PURPOSE-LED PUBLISHING™

PAPER • **OPEN ACCESS**

Two fitness inference schemes compared using allele frequencies from 1068 391 sequences sampled in the UK during the COVID-19 pandemic

To cite this article: Hong-Li Zeng *et al* 2025 *Phys. Biol.* **22** 016003

IOP Publishing Phys. Biol. 22 (2025) 016003 https://doi.org/10.1088/1478-3975/ad9213

Physical Biology

OPEN ACCESS

RECEIVED
11 June 2024

REVISED
31 October 2024

ACCEPTED FOR PUBLICATION
13 November 2024

PUBLISHED
21 November 2024

PAPER

Two fitness inference schemes compared using allele frequencies from 1068 391 sequences sampled in the UK during the COVID-19 pandemic

Hong-Li Zeng[1], **Cheng-Long Yang**[1], **Bo Jing**[1], **John Barton**[2] and **Erik Aurell**[3,*]

[1] School of Science, Nanjing University of Posts and Telecommunications, Key Laboratory of Radio and Micro-Nano Electronics of Jiangsu Province, Nanjing 210023, People's Republic of China
[2] Department of Computational & Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15260, United States of America
[3] Department of Computational Science and Technology, AlbaNova University Center, SE-106 91 Stockholm, Sweden
[*] Author to whom any correspondence should be addressed.

E-mail: eaurell@kth.se, hlzeng@njupt.edu.cn and jpbarton@pitt.edu

Keywords: SARS-CoV-2, allele frequency time series, fitness inference, transient quasi-linkage equilibrium (tQLE), marginal path likelihood (MPL)

Abstract

Throughout the course of the SARS-CoV-2 pandemic, genetic variation has contributed to the spread and persistence of the virus. For example, various mutations have allowed SARS-CoV-2 to escape antibody neutralization or to bind more strongly to the receptors that it uses to enter human cells. Here, we compared two methods that estimate the fitness effects of viral mutations using the abundant sequence data gathered over the course of the pandemic. Both approaches are grounded in population genetics theory but with different assumptions. One approach, tQLE, features an epistatic fitness landscape and assumes that alleles are nearly in linkage equilibrium. Another approach, MPL, assumes a simple, additive fitness landscape, but allows for any level of correlation between alleles. We characterized differences in the distributions of fitness values inferred by each approach and in the ranks of fitness values that they assign to sequences across time. We find that in a large fraction of weeks the two methods are in good agreement as to their top-ranked sequences, i.e. as to which sequences observed that week are most fit. We also find that agreement between the ranking of sequences varies with genetic unimodality in the population in a given week.

1. Introduction

The COVID-19 pandemic had the largest impact on world-wide human health by an infectious disease agent since the Spanish flu more than a century ago [1]. After more than three years of at times high infection rates in practically all countries in the world, the disease has reached an endemic state, and the virus will likely remain in circulation in the foreseeable future. The spread of SARS-CoV-2 was accompanied by the emergence of many variants, some of which successfully replaced earlier variants. These variants differed in their virulence, infectiousness, and resistance to vaccines. They also differed in their exact genotypes, as determined by many high-quality whole-genome sequences deposited in repositories such as GISAID [2].

The unprecedented amount of genomic time series data collected for SARS-CoV-2 allows for analysis that was previously impossible. In particular, this data enables the development and comparison of prediction and/or inference methods that may be useful in a future pandemic, an event that is likely unavoidable even if challenging to predict. Genomic time series analysis also allows for feature discovery, which can help shed light on the biology of the virus in its newly conquered environment, i.e. the human population.

We will here compare and contrast two recently developed approaches for fitness inference from genetic time series data. One approach is based on the quasi-linkage equilibrium (QLE) theory of Kimura [3] and Neher and Shraiman [4, 5], which we will here use in a dynamic (non-stationary) version which

we call tQLE. The basic idea of tQLE is to infer parameters of models in exponential families describing the distribution of genotypes in a population from sequence data with time stamps. Due to the high sampling world-wide during the pandemic, sequence data can be sampled precisely in time, down to periods of even a single week. In this approach, the SARS-CoV-2 genotype distribution is first described by Potts parameters $h_i(t)$ and $J_{ij}(t)$ where t is the sample time (metadata available in GISAID), and where the data can optionally also be stratified by the region of origin for each sample. QLE theory relates epistatic fitness [parameters f_{ij}] to Potts parameters J_{ij}. tQLE additionally gives a relationship between the contribution of additive fitness from variation at genomic position i [parameter f_i], Potts parameters h_i and J_{ij}, and the time derivative $\dot{h}_i$.

The second approach called marginal path likelihood (MPL) was recently developed by Barton *et al* [6], and applied by them on SARS-CoV-2 data up to August 2021 [7]. The main idea of MPL is to estimate the probability of an observed history of allele frequencies from a Wright-Fisher model (or, in the case of SARS-CoV-2, a branching process epidemiological model [7]), including recombination, and then maximize this probability over model parameters. The resulting formula involves frequencies, mutational pressure, and linkage disequilibrium (correlation) between alleles. This approach has some similarities to an inference formula from a time series for neuroscience applications developed by two of us some time ago [8].

The main conceptual difference between the two methods is that MPL is derived from the finite-N noisy dynamics of single-locus frequencies while tQLE is derived from the infinite-N noise-less dynamics of single-locus frequencies and two-locus pair frequencies. tQLE allows for both additive and pairwise epistatic contributions to fitness, if the additive contribution dominates, so that the instantaneous distribution is close to linkage equilibrium. A scenario when this happens, and which is assumed in the version of tQLE used in this work, is when recombination is a faster process than selection and mutations [5, 9]. Other scenarios are however also possible [10], and could be used as a basis for other variants of tQLE. The version of MPL used here is derived from single-locus frequencies, and it can only be used to infer additive fitness. On the other hand, the incorporation of stochastic (finite-N) effects and not requiring the assumption of QLE are advantages of MPL.

Here, we compared the fitness values inferred by tQLE and MPL for weekly batches of SARS-CoV-2 sequences, collected over the first few years of the pandemic. We found good agreement between the two methods on the relative ranks of sequences in terms of their fitness. The similarity of the rankings is especially notable given the differences in modeling assumptions for tQLE and MPL.

2. Materials

2.1. Data preparation

The data utilized here was sourced from the GISAID repository, spanning from the beginning of the COVID-19 pandemic until 12 August 2023. Subsequently, a remarkable decline in the number of genomes uploaded to the database was observed. The inclusion criteria for our analysis involved selecting only high-quality and full-length genomes, as per the defined standards outlined on the GISAID website. All retained genomes possess a length not exceeding 29 903 base pairs. Each genome is labeled based on its sample collection date, a metadata parameter provided by GISAID. Given the observable bias in alternative submission times to GISAID [11], sequences are systematically stratified weekly, resulting in a total of 179 datasets encompassing 5644 661 sequences. Due to a significant geographical imbalance in the collected samples of SARS-CoV-2, the main analysis reported here is focused on the UK region. For robustness, we show in the appendix also corresponding results for three geographic regions (Colorado, Florida, and Japan) of similar size as the UK, and for most of the time also having sufficiently many sequences per week [12].

2.2. Data processing

The data for the regions in figure 1 satisfy the following minimal criteria:

- In any period of 5 days within the time series, there are at least 20 total samples.
- The number of days in the time series is greater than 20.

