-2. *Nature* **602**, 487–495. https://doi.org/10.1038/s41586-021-04352-y (2022).
23. Adams, R. M., Kinney, J. B., Walczak, A. M. & Mora, T. Epistasis in a Fitness Landscape Defined by Antibody-Antigen Binding Free Energy. *Cell. Syst.* **8**, 86–93e83. https://doi.org/10.1016/j.cels.2018.12.004 (2019).
24. Lyons, D. M., Zou, Z., Xu, H. & Zhang, J. Idiosyncratic epistasis creates universals in mutational effects and evolutionary trajectories. *Nat. Ecol. Evol.* **4**, 1685–1693. https://doi.org/10.1038/s41559-020-01286-y (2020).
25. Chen, C. C. et al. Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8**, 1580–1591. https://doi.org/10.1109/TCBB.2011.46 (2011).
26. Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* **10**, 392–404. https://doi.org/10.1038/nrg2579 (2009).
27. Ueki, M. & Cordell, H. J. Improved statistics for genome-wide interaction analysis. *PLoS Genet.* **8**, e1002625. https://doi.org/10.1371/journal.pgen.1002625 (2012).
28. Kimura, M. Attainment of Quasi Linkage Equilibrium When Gene Frequencies Are Changing by Natural Selection. *Genetics* **52**, 875–890. https://doi.org/10.1093/genetics/52.5.875 (1965).
29. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U S A.* **106**, 67–72. https://doi.org/10.1073/pnas.0805923106 (2009).
30. Neher, R. A. & Shraiman, B. I. Statistical genetics and evolution of quantitative traits. *Rev. Mod. Phys.* **83**, 1283 (2011).
31. Louie, R. H. Y., Kaczorowski, K. J., Barton, J. P., Chakraborty, A. K. & McKay, M. R. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proc. Natl. Acad. Sci. U S A.* **115**, E564–E573. https://doi.org/10.1073/pnas.1717765115 (2018).
32. Wei, W. H., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **15**, 722. https://doi.org/10.1038/nrg3747 (2014). https://www.nature.com/articles/nrg3747#supplementary-information
33. Fisher, R. A. *The genetical theory of natural selection* (Clarendon, 1958, 1930).
34. Pedruzzi, G. & Rouzine, I. M. Epistasis detectably alters correlations between genomic sites in a narrow parameter window. *PLoS One*. **14**, e0214036. https://doi.org/10.1371/journal.pone.0214036 (2019).
35. Rouzine, I. M. Long-range linkage effects in adapting sexual populations. *Sci. Rep.* **13**, 12492. https://doi.org/10.1038/s41598-023-39392-z (2023).

