Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Biological modelling and information

Inferring epistasis from genomic data with comparable mutation and outcrossing rate

Hong-Li Zeng[1,7,*], Eugenio Mauri[2,7], Vito Dichio[3,4,5], Simona Cocco[2], Rémi Monasson[2] and Erik Aurell[6]

[1] School of Science, and New Energy Technology Engineering Laboratory of Jiangsu Province, Nanjing University of Posts and Telecommunications, Nanjing 210023, People's Republic of China
[2] Laboratory of Physics of the Ecole Normale Supérieure, CNRS UMR 8023 and PSL Research, 24 rue Lhomond, 75231 Paris cedex 05, France
[3] Institut du Cerveau et de la Moelle épinière, ICM, F-75013, Paris, France
[4] Inria, Aramis project-team, F-75013, Paris, France
[5] Sorbonne Université, F-75013, Paris, France
[6] KTH-Royal Institute of Technology, AlbaNova University Center, SE-106 91 Stockholm, Sweden
E-mail: hlzeng@njupt.edu.cn

Received 31 December 2020
Accepted for publication 1 June 2021
Published 13 August 2021

Online at stacks.iop.org/JSTAT/2021/083501
https://doi.org/10.1088/1742-5468/ac0f64

Abstract. We consider a population evolving due to mutation, selection and recombination, where selection includes single-locus terms (additive fitness) and two-loci terms (pairwise epistatic fitness). We further consider the problem of inferring fitness in the evolutionary dynamics from one or several snapshots of the distribution of genotypes in the population. In recent literature, this has been done by applying the quasi-linkage equilibrium regime, first obtained by Kimura in the limit of high recombination. Here, we show that the approach also works in the interesting regime where the effects of mutations are comparable to or larger than recombination. This leads to a modified main epistatic fitness

*Author to whom any correspondence should be addressed.
[7]These authors contributed equally to this work.

1742-5468/21/083501+24$33.00

inference formula where the rates of mutation and recombination occur together. We also derive this formula using by a previously developed Gaussian closure that formally remains valid when recombination is absent. The findings are validated through numerical simulations.

Keywords: computational biology, evolutionary and comparative genomics, population dynamics

Contents

1. Introduction

Fitness as understood in this paper is the propensity of an organism to pass on its genotype to the next generation, described by a fitness value of each genotype. A set of such values is called a fitness landscape. Evolution is a process whereby nature tends towards populating the peaks in the landscape [1]. Motion in fitness landscapes describes the evolution of a population of one species in a roughly constant environment. Prime

J. Stat. Mech. (2021) 083501

examples of this are pathogens and parasites colonizing a host evolving on a much slower time scale. The most fit pathogen is then one that is best able to exploit the opportunities and weaknesses of a typical host to grow, multiply and eventually spread to other hosts. Excluded from the concept of fitness as considered here are aspects of games of competition and cooperation in evolution [2, 3].

Sequencing of genomes of human pathogens today happens on a massive scale. In an extreme example, samples of SARS-CoV-2, the etiological agent of the disease COVID-19, have by now been sequenced more than 1,200,000 times (accessed on 23 April 2021), and is being sequenced many thousands of times daily [4–6]. This virus in the betacoronavirus family has only been known to science for about 16 months.

It is clear that much information about the evolutionary process must be contained in such data. In particular, if genetic variants in different positions contribute synergistically to fitness this should be reflected in the distribution over genotypes. The goal of this paper is to address the basis of such an approach, and to develop tools to use it better in the future. In two recent contributions [7, 8], we have argued that a natural setting is the quasi-linkage equilibrium (QLE) phase of Kimura [9], surveyed by Kirkpatrick, Johnson and Barton [10], and more recently studied by Neher and Shraiman [11, 12]. When recombination (the exchange of genomic material between individuals, or sex) is a much faster process than mutations or selection due to fitness the stationary distribution over genotypes is the Gibbs–Boltzmann distribution of an Ising or Potts model. The inverse Ising/Potts [13, 14] or direct coupling analysis (DCA) [15–17] methods have been invented to infer the parameters of such distributions from samples. The quantitative properties of QLE allow us to go one step further, and relate those effective couplings to the parameters of the evolutionary dynamics, which we will call the Kimura–Neher–Shraiman (KNS) theory. In [8], we showed that it is indeed possible to retrieve synergistic contributions to fitness from simulated population data by KNS theory.