Applying these criteria, our dataset for the UK region spans 170 weeks, ranging from 2020-03-14 to 2023-06-04. The dataset for the UK comprises 1068 391 sequences in total.

For the sequences in each week, a Multiple Sequence Alignment was constructed through the `MAFFT` software [13, 14]. Sequences from each week are aligned separately to the reference sequence 'Wuhan-Hu-1', with GISAID accession number EPI_ISL_40 2125 [15]. Note that this is different from the procedure in [16], where pre-aligned MSAs were used. The total number of sequences in that study was much less than that used here.

Each MSA is a matrix $\boldsymbol{\sigma} = \{\sigma_i^n | i = 1, \ldots, L, n = 1, \ldots, N\}$, where N represents the number of genomic sequences in a week while L represents the number of loci in an aligned sequence [17, 18]. Thus, all aligned sequences have a length of $L = 29\,903$, the same as

that of the reference sequence, while N by construction varies from week to week, see figure 1. The loci between 256 and 29 674 are referred to as *coding region*, since they code for the protein-coding genes in the SARS-CoV-2 genome. Each entry σ_i^n of the MSA $\boldsymbol{\sigma}$ is either one of the 4 nucleotides (A,C,G,T), or the alignment gap '-', the minorities like 'KYF…' are changed to the sign of '-' for the sake of simplicity of the following allele frequency analysis.

2.3. Data filtering

In figure 2 we show the allele frequency time series for all loci in the UK data, and in figure 3 only for loci in the coding region. These two figures show that at a majority of loci all sequences contain the same symbol, and most of the remaining variation is in the non-coding regions. In a first filtering step we retain in the analysis loci where the most frequent mutation away from wild-type is classified as 'Non-synonymous mutation' as defined in [6] and where the largest mutant frequency is at least 1%. In a second step we retain only those loci which meet the criteria for all weeks. For the UK data used in this study there remains 209 loci.

For the other data sets shown in figure 1(lower panel) there would remain respectively 1063 loci in 'Global', 173 loci in 'EU', 328 loci in 'NA' (North America) and 225 loci in 'Asia'. For consistency, the average Hamming distances are however for all computed from the variability at the same 209 loci as in the UK data set.

3. Methods

3.1. The driving forces of evolution: selection, mutation, recombination and genetic drift

In both approaches to be considered, the driving forces of evolution are assumed (Darwinian) selection, mutation, recombination, and genetic drift (finite population effects). Effects excluded from consideration are hence e.g. spatial barriers (island models). The genome $\mathbf{x} = (x_1, x_2, \ldots, x_L)$ where each x_i is an indicator variable of the allele (nucleotide in the set $\{-, N, A, C, G, T\}$ at locus i (position i in the MSA).

(a) *Fitness* is assumed to be a function

$$F(\mathbf{x}) = \sum_{i=1}^{L}\sum_{a} f_{i,a}^{(1)} \mathbf{1}_{x_i,a} + \sum_{i=1<j}^{L}\sum_{a,b} f_{ij,ab}^{(2)} \mathbf{1}_{x_i,a}\mathbf{1}_{x_j,b}. \tag{1}$$

The coefficients $f_{i,a}^{(1)}$ are called *additive fitness* and parameterize the selective advantage of allele a at locus i with respect to wild-type. The coefficients $f_{ij,ab}^{(2)}$ are called (pair-wise) *epistatic fitness* and parameterize the selective advantage of alleles a and b at loci i and j beyond what they contribute separately. In tQLE $f_{ij,ab}^{(2)}$ are adjustable parameters inferred from the data, which for self-consistency however cannot be too large. In MPL $f_{ij,ab}^{(2)}$ are absent.

The evolution of genotypes in the population over a short time Δt due to fitness is on the level of a normalized distribution given by

$$P(\mathbf{x}, t+\Delta t)\,|_{\text{fit.}} = \frac{e^{\Delta t F(\mathbf{x})}}{\sum_{\mathbf{x}} e^{\Delta t F(\mathbf{x})} P(\mathbf{x}, t)} P(\mathbf{x}, t)\,. \tag{2}$$

In both approaches our goal is to estimate the coefficients $f_{i,a}^{(1)}$.

(b) *Mutations* are random changes of single alleles. In general, they could be parametrized acting as

$$P(\mathbf{x}, t+\Delta t)\,|_{\text{mut.}} = P(\mathbf{x}, t) + \Delta t \sum_i \sum_{ab} \mathbf{1}_{x_i,a} \times \left(\mu_i^{ba} P\left(M_i^{ab}\mathbf{x}, t\right) - \mu_i^{ab} P(\mathbf{x}, t)\right) \tag{3}$$

where M_i^{ab} is the flip operator which changes allele a at locus i to b, and μ_i^{ab} rate of this process. These rates are only partially known, and only parametrise a fraction of naturally occurring mutations. As our focus is here on fitness we will follow the original theoretical literature [5, 6] and take them all equal to one overall mutation rate μ.

(c) *Recombination* is modelled as a process whereby two genomes combine and give rise to a third. On the level of distributions that is given by

$$P(\mathbf{x}, t+\Delta t)\,|_{\text{recomb}} = P(\mathbf{x}, t)\,(1 - r\Delta t) + r\Delta t \sum_{\mathbf{x}_m, \mathbf{x}_f} C\left(\mathbf{x}; \mathbf{x}_m, \mathbf{x}_f\right) \times P(\mathbf{x}_m, t)\, P\left(\mathbf{x}_f, t\right). \tag{4}$$

In above r is an overall recombination rate with dimension inverse time. The function $C(\mathbf{x}; \mathbf{x}_m, \mathbf{x}_f)$ is the specific rate at which genomes $\mathbf{x}_m$ and $\mathbf{x}_f$ combine to yield $\mathbf{x}$. In tQLE it only enters through the derived quantity c_{ij} which is the probability that alleles at loci i and j are inherited from different parents [5, 10, 19]. In MPL recombination drives the evolution and influences the distribution of evolutionary trajectories, but does not directly enter in the inference formulae. The factor $(1 - r\Delta t)$ serves to normalize the distribution. More general models of recombination in the same context are discussed in [19].

(d) *Genetic drift* is the term of stochastic effects due to a finite population. All three evolution equations (2)–(4) are valid on the ensemble level, and can be simulated by evolving several populations in parallel, and then averaging.

Single-locus frequencies and two-loci pair frequencies (and other characteristics) will evolve due to both deterministic drift and random noise. In QLE the corresponding stochastic differential equations for single- and two-loci frequencies are derived and discussed [5]. In MPL the stochastic differential equations for single-locus frequencies are central in deriving the path probabilities as discussed below.