36. Pedruzzi, G., Barlukova, A. & Rouzine, I. M. Evolutionary footprint of epistasis. *PLoS Comput. Biol.* **14**, e1006426. https://doi.org/10.1371/journal.pcbi.1006426 (2018).
37. Kryazhimskiy, S., Dushoff, J., Bazykin, G. A. & Plotkin, J. B. Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet.* **7**, e1001301. https://doi.org/10.1371/journal.pgen.1001301 (2011).
38. Neverov, A. D., Kryazhimskiy, S., Plotkin, J. B. & Bazykin, G. A. Coordinated Evolution of Influenza A Surface Proteins. *PLoS Genet.* **11**, e1005404. https://doi.org/10.1371/journal.pgen.1005404 (2015).
39. Pedruzzi, G. & Rouzine, I. M. An evolution-based high-fidelity method of epistasis measurement: Theory and application to influenza. *PLoS Pathog*. **17**, e1009669. https://doi.org/10.1371/journal.ppat.1009669 (2021).
40. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645. https://doi.org/10.1101/gr.092759.109 (2009).
41. Tao, K. et al. The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.* **22**, 757–773. https://doi.org/10.1038/s41576-021-00408-x (2021).
42. Johnson, B. A. et al. Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. *Nature* **591**, 293–299. https://doi.org/10.1038/s41586-021-03237-4 (2021).
43. Kanakan, A. et al. Clinico-Genomic Analysis Reiterates Mild Symptoms Post-vaccination Breakthrough: Should We Focus on Low-Frequency Mutations? *Front. Microbiol.* **13**, 763169. https://doi.org/10.3389/fmicb.2022.763169 (2022).
44. Wu, H. et al. Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. *Cell. Host Microbe*. **29**, 1788–1801e1786. https://doi.org/10.1016/j.chom.2021.11.005 (2021).
45. Li, Y. et al. The R203M and D377Y mutations of the nucleocapsid protein promote SARS-CoV-2 infectivity by impairing RIG-I-mediated antiviral signaling. *PLoS Pathog*. **21**, e1012886. https://doi.org/10.1371/journal.ppat.1012886 (2025).
46. Cruz, C. A. K. & Medina, P. M. B. Temporal changes in the accessory protein mutations of SARS-CoV-2 variants and their predicted structural and functional effects. *J. Med. Virol.* **94**, 5189–5200. https://doi.org/10.1002/jmv.27964 (2022).
47. Bonhoeffer, S., Chappey, C., Parkin, N. T., Whitcomb, J. M. & Petropoulos, C. J. Evidence for positive epistasis in HIV-1. *Science* **306**, 1547–1550. https://doi.org/10.1126/science.1101786 (2004).
48. Zhang, T. H. et al. Predominance of positive epistasis among drug resistance-associated mutations in HIV-1 protease. *PLoS Genet.* **16**, e1009009. https://doi.org/10.1371/journal.pgen.1009009 (2020).
49. Althaus, C. L. & De Boer, R. J. Dynamics of immune escape during HIV/SIV infection. *PLoS Comput. Biol.* **4**, e1000103. https://doi.org/10.1371/journal.pcbi.1000103 (2008).
50. Batorsky, R., Sergeev, R. A. & Rouzine, I. M. The route of HIV escape from immune response targeting multiple sites is determined by the cost-benefit tradeoff of escape mutations. *PLoS Comput. Biol.* **10**, e1003878. https://doi.org/10.1371/journal.pcbi.1003878 (2014).
51. Smith, C. A. & Ashby, B. Antigenic evolution of SARS-CoV-2 in immunocompromised hosts. *Evol. Med. Public. Health*. **11**, 90–100. https://doi.org/10.1093/emph/eoac037 (2023).
52. Crona, K. Recombination and peak jumping. *PLoS One*. **13**, e0193123. https://doi.org/10.1371/journal.pone.0193123 (2018).
53. Rouzine, I. M. *Mathematical Modeling of Evolution. Volume 1. One-Locus and Multi-Locus Theory and Recombination* (De Gruyter, 2020).
54. Rouzine, I. M. *Mathematical Models of Evolution. Volume 2: Fitness Landscape, Red Queen, Evolutionary Enigmas, and Applications to Virology* (De Gruyter, 2023).
55. Rouzine, I. M. & Weinberger, L. S. The quantitative theory of within-host viral evolution. *J. Stat. Mech. Theory Exp.* https://doi.org/10.1088/1742-5468/2013/01/P01009 (2013).
56. Rouzine, I. M. & Rozhnova, G. Antigenic evolution of viruses in host populations. *PLoS Pathog*. **14**, e1007291. https://doi.org/10.1371/journal.ppat.1007291 (2018).
57. Yan, L., Neher, R. A. & Shraiman, B. I. Phylodynamic theory of persistence, extinction and speciation of rapidly adapting pathogens. *Elife* **8** https://doi.org/10.7554/eLife.44205 (2019).
58. Marchi, J., Lassig, M., Walczak, A. M. & Mora, T. Antigenic waves of virus-immune coevolution. *Proc. Natl. Acad. Sci. U S A*. **118** https://doi.org/10.1073/pnas.2103398118 (2021).
59. Muller, H. G. Some genetic aspects of sex. *Am. Nat.* **66**, 118–128 (1932).
60. Rice, S. *Evolutionary Theory: Mathematical and Conceptual Foundations* (Sinauer Associated, 2004).
61. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78**, 737–756. https://doi.org/10.1093/genetics/78.2.737 (1974).
62. Rouzine, I. M., Wakeley, J. & Coffin, J. M. The solitary wave of asexual evolution. *Proc. Natl. Acad. Sci. U S A*. **100**, 587–592. https://doi.org/10.1073/pnas.242719299 (2003).
63. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
64. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102–103**, 127–144 (1998).
65. Desai, M. M. & Fisher, D. S. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* **176**, 1759–1798. https://doi.org/10.1534/genetics.106.067678 (2007).
66. Hallatschek, O. The noisy edge of traveling waves. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1783–1787 (2010).

67. Rodriguez-Rivas, J., Croce, G., Muscat, M. & Weigt, M. Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. *Proc. Natl. Acad. Sci. U S A*. **119** https://doi.org/10.1073/pnas.2113118119 (2022).
68. Zeng, H. L., Dichio, V., Rodríguez Horta, E., Thorell, K. & Aurell, E. Global analysis of more than 50,000 SARS-CoV-2 genomes reveals epistasis between eight viral genes. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 31519–31526. https://doi.org/10.1073/pnas.2012331117 (2020).
69. Batorsky, R. et al. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. *Proc Natl Acad Sci U S A*108, 5661–5666 https://doi.org:1102036108 (2011). https://doi.org/10.1073/pnas.1102036108
70. Gonzalez-Vazquez, L. D. & Arenas, M. Molecular Evolution of SARS-CoV-2 during the COVID-19 Pandemic. *Genes (Basel)*. 14. https://doi.org/10.3390/genes14020407 (2023).
71. Lagunov, S. V. & Rouzine, I. M. in *BioRxiv* (2025).
72. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform*. **20**, 1160–1166. https://doi.org/10.1093/bib/bbx108 (2019).
73. Rouzine, I. M. & Coffin, J. M. Evolution of human immunodeficiency virus under selection and weak recombination. *Genetics* **170**, 7–18. https://doi.org/10.1534/genetics.104.029926 (2005).
74. Neher, R. A. & Leitner, T. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput. Biol.* **6**, e1000660. https://doi.org/10.1371/journal.pcbi.1000660 (2010).
75. Rast, L. I. et al. Conflicting Selection Pressures Will Constrain Viral Escape from Interfering Particles: Principles for Designing Resistance-Proof Antivirals. *PLoS Comput. Biol.* **12**, e1004799. https://doi.org/10.1371/journal.pcbi.1004799 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

I. V. Likhachev[1] · I. M. Rouzine[1]

✉ I. M. Rouzine
igor.rouzine@iephb.ru

[1] Sechenov Institute of Evolutionary Physiology and Biochemistry, Russian Academy of Sciences, St Petersburg, Russia