In the following, we will present an extension where we relax the requirement that recombination has to be the fastest process in the problem. Instead we allow for either recombination or mutation being the fastest process and derive a new modified epistatic inference formula. We do this both by adapting the argument from QLE [12] and by a Gaussian closure recently developed by three of us [18, 19]. We will show that this new theory allows for retrieving synergistic contributions to fitness in much wider parameter ranges. Recombination (sex) is hence no longer required to be a much stronger process than mutations, but could in the Gaussian closure actually be set to zero. The conditions on recombination compared to variations in synergistic contributions to fitness are also much less strict in the new theory.

The paper is organized as follows. In section 2, we summarize evolution driven by selection, recombination and mutations, and contrast the different epistasis inference formulae. In section 3, we derive the formula at high mutation but not necessarily high recombination within QLE, while in section 4 we do it from the Gaussian closure ansatz. In section 5, we summarize our model and simulation strategies, and in sections 5.1 and 5.2 we compare how well we are able to infer fitness when varying mutation rate, the strength of fitness variations, and the rate of recombination. In section 6, we summarize and discuss our results. Appendices contain additional material. Appendix A

J. Stat. Mech. (2021) 083501

computes higher order terms for the inference formula in the Gaussian closure scheme. Appendix B contains parameter settings for simulations of an evolving population using the FFPopSim software [20], and in appendix C we give details on the DCA method we have used in this work. Appendix D presents the comparisons of equations obtained from QLE and Gaussian closure. Appendix E shows the effects of genetic drift on the epistasis inference. Appendix F provides the epistasis inference with Gaussian distributed additive effects.

2. Evolutionary dynamics and epistasis inference

The forces of evolution in classical population genetics are selection, mutations and genetic drift [21, 22]. Selection confers an advantage on individuals with certain characteristics, so that they tend to have more descendants. Mutations are random changes of the genomes. Genetic drift is the element of chance as to which individual survives, and which does not. Common to these three forces is that they all act on the single genotype level: an organism survives to the next generation or it does not. If it does, it will have a number of descendants 'children', 'grand-children', etc. The distribution of individuals over genotypes can then formally be written as a gain-loss process

$$\partial_t P(\mathbf{g},t) = \sum_{\mathbf{g}'} \left(k_{\mathbf{g}',\mathbf{g}} P(\mathbf{g}',t) - k_{\mathbf{g},\mathbf{g}'} P(\mathbf{g},t)\right), \tag{1}$$

where the rates $k_{\mathbf{g}',\mathbf{g}}$ encode selection and mutation. Genetic drift cannot be described by equation (1) directly, which is valid in the infinite population size limit, but appears in Monte Carlo simulation naturally through finite N effects. The details of relevant equations are discussed in great detail in [12] as well as more recently in [7, 8].

Recombination (or sex) is the process by which two genotypes combine to give a third one in the next generation. It cannot be expressed in the form of equation (1). Instead, in general terms it looks like

$$\partial_t P(\mathbf{g},t) = \cdots + \sum_{\mathbf{g}',\mathbf{g}''} C_{\mathbf{g},\mathbf{g}',\mathbf{g}''} P_2(\mathbf{g}',\mathbf{g}'',t), \tag{2}$$

where P_2 stands for the joint probability of two genotypes $\mathbf{g}'$ and $\mathbf{g}''$, and $C_{\mathbf{g},\mathbf{g}',\mathbf{g}''}$ is the rate at which these two produce an offspring $\mathbf{g}$. Equation (2) is not closed; there would be an equation for $\partial_t P_2$, which would depend on the three-genotype distribution P_3, and so on. A standard way to close such a BBKGY-like hierarchy is to assume random mating (random collisions), i.e. $P_2(\mathbf{g}',\mathbf{g}'',t) = P(\mathbf{g}',t)P(\mathbf{g}'',t)$. Combining (1) and (2), we hence get the evolution of a population as a non-linear differential equation analogous to a Boltzmann equation.

In (1) and (2), each genotype $\mathbf{g}$ is seen as a sequence of positions (or loci) of length L, $\mathbf{g} \equiv \{s_0, s_1, \ldots, s_{L-1}\}$. The variable at each position (the allele) s_i can be in one out of n_i states. In the following discussion, we simplify by taking $n_i = 2$ such that s_i is a binary variable. Following the conventions in the physical literature, and in particular [12], we set $s_i = \pm 1$.