3.2. Quasi-linkage equilibrium (QLE)

The phase of QLE was discovered by Kimura in the study of a two-locus biallelic model [3]. The extension to many loci was investigated by Neher and Shraiman [4, 5]. The generalization to more than two alleles per locus was given in [19]. The two defining properties of QLE (formalized in [10]) are

1. Multi-genome probability distributions factorize such that $P_n(\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(n)}) = P(\mathbf{x}^{(1)})P(\mathbf{x}^{(2)})\cdots P(\mathbf{x}^{(n)})$. This property is especially important for $n=2$ as it allows to model the effects of recombination as a molecular collision in kinetic gas theory (Boltzmann's *Stosszahlansatz*).
2. The single-genome probability distributions are Gibbs distribution with terms no higher than in fitness. For (1) this means the Ising-Potts distributions of equilibrium statistical mechanics

$$P(\mathbf{x}) = \frac{1}{Z(\{h_t\},\{J_t\})}\exp\left(\sum_{i,a} h_{i,t}(a)\,\mathbf{1}_{x_i,a} + \sum_{i<j,ab} J_{ij,t}(a,b)\,\mathbf{1}_{x_i,a}\mathbf{1}_{x_j,b}\right). \quad (5)$$

In (5) we have included a time argument t of the Ising-Potts parameters $h_{i,t}$ and $J_{ij,t}$. This is first to emphasize that (5) is not based on assumptions that lead to thermal equilibrium where h_i and J_{ij} are constant. Second, and more importantly, when (5) is fit to time-ordered data, as will be described, the inferred parameters $h_{i,t}, J_{ij,t}$ do depend on time t. For simplicity of notation we will from now on however not explicitly write this argument t. Relations between fitness parameters $\{f^{(1)}\}$ and $\{f^{(2)}\}$ in (1) and stationary Ising/Potts model parameters $(\{h\},\{J\})$ (5) are key quantitative results of QLE theory. In [3] and [5] they were derived in the limit where overall recombination rate r is larger than both overall mutation rate μ and variations in fitness. Direct tests using scatter plots were first given in [9]. Several alternative relations were derived for larger mutation rates in [20], of which one was tested in [20], see also [10].

3.3. Transient QLE (tQLE) and fitness inference from time series data

As already stated above, QLE is a dynamic theory where parameters $(\{h\},\{J\})$ in general change in time. We here introduce the derived abbreviation tQLE to emphasize that we use the formulas for inference on such in principle (and generally in practice) time-changing data.

The equation for $(\{J\})$ is of the relaxation type, and in the theory of [5]

$$\dot{J}_{ij}(a,b) = f^{(2)}_{ij}(a,b) - rc_{ij}J_{ij}(a,b). \quad (6)$$

For large enough r the Potts parameters will hence relax to a stable fixed point, i.e. to $J^*_{ij}(a,b) = \frac{f^{(2)}_{ij}(a,b)}{rc_{ij}}$, which allows to *infer* epistatic fitness parameters from Potts parameters computed from the data through the formula

$$f^{(2),*}_{ij}(a,b) = rc_{ij}J^*_{ij}(a,b). \quad (7)$$

This relation was derived in [5], and tested (in stationary state) in [9]. As discussed in [21] since (7) only relates pair-wise quantities, it can also work when the single-nucleotide frequencies change. This could for instance be the case of additive fitness changes in time, say by a change of the fitness landscape of which one example could be the introduction of widespread vaccination against SARS-CoV-2 in the COVID19 pandemic.

The equation for $(\{h\})$ is on the other hand not of the relaxation type [5], equation (24)

$$\dot{h}_i(a) = f^{(1)}_i(a) + r\sum_{j,b} c_{ij}J_{ij}(a,b)\,m_j(b) \quad (8)$$

where $m_j(b) = \sum_{\mathbf{x}} P(\mathbf{x})\mathbf{1}_{x_j,b}$ is the frequency of allele b at locus j. Combining (7) and inferred values of $\{h\}$ at two consecutive time intervals lead to the inference formula

$$f^{(1),*}_i(a,\Gamma) = \frac{1}{\Delta t}\left[h_i(a,\Gamma+\Delta t) - h_i(a,\Gamma)\right] - \sum_{j,b} f^{(2),*}_{ij}(a,b,\Gamma)\,m_j(b,\Gamma), \quad (9)$$

with Γ indicating the discrete time.

3.4. Loss of QLE

The QLE state is lost when the distribution no longer fulfills the two listed criteria. A well-studied loss channel at very low mutation rate and sufficiently low recombination rate is through the emergence of clones [5, 22]. These are groups of identical, high-fitness genomes related by common descent. Instead of an exponential model (as in QLE), the distribution of genotypes is instead a mixture of clones, with one separate distribution for each clone. There may also be only a single clone, in which case all (or most) of

the genotypes are the same. Quantitative predictions on the threshold between a QLE phase and a clone-dominated phase were derived in [22]. One aspect of the transition between QLE and a clone-dominated phase is that QLE can only exist as a transient state in a finite population with strictly no mutations. The reason is that in this setting sooner or later the most fit genome takes over as a single dominating clone, see [19] and [9] for a discussion. Hence QLE loss at non-zero mutation rate is relevant. A second loss channel observed at a higher mutation rate leads to a phase of 'noisy clones' coexisting with a QLE-like state. In [10] this new phase was named Non-Random Coexistence (NRC). The transition from QLE to NRC goes through an intermittent phase where the state of the population jumps between QLE and NRC, illustrating the complexity of phenomena not yet completely mapped out even in theoretical models. The dependence of the jump rates on population size N was investigated in [10], and leads to a qualitatively similar behaviour as [22], i.e. for a sufficiently large population only the NRC phase is stable.

3.5. The marginal path likelihood (MPL) method

The marginal path likelihood (MPL) method [6] is based on the evolution of nucleotide frequencies in Kimura's diffusion approximation [23, 24]. The starting point is thus the joint probability $P(\{m\}^{(1)},\{m\}^{(2)},\ldots,\{m\}^{(L)})$ where $m_a^{(i)}$ is the frequency of allele a on locus i, normalized as $\sum_a m_a^{(i)} = 1$. In the diffusion approximation, this probability satisfies a Fokker-Planck equation

$$\partial_t P = -\sum_{i,a} \frac{\partial}{\partial m_a^{(i)}}\left(u_a^{(i)} P\right) + \sum_{ij,ab} \frac{\partial^2}{\partial m_a^{(i)} \partial m_b^{(j)}}\left(D_{ab}^{(ij)} P\right) \tag{10}$$

where the drift vector and diffusion matrix are given by ([6], equations (6) and (S9) and following, notation aligned with the present presentation)

$$u_a^{(i)} = m_a^{(i)}\left(1 - m_a^{(i)}\right) f_i^{(1)}(a) + \mu\left(1 - 2m_a^{(i)}\right) + \sum_{j,b}\left(m_{ab}^{(ij)} - m_a^{(i)} m_b^{(j)}\right) f_j^{(1)}(b) \tag{11}$$

$$D_{ab}^{(ij)} = \begin{cases} m_a^{(i)} m_b^{(i)} & i = j \\ m_{ab}^{(ij)} - m_b^{(i)} m_b^{(j)} & i \neq j \end{cases}. \tag{12}$$

The Fokker–Planck equation (10) corresponds to a multidimensional Langevin equation for which the probability of a path sampled at discrete times can be estimated by standard arguments. Maximizing this path probability with a Gaussian prior leads to the central inference formula in MPL

$$f_i^{(1),*}(a) = \sum_{j,b}\left[\sum_{k=1}^{K} \Delta t_k D_{ab}^{(ij)}(t_k) + \gamma \mathbf{1}_{ia,jb}\right]^{-1}_{ia,jb} \times \left[m_b^{(j)}(t_K) - m_b^{(j)}(t_0) - \mu \sum_{k=1}^{K-1} \Delta t_k \left(1 - 2m_b^{(j)}(t_k)\right)\right]. \tag{13}$$

In above a time interval $[t_0, t_K]$ has been divided up in K sampling intervals and the allele frequencies ($m_a^{(i)}$) and drift and diffusion terms (from (11) and (12)) estimated for each. The sampling interval times are defined as $\Delta t_k = t_{k+1} - t_k$. γ is the width of the Gaussian prior, and acts as a regularizer.

In (9) the inferred additive fitness depends on time and is linear in the time derivative of one inferred Potts parameters $h_i(a)$. This parameter is in itself a (complicated) function of the single-nucleotide and pair-wise frequencies at that time. In (13) the inferred additive fitness also depends on time, more specifically on a time interval, and is linear in $m_a^{(i)}(t_K) - m_a^{(i)}(t_0)$, the change in all the single-nucleotide frequencies over that interval. Pair-wise frequencies also enter in (13), through the dependence of $D_{ab}^{(ij)}$ as in (12).

3.6. Recombination in coronaviruses and in SARS-CoV-2

SARS-CoV-2 is in classifications such as Pango [25, 26], assumed to evolve by descent, and the growth and subsequent decay of SARS-CoV-2 variants [27–29] is well-known. This can be taken to be the standard view of SARS-CoV-2 evolution, analogous to the evolution of other viruses such as influenza. A complicating factor is that many viral variants seem themselves to evolve, to split into sub-variants and perhaps to recombine [30, 31]. On the other hand, coronaviruses in general exhibit recombination [32–35], a process which has also been observed to occur in SARS-CoV-2 [36–40].

Whether a propensity for recombination between pairs of viral genomes results in observed recombination in a viral population is not a simple question; two viruses must meet, typically in the same host, to recombine. The effectiveness and importance of recombination in the general SARS-CoV-2 viral population has accordingly been questioned [41]. The classic QLE phase (Kimura and Neher & Shraiman) assumes a fully mixing population, which the global SARS-CoV-2 viral population during the pandemic clearly was not. If a QLE-like phase exists in a not fully mixing population, and at which parameter values, has not, as far as we are aware of, been systematically studied in the literature. On the other hand, as found by two of us in previous theoretical work

[20], large recombination rate is not the only parameter range that leads to a QLE state in a fully mixing population; a large mutation rate with some recombination rate is also sufficient. An interesting avenue for further theoretical studies would be to try to find out if this mechanism also operates in not fully mixing populations.

3.7. Assumed values of model parameter *r* and their meaning

The recombination rate r sets a scale for the epistatic contributions to fitness which does not change the order of fitness of sequences as long as inferred epistatic fitness dominates over inferred additive fitness. The MPL inference scheme includes a regularization parameter which plays a similar role; this seems at present to be unavoidable in these studies. Here the value of r is set to 0.027 as indicated in [42].

3.8. Rank order comparisons

Given two lists of sequences ordered as to fitness, we compare the rankings by the Spearman correlation coefficient. The Spearman correlation coefficient is a measure of rank correlation obtained by computing the Pearson correlation between the rankings of the values in each vector. Thus, it can robustly measure nonlinear associations between fitness values as well as linear ones. This is computed with the Matlab function 'corr()' with type argument 'Spearman'.

3.9. Manipulations of tQLE and MPL scheme on SARS-CoV-2 genomics

The MSAs over weeks are prepared and filtered according to the processes described in section 2. For tQLE, the maximization of the pseudo-likelihood (PLM) [43, 44] is used on each MSAs to learn the Ising-Potts parameters $h_{i,t}(a)$ and $J_{ij,t}(a,b)$ in equation (5). Then the epistatic fitness parameters $f^2_{ij}(a,b)$ are computed according to equation (7). In which the recombination rate $r = 0.027$ [42] and the probability that alleles at loci i and j are inherited from different parents $c_{ij} = 0.5$ [9] respectively. With the inferred $h_{i,t}(a)$s and $J_{ij,t}(a,b)$s by PLM, the additive fitness parameters for each week are obtained according to equation (9).

For MPL, the set of all weekly data defines the path. Thus we use the entire path (data set) to compute the additive fitness parameters [7]. The MPL library is available at [45].

The fitness values for each sequence in every week are computed using equation (1), using the week additive fitness parameters from tQLE and the whole set from MPL respectively. Then the rank orders for the fitness of sequences in each week are compared through their Spearman correlation to check the inference similarities of these two schemes.

4. Results

Globally, 5644 661 sequences were obtained from the GISAID database. Upon weekly stratification of the data, we obtain 179 weeks in total. Notably, the sample collections of SARS-CoV-2 reveal a pronounced geographical imbalance, as depicted in figure 1. Consequently, our analysis focuses exclusively on genomic data originating from three regions in the UK (England, Wales, and Scotland). The number of sequences from the UK is 1068 391 in total.

4.1. Amount and diversity of collected sequences over time

The weekly stratified datasets are geographically segmented for a detailed analysis. As illustrated in figure 1(upper panel), the number of sequences undergoes dynamic changes across weeks for different regions, including Europe (dashed orange lines), North America (dashed yellow lines), Asia (dashed purple line), three distinct regions of the UK (green line), and the global dataset (blue line). To account for the impact of geographic separation, data from North Ireland is intentionally excluded. Moreover, figure 1(upper panel) reveals a significant downturn in the number of collected samples in 2022, with Europe emerging as the primary source of sequences. The diversity of collected sequences is in figure 1(lower panel) quantified by average Hamming distance. One observes increased diversity after the emergence of Alpha, Delta and Omicron (marked in figure).

4.2. Allele frequencies

We computed allele frequencies over time from SARS-CoV-2 multiple sequence alignments from the UK. Subsequently, the allele frequencies for the nucleotides in the 'Wuhan-Hu-1' reference sequence are selected. Figure 2 illustrates the allele frequencies over all loci ($L = 29,903$ base pairs in total), while figure 3 specifically shows the allele frequencies within the coding region, spanning from the 256th to the 29 674th sites in the sequence. There are sustained fluctuations at the bottom of figure 2, which hence mainly originate from the non-coding region (3′-UTR and 5′-UTR parts) of SARS-CoV-2. Both plots display variations at specific loci within the coding region, of which many of them can be related to listed mutations in known Variants of Concern, compare monthly data and a relation to Omicron reported in [46] and [21] (figure 5). Furthermore, the increased variability in allele frequencies after Omicron took over (after 1Q22) matches the broad peak in sequence diversity shown in figure 1(lower panel).

4.3. Fitness predicted by tQLE

Epistatic fitness or covariation selection coefficients f_{ij}s in QLE follow from the theory developed in [3, 5]

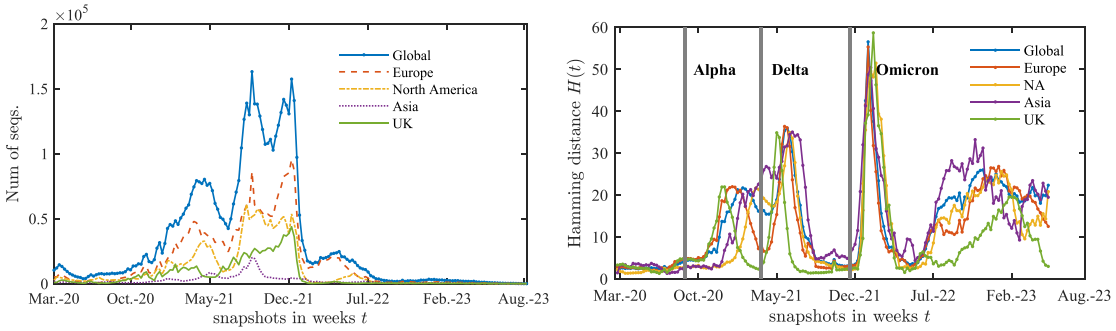

Figure 1. Weekly dynamics of the number of sequences downloaded from the GISAID portal, portraying global trends (solid blue line), European contributions (dashed orange line), North American submissions (dashed yellow line), Asian datasets (purple), and those specifically from the UK excluding Ireland (solid green line). Europe and North America consistently emerge as the primary contributors to GISAID, with Europe leading in sequence submissions throughout most weeks. In 2022, Europe maintained its prominence by contributing the highest number of sequences. Lower panel: sequence diversity per week quantified by average Hamming distance between sequences collected and retained in the analysis that week. Vertical lines mark the first observations of Variants of Concern (VoCs) Alpha, Delta, and Omicron, each of which is followed by a peak in diversity approximately when the variant rose to dominance.