J. Stat. Mech. (2021) 083501

We will from now on limit ourselves to fitness landscapes that contain linear and quadratic terms in the allele variables. This means that the fitness of a genotype is given by a function

$$F(\mathbf{g}) = \sum_i f_i s_i + \sum_{ij} f_{ij} s_i s_j. \tag{3}$$

The linear term f_i is called an *additive contribution to fitness*, while the quadratic f_{ij} is an *epistatic contribution to fitness*. The goal of the line of research pursued in this paper is to find ways to retrieve the f_{ij} from the distribution of genotypes in a population.

The QLE theory is based on approximating the genome distribution P as a Gibbs–Boltzmann distribution of the Ising/Potts type:

$$\log P(\mathbf{g}, t) = \Phi(t) + \sum_i \phi_i(t) s_i + \sum_{i<j} J_{ij}(t) s_i s_j, \tag{4}$$

In the above, $\Phi(t)$ is a normalization factor playing the same role as $-\beta F(\beta)$ in statistical mechanics. By expressing the evolution equations for $P(\mathbf{g}, t)$ in terms of the effective parameters $\Phi(t)$, $\phi_i(t)$ and $J_{ij}(t)$, it was shown in [9, 11, 12] that the distribution (4) is stable at a high rate of recombination. The values of the parameters Φ, ϕ_i and J_{ij} in stationary state are then related to the model parameters as discussed in detail in [7, 12]. In particular, J_{ij} is simply proportional to f_{ij} which can be turned around to the *KNS fitness inference formula*

$$f^*_{ij} = J^*_{ij} \cdot r c_{ij}. \tag{5}$$

The stars on both sides indicate that these are inferred quantities, and the proportionality parameters r and c_{ij} are discussed below.

In this work, we will extend the above analysis to the regime where recombination is not necessarily high, but mutation remains a faster process than selection. In section 3, we derive this within QLE, and in section 4 we do it by Gaussian closure. Here, we discuss and contrast these different (though related) formulae.

In QLE with mutation comparable to or larger than recombination, the relevant inference formula changes to

$$f^*_{ij} = J^*_{ij} \cdot (4\mu + r c_{ij}), \tag{6}$$

where μ is the rate of mutations assumed to be the same at all loci and in both directions. In both (5) and (6), the Gibbs–Boltzmann parameter J^*_{ij} is not directly observed, but has to be inferred from the data. All such procedures, collectively known either as inverse Ising/Potts or as DCA, have to make a trade-off between accuracy and computability. Let us mention here the benchmark statistical method of maximum likelihood, which is accurate but not efficiently computable in large systems, and naive mean-field (nMF) inference (described in appendix C), which amounts to matrix inversion of the empirical correlation matrix. Other procedures were reviewed in [13, 14] (see [15–17]). A particular DCA procedure introduced in [12] is small interaction expansion (SIE)

$$J^{*,\mathrm{SIE}}_{ij} = \chi_{ij} / \left(\left(1 - \chi_i^2\right) \left(1 - \chi_j^2\right) \right) \tag{7}$$

J. Stat. Mech. (2021) 083501

where $*$ stands for the type of inference used, and $\chi_i \equiv \langle s_i \rangle$ and $\chi_{ij} \equiv \langle s_i s_j \rangle - \chi_i \chi_j$ are the (connected) first and second order correlation functions. Inference formula (7) is not very accurate as a general DCA method [7, 12], but has the advantage of being eminently computable. Substituting (7) in (6), one gets

$$f_{ij}^{*,\mathrm{SIE}} = \frac{\chi_{ij}}{\left((1-\chi_i^2)\left(1-\chi_j^2\right)\right)} \cdot (4\mu + r c_{ij}) \,. \tag{8}$$

As it will turn out, (8) is also the formula which appears directly in Gaussian closure. We hence can derive (8) in two different ways.

In all of the above, the parameters μ, r and c_{ij} have the same meaning as in [12] and stand for mutation rate (assumed uniform), recombination rate (assumed uniform) and the probability of off-springs inheriting the genetic information from different parents. For high-recombination organisms, c_{ij} depends on the cross-over rate ρ and the genomic distance between loci i and j [8], except when loci i and j are very closely spaced on the genome.

$$c_{ij} \approx \frac{1}{2}\left(1 - \mathrm{e}^{-2\rho|i-j|}\right) . \tag{9}$$

When comparing (5) and (8) in numerical testing, we simulate an evolving population at the same parameter values, and then either use the genotype information to compute empirical correlations, or to infer Ising/Potts parameters by DCA. For simplicity, we will in the following only present results obtained by DCA nMF inference. The results are very similar for other common variants of DCA.