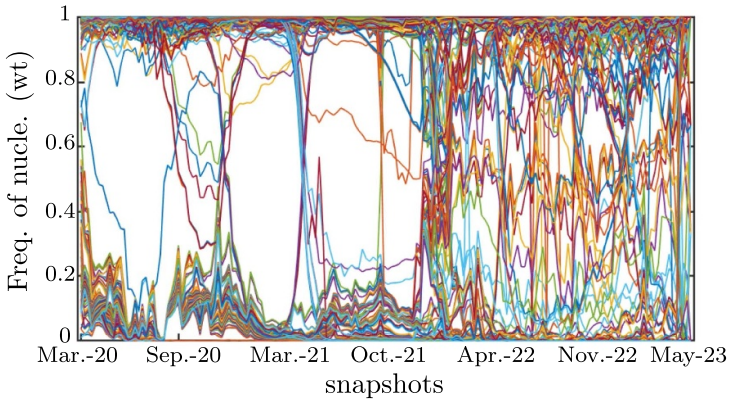

Figure 2. Allele frequencies of all 29 903 sites of the nucleotides in the 'Wuhan-Hu-1' sequence for the UK datasets over weeks. The frequencies located at the bottom mainly originate from the non-coding region (3'-UTR and 5'-UTR) of SARS-CoV-2. There are more oscillations in 2022 than those in 2020 and 2021.

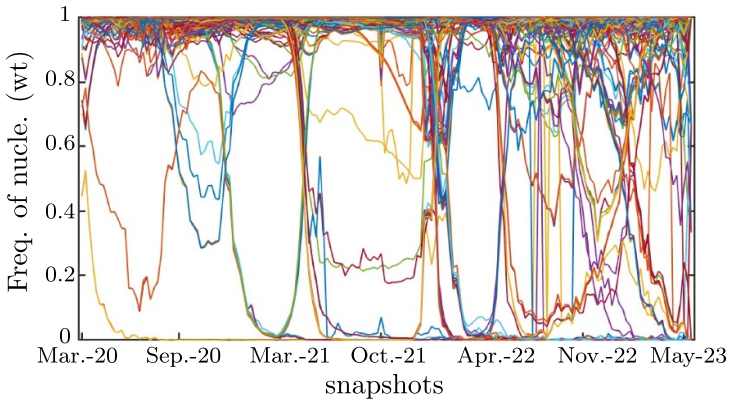

Figure 3. Allele frequencies for all sites in the coding region (sites 266 to 29 674) for the UK datasets of the wild-type nucleotides from the 'Wuhan-Hu-1' sequence. The fluctuations as shown at the bottom of figure 2 disappear here. The oscillations depicted here are more visible than those in figure 2.

(in the version for bi-allelic loci). This theory generalizes directly to multi-allelic loci [9, 19, 21], and leads to the inference formulae equation (7). The additive fitness or selection coefficients f_is are in tQLE analogously obtained as the differences between the time derivative of an additive term ($\dot{h}_i$) and a combination

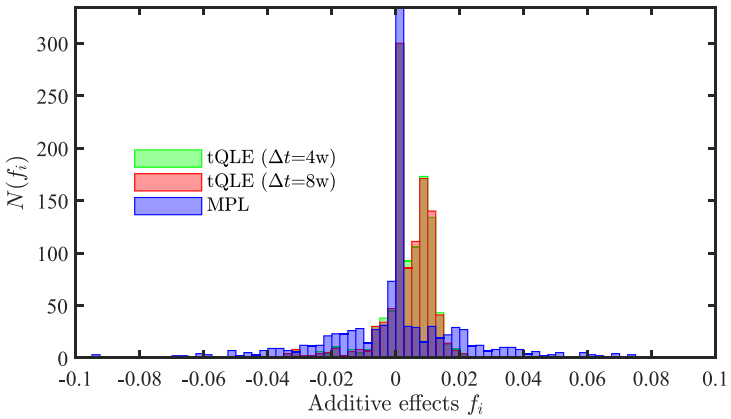

Figure 4. Histograms of the additive fitness from the tQLE (green and red bars) and MPL (blue bars) approaches. For the tQLE method, two distinct time intervals, namely $\Delta t = 4$ weeks (denoted by light green) and $\Delta t = 8$ weeks (denoted by light red), are considered in the additive term of equation (9). This illustrative example showcases the outcomes for the week spanning from 2021-12-19 to 2021-12-25 of the UK dataset.

of the epistatic terms and allele frequencies ($\sum_j f_{ij} m_j$). In the generalization to multi-allelic loci, and when time derivatives are approximated as discrete time differences, this leads to inference formula equation (9) where t and $t + \Delta$ stand for two different weeks.

As shown in figure 4, the additive fitness by tQLE with the time interval of $\Delta t = 4$ (green) and 8 (red) weeks are consistent with each other. This indicates that equation (9) tQLE fitness values do not strongly depend on the choice of Δt values, which is further shown in figure 5.

Figure 5 shows the epistatic terms versus f_is inferred by the tQLE for $\Delta t = 4$ weeks. The red squares lay closely around the diagonal. Such high consistency demonstrates that the epistatic term dominates over the time derivative $\dot{h}_i$ in the total inference formula for f_is. This pattern is consistent across all times in the present data set.

4.4. Additive fitness inferred by MPL

The additive fitness inferred by MPL validates its stability and reliability for the SARS-CoV-2 datasets. Here, even with a different stratification strategy for the UK datasets, similar results are obtained with those in [7]. While the great majority of fitness effects of mutations are inferred to be nearly neutral, MPL infers both strongly beneficial and deleterious mutations, as illustrated by the conspicuous deviation of the blue bars in figure 4 from the neutral zero point.

4.5. Comparison between MPL and tQLE

Figure 4 shows the histograms of the additive fitness f_is by the tQLE (green and red bars) and MPL (blue bars) models, respectively. The f_is anticipated by the tQLE model exhibit a distribution proximate to the zero point, constraining in a relatively narrow range. In contrast, fitness effects inferred by MPL are mostly concentrated around zero, but with large deviations for a small number of mutations.

To assess the fitness estimates derived from the tQLE and MPL methods, we used both methods to rank sequences within a time window according to their fitness, and then compared these rankings. The fitness score of a sequence is defined in equation (1), in which the f_is from the first term are the fitness effect while the f_{ij}s from the second term are the epistatic term. The total fitness of a sequence from the MPL corresponds to the first term of equation (1) while that for the tQLE method is given by both terms of equation (1). For each time window (one week) the tQLE and MPL fitness scores for each sequence are computed and subsequently arranged in descending order.

In figure 6 we show the Spearman rank correlation coefficient c between the two rankings, stratified as to time (mean value per week, upper panel) and as to overall distribution (histogram, lower panel). The agreement is generally good ($c > 0.8$ for 105 out of 166 weeks). MPL and tQLE hence order sequences as to fitness in closely similar ways, for these UK-sampled SARS-CoV-2 sequences obtained from GISAID.

4.6. Change of agreement over time

Figure 6 also relates agreement/disagreement in rankings to diversity (or lack thereof) in the sequences obtained in one week. For easier visualization we have plotted Spearman correlation $c(t)$ together with the relative Hamming distance deficit $rH^{(-)}$ (defined in caption to figure 6). The two curves move in concert. This indicates a contravariation of $c(t)$ with average Hamming distance $H(t)$, i.e. that periods of high (low) Spearman correlation are related to periods of low (high) sequence diversity. A possible explanation is that in this data set periods of high sequence variability (large average Hamming distance) appeared when one VoC was in the process of taking over the population, see figure 1(lower panel). At these times the UK viral population resembled a mixture of two

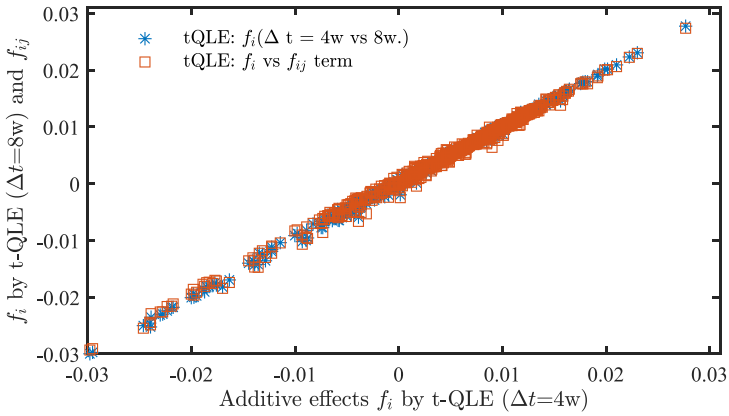

Figure 5. Scatter plots for the additive fitness f_i obtained from the tQLE. The presented results are derived from the dataset as of 15 December 2021, for the UK. Blue stars represent the f_i values corresponding to a time interval of $\Delta t = 8$ weeks (on the y-axis) against those for $\Delta t = 4$ weeks (on the x-axis), while red squares represent the epistatic term $-\sum_{j,b} f_{ij}(a,b)m_j(b)$ in equation (9) against the f_i values for $\Delta t = 4$ weeks. Notably, both cases exhibit a close alignment with the diagonal, indicating a strong correlation between the compared terms.

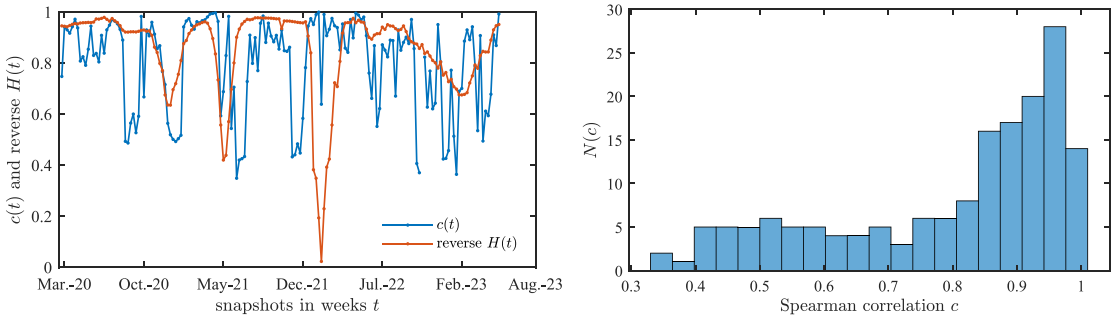

Figure 6. Upper panel: Weekly Spearman correlations $c(t)$ and relative Hamming distance deficit $rH^{(-)}$. (Blue curve) $c(t)$ for same UK data as in figure 1 (England, Scotland, Wales) from late February 2020 to early June 2023. First data point is for the time interval 24 February 2020 and 14 March 2020. (Red curve) $rH^{(-)}(t) = (H_{\max} - H(t))/H_{\max}$ where $H(t)$ is the average Hamming distance between pairs of SARS-CoV-2 genomes sampled in week t and $H_{\max}$ (about 60 in this data set) is the maximum of $H(t)$ over all weeks. Lower panel: Distribution of the Spearman correlations c between the top 10% fitness provided by the MPL and the tQLE approach per week. In 105 out of 166 weeks $c > 0.8$.

clones, different from the one Gibbs-Boltzmann distribution posited in QLE theory from which the tQLE inference method has been derived, see section 3.

5. Discussion

Fitness is the central notion of population genetics going back to the beginning of the field. Inferring fitness has been cumbersome, and much of the literature has been dominated by theoretical investigations. The ongoing sequencing revolution has the potential to change this state of affairs, if fitness can be reliably inferred from large-scale population-wide whole-genome sequence data.

In previous work, physics-inspired models and methods have been applied to study the evolution of viruses. Prominent examples include simple physically-inspired fitness models on both smooth [47] and rugged [48] fitness landscapes. Physical methods have also been applied to phylogenetic trees or sequence distributions to study the evolution of viruses like influenza [49, 50] and HIV [51–53].

We have here compared two approaches, MPL and tQLE, to infer fitness from time-stratified snapshots of an evolving population, applied to UK data on SARS-CoV-2 during the COVID-19 pandemic. These methods differ from prior examples in that they attempt to learn the fitness effects of mutations directly from evolutionary dynamics. The focus on dynamics also differs from most machine learning-based approaches, including large language models, which typically focus on large sequence ensembles without a temporal component [54, 55]. Comparisons of the inferred fitness of SARS-CoV-2 sequences during discrete time windows reveal a varying level of correlation between the two approaches. During a large fraction of time windows the agreement between the two approaches is quite strong, as shown in figure 6.

The gold standard for the accuracy of inferred fitness is comparison to experiments. Nevertheless,

different inference schemes can be compared between themselves, and inter-scheme agreement is a proxy when experiments are not available, as they are not on a population-wide and global genomic scale for SARS-CoV-2. The two approaches rest on simplifying assumptions of different kinds. In MPL, the fitness landscape is assumed to contain only additive components of fitness, which is known to be a simplification. In tQLE, on the other hand, the instantaneous state of the population is assumed to be in a QLE state, i.e., as in a Gibbs-Boltzmann distribution with effective energy terms dependent on fitness. It is not *a priori* clear which of the two sets of assumptions is the strongest for a strongly recombining virus like SARS-CoV-2, and their relative strengths could also have varied during the pandemic. The agreement between these two approaches suggests that, at least in some cases, sufficient data exists to make meaningful inferences about fitness that are not strongly dependent on the details of the underlying model.

We posit that for future pandemics sequence data is very likely to be abundantly available and available much sooner than experimentally determined fitness scores. Fitness parameters systematically inferred from data may then yield predictions useful in the analysis and the understanding of how the pandemic evolves, and to the choice and evolution of countermeasures. Furthermore, our results suggest the possibility of combining the two methods by taking the stochastic dynamics in a QLE state as developed in [5] into account. Recently, MPL was extended to consider epistatic interactions [56], but the resulting expressions are computationally intensive and require the calculation of fourth-order correlations. This has made epistatic inference with MPL challenging for populations with large numbers of mutations, as observed for SARS-CoV-2. Introducing ideas from QLE could then reduce the computational burden and widen the scope of MPL.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/hlzeng/UK_header_collection_submission_date_per_week_csv.

Acknowledgments

HLZ and EA thank Dr Vito Dichio for constructive remarks. The work of HLZ was sponsored by National Natural Science Foundation of China (11705097), Natural Science Foundation of Nanjing University of Posts and Telecommunications (Grant No. 221101, 222134). The work of JPB reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM138233. EA acknowledges support of the Swedish Research Council through Grant 2020-04980.

Appendix A. Two schemes for different regions

To strengthen our argument, we chose to test three geographic regions of similar size as the UK, and at least some of the time also sufficient data: 1. Colorado; 2. Florida; 3. Japan. The distributions of the Spearman correlations are shown in figure 7. As one can see, these results are qualitatively the same as the inferred results for the UK. However, as the number of genomes in these three regions for some weeks is quite small (i.e. a few tens or less), we have to use all the genomes instead of the tops for the UK case.

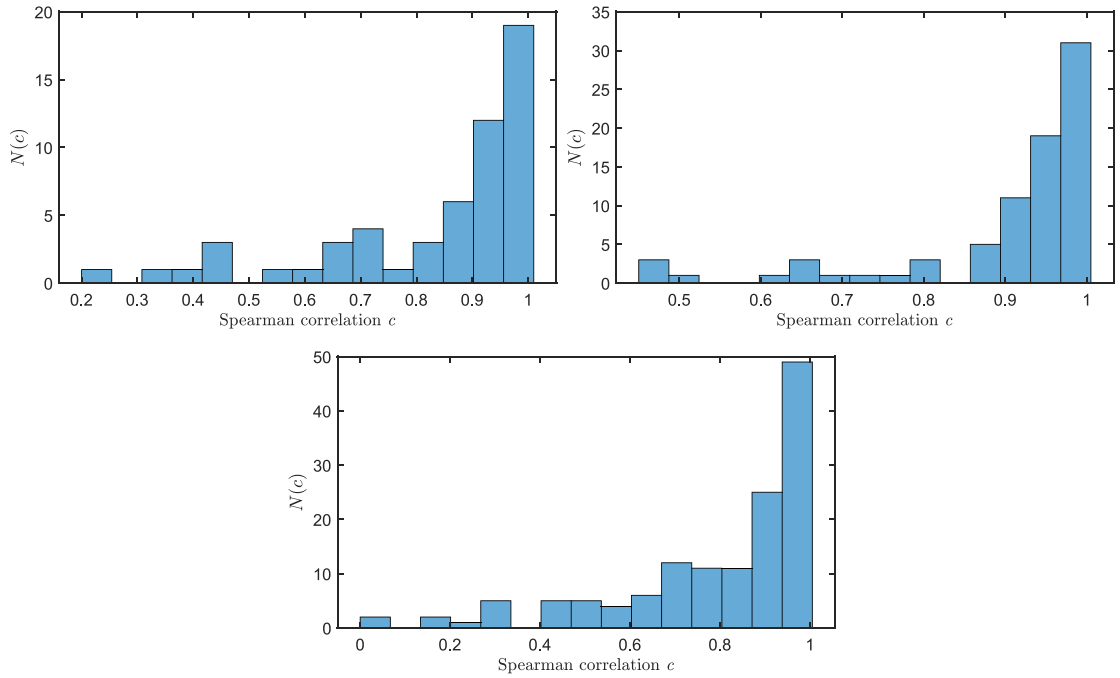

Figure 7. Distributions of Spearman correlations over weeks of three regions far from the UK. Top to bottom: Colorado, Florida, and Japan respectively.

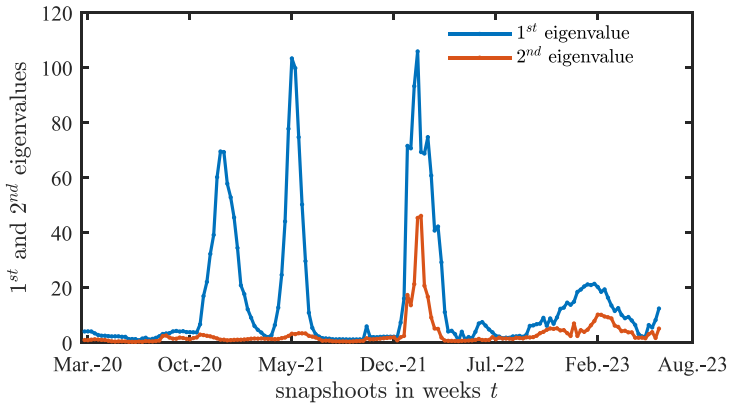

Figure 8. The eigenvalues of the first and second component with the PCA projection of weekly MSAs. Blue is the eigenvalues of the 1st component, while red is that of the 2nd component.

Appendix B. PCA analysis on MSAs

We have tried the PCA projection on the UK dataset in addition to the Hamming distance between sequences as we used in the main context. Specifically, we computed the first and second eigenvalues for different weeks. As shown in figure 8, the first eigenvalue (leading PCA component) displays a similar behaviour over weeks as the average Hamming distance. For comparison, the second PCA component displays a different type of behaviour which we have not tried to interpret.

ORCID iDs

Hong-Li Zeng https://orcid.org/0000-0003-1764-4657
Cheng-Long Yang https://orcid.org/0009-0000-4011-0990
John Barton https://orcid.org/0000-0003-1467-421X
Erik Aurell https://orcid.org/0000-0003-4906-3603

References

[1] World Health Organization 2022 *Who Coronavirus (Covid-19) Dashboard* (available at: https://covid19.who.int/) (Accessed 17 May 2022)
[2] Shu Y and McCauley J 2017 GISAID: global initiative on sharing all influenza data – from vision to reality *Eurosurveillance* **22** 30494
[3] Kimura M 1965 Attainment of quasi linkage equilibrium when gene frequencies are changing by natural selection *Genetics* **52** 875

[4] Neher R A and Shraiman B I 2009 Competition between recombination and epistasis can cause a transition from allele to genotype selection *Proc. Natl Acad. Sci.* **106** 6866
[5] Neher R A and Shraiman B I 2011 Statistical genetics and evolution of quantitative traits *Rev. Mod. Phys.* **83** 1283
[6] Sohail M S, Louie R H Y, McKay M R and Barton J P 2021 MPL resolves genetic linkage in fitness inference from complex evolutionary histories *Nat. Biotechnol.* **39** 472
[7] Lee B, Sohail M S, Finney E, Ahmed S F, Quadeer A A, McKay M R and Barton J P 2022 *medRxiv* (https://doi.org/10.1101/2021.12.31.21268591) (posted online 30 September 2024)
[8] Zeng H-L, Alava M, Aurell E, Hertz J and Roudi Y 2013 Maximum likelihood reconstruction for ising models with asynchronous updates *Phys. Rev. Lett.* **110** 210601
[9] Zeng H-L and Aurell E 2020 Inferring genetic fitness from genomic data *Phys. Rev.* E **101** 052409
[10] Dichio V, Zeng H-L and Aurell E 2023 Statistical genetics in and out of quasi-linkage equilibrium *Rep. Prog. Phys.* **86** 052601
[11] Kalia K, Saberwal G and Sharma G 2021 The lag in SARS-CoV-2 genome submissions to GISAID *Nat. Biotechnol.* **39** 1058
[12] Zeng H L 2024 UK_header_collection_submission_date_per_week_cs *Github* (available at: https://github.com/hlzeng/UK_header_collection_submission_date_per_week_csv)
[13] Katoh K, Rozewicki J and Yamada K D 2017 MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization *Briefings Bioinform.* **20** 1160
[14] Kuraku S, Zmasek C M, Nishimura O and Katoh K 2013 aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity *Nucleic Acids Res.* **41** W22
[15] Chen J *et al* 2020 Structural basis for helicase-polymerase coupling in the SARS-CoV-2 replication-transcription complex *Cell* **182** 1560
[16] Zeng H-L, Dichio V, Rodriguez Horta E, Thorell K and Aurell E 2020 Global analysis of more than 50,000 SARS-CoV-2 genomes reveals epistasis between eight viral genes *Proc. Natl Acad. Sci.* **117** 31519
[17] Cocco S, Feinauer C, Figliuzzi M, Monasson R and Weigt M 2018 Inverse statistical physics of protein sequences: a key issues review *Rep. Prog. Phys.* **81** 032601
[18] Horta R, Barrat-Charlaix P and Weigt M 2019 *Entropy* **21** 1
[19] Gao C-Y, Cecconi F, Vulpiani A, Zhou H-J and Aurell E 2019 DCA for genome-wide epistasis analysis: the statistical genetics perspective *Phys. Biol.* **16** 026002
[20] Zeng H-L, Mauri E, Dichio V, Cocco S, Monasson R and Aurell E 2021 Inferring epistasis from genomic data with comparable mutation and outcrossing rate *J. Stat. Mech.* **2021** 083501
[21] Zeng H-L, Liu Y, Dichio V and Aurell E 2022 Temporal epistasis inference from more than 3500000 SARS-CoV-2 genomic sequences *Phys. Rev.* E **106** 044409
[22] Neher R A, Vucelja M, Mezard M and Shraiman B I 2013 Emergence of clones in sexual populations *J. Stat. Mech.* **2013** 01008
[23] Kimura M 1956 A model of a genetic system which leads to closer linkage by natural selection *Evolution* **10** 278
[24] Kimura M 1964 Diffusion models in population genetics *J. Appl. Probab.* **1** 177–232
[25] Pango lineages: latest epidemiological lineages of sars-cov-2 2022 (available at: https://cov-lineages.org/) (Accessed 17 May 2022)
[26] Rambaut A, Holmes E C, O'Toole Á, Hill V, McCrone J T, Ruis C, du Plessis L and Pybus O G 2020 A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology *Nat. Microbiol.* **5** 1403–7
[27] Chand M *et al* 2021 *Investigation of Sars-CoV-2 Variants of Concern in England* (Public Health England)
[28] Tegally H *et al* 2021 Detection of a SARS-CoV-2 variant of concern in South Africa *Nature* **592** 438–43
[29] Chand M, Hopkins S, Dabrera G, Achison C, Barclay W, Ferguson N, Volz E, Loman N, Rambaut A and Barrett J 2020 *Sars-cov-2 Variants of Concern and Variants Under Investigation in England Technical Briefing 29* (Public Health England)
[30] Rono E K 2021 (posted online 28 October 2021) *bioRxiv* (https://doi.org/10.1101/2021.10.08.463334)
[31] Duerr R *et al* 2022 Clinical and genomic signatures of SARS-CoV-2 Delta breakthrough infections in New York *eBioMedicine* **82** 104141
[32] Lai M M and Cavanagh D 1997 The molecular biology of coronaviruses *Adv. Virus Res.* **48** 1–100
[33] Graham R L and Baric R S 2010 Recombination, reservoirs and the modular spike: mechanisms of coronavirus cross-species transmission *J. Virol.* **84** 3134
[34] Hartenian E, Nandakumar D, Lari A, Ly M, Tucker J M and Glaunsinger B A 2020 The molecular virology of coronaviruses *J. Biol. Chem.* **295** 12910
[35] Li X, Giorgi E E, Marichannegowda M H, Foley B, Xiao C, Kong X-P, Chen Y, Gnanakaran S, Korber B and Gao F 2020 *Sci. Adv.* **6** eabb9153
[36] Choi B *et al* 2020 Persistence and evolution of SARS-CoV-2 in an immunocompromised host *New Engl. J. Med.* **383** 2291
[37] Baang J H *et al* 2021 Prolonged severe acute respiratory syndrome coronavirus 2 replication in an immunocompromised patient *J. Infect. Dis.* **223** 23
[38] Gribble J, Stevens L J, Agostini M L, Anderson-Daniels J, Chappell J D, Lu X, Pruijssers A J, Routh A L, Denison M R and Pekosz A 2021 The coronavirus proofreading exoribonuclease mediates extensive viral recombination *PLoS Pathog* **17** e1009226
[39] Hensley M K *et al* 2021 *Clin. Infect. Dis.* **28** ciab072
[40] Kemp S A *et al* 2021 *Nature* **592** 277–82
[41] VanInsberghe D, Neish A S, Lowen A C and Koelle K 2021 Recombinant SARS-CoV-2 genomes circulated at low levels over the first year of the pandemic *Virus Evol.* **7** veab059
[42] Turakhia Y *et al* 2022 Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape *Nature* **609** 994
[43] Ekeberg M, Hartonen T and Aurell E 2014 Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences *J. Comput. Phys.* **276** 341
[44] Gao C-Y 2018 Gaochenyi/cc-plm Github http://github.com/gaochenyi/CC-PLM
[45] Barton J and Lee B 2021 Barton/mpl Github https://github.com/bartonlab/paper-SARS-CoV-2-inference
[46] Zeng H-L, Liu Y, Dichio V, Thorell K, Nordén R and Aurell E 2021 arXiv:2109.02962
[47] Tsimring L S, Levine H and Kessler D A 1996 RNA virus evolution via a fitness-space model *Phys. Rev. Lett.* **76** 4440
[48] Fontana W, Schnabl W and Schuster P 1989 Physical aspects of evolutionary optimization and adaptation *Phys. Rev.* A **40** 3301
[49] Łuksza M and Lässig M 2014 A predictive fitness model for influenza *Nature* **507** 57
[50] Neher R A, Russell C A and Shraiman B I 2014 Predicting evolution from the shape of genealogical trees *eLife* **3** e03568
[51] Ferguson A L, Mann J K, Omarjee S, Ndung'u T, Walker B D and Chakraborty A K 2013 Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design *Immunity* **38** 606
[52] Mann J K, Barton J P, Ferguson A L, Omarjee S, Walker B D, Chakraborty A, Ndung'u T and Regoes R R 2014 The fitness landscape of HIV-1 Gag: Advanced Modeling Approaches and Validation of Model Predictions by In Vitro Testing *PLOS Comput. Biol.* **10** 1

[53] Barton J P, Goonetilleke N, Butler T C, Walker B D, McMichael A J and Chakraborty A K 2016 Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable *Nat. Commun.* **7** 11660
[54] Hie B, Zhong E D, Berger B and Bryson B 2021 Learning the language of viral evolution and escape *Science* **371** 284
[55] Thadani N N, Gurev S, Notin P, Youssef N, Rollins N J, Ritter D, Sander C, Gal Y and Marks D S 2023 Learning from prepandemic data to forecast viral escape *Nature* **622** 818
[56] Sohail M S, Louie R H Y, Hong Z, Barton J P, McKay M R and Townsend J 2022 Inferring epistasis from genetic time-series data *Mol. Biol. Evol.* **39** msac199