3. Quasi-linkage equilibrium outside high recombination

In this section, we introduce the model defined in [12] for the evolution of the distribution of genomes $P(\mathbf{g}, t)$ and discuss the high mutation regime in order to recover the inference formula for epistatic interactions (6). Throughout, we assume an infinite population; genetic drift is therefore not considered.

Selection is the first fundamental ingredient and works as follows: each possible sequence $\mathbf{g}$ grows inside the population with a certain growth-rate $F(\mathbf{g})$, called fitness, which can be described as a function of the specific sequence $\mathbf{g}$. As stated above in equation (3), we will approximate any fitness function F as the sum of linear terms f_i, called additive fitness, and pairwise interactions f_{ij}, called epistatic fitness. Note that in general, one can also include higher order terms of the form $f_{i_1 \ldots i_n} s_{i_1} \ldots s_{i_n}$.

The second ingredient for the population evolution is mutations. We assume that in each small time interval $\Delta t \ll 1$ a fraction $\mu \Delta t$ of all the alleles inside the population (L for each individual) mutates by a single spin-flip; μ is therefore named mutation rate. We describe the process of a spin flip by introducing an operator M_i acting on a sequence by changing the sign of the ith spin. To understand how the frequency of a certain sequence $\mathbf{g}$ changes in the interval Δt, we should count how many individuals have mutated into the sequence $\mathbf{g}$ and how many sequences have instead mutated from away this state.

J. Stat. Mech. (2021) 083501

The last element to consider is recombination between different sequences. At each small time interval Δt, a fraction $r\Delta t$ of the individuals (where r is the recombination rate) encounters random pairing and crossing-over, giving rise to new genomes. The evolution of the distribution P in the interval Δt due to recombination is given by

$$P(\mathbf{g}, t+\Delta t) = (1 - r\Delta t)P(\mathbf{g}, t) + r\Delta t \sum_{\{s_i'\}\{\xi_i\}} C(\{\xi\})P(\mathbf{g}^{(m)}, t)P(\mathbf{g}^{(f)}, t). \tag{10}$$

The first term counts for those individuals that did not recombine during the time interval Δt. When two individuals recombine, a new genotype is formed by inheriting some loci from the mother with genotype $\mathbf{g}^{(m)}$ and the complement from the father with genotype $\mathbf{g}^{(f)}$. The parts of the genomes of the mother and the father not inherited by the child (and hence discarded) are denoted as $\mathbf{g}'$. The crossover can be described by a vector $\{\xi_i\}$, with $\xi_i \in \{0, 1\}$. If $\xi_i = 1$, the ith locus is inherited from the mother, otherwise from the father. Turning around the relation, we have $s_i^{(m)} = s_i\xi_i + s_i'(1-\xi_i)$ and $s_i^{(f)} = s_i'\xi_i + s_i(1-\xi_i)$ where s_i is the allele of the child at locus i, and s_i' is the discarded allele. The probability of each realization of $\{\xi_i\}$ is given by $C(\{\xi\})$. Subsequently, we need to sum over all the possible genomes that are not passed on the offspring ($\mathbf{g}'$), as well as all the possible crossover patterns $\{\xi\}$ [7, 12].

Merging together all of the ingredients, we obtain the following non-linear differential equation for the time derivative of the genotype distribution P:

$$\begin{aligned}\dot{P}(\mathbf{g}, t) &= \frac{\mathrm{d}}{\mathrm{d}t}|_{\text{fitness}}\, P(\mathbf{g}, t) + \frac{\mathrm{d}}{\mathrm{d}t}|_{\text{mut}}\, P(\mathbf{g}, t) + \frac{\mathrm{d}}{\mathrm{d}t}|_{\text{rec}}\, P(\mathbf{g}, t) \\ &= \left[F(s) - \langle F\rangle\right] P(\mathbf{g}, t) + \mu \sum_{i=0}^{L-1} \left[P(M_i\mathbf{g}, t) - P(\mathbf{g}, t)\right] \\ &\quad + r \sum_{\{s_i'\}\{\xi_i\}} C(\{\xi\}) \left[P(\mathbf{g}^{(m)}, t)P(\mathbf{g}^{(f)}, t) - P(\mathbf{g}, t)P(\mathbf{g}', t)\right].\end{aligned} \tag{11}$$

Now, we want to study the stationary solutions of this master equation. In particular, we seek to extend Neher and Shraiman's argument [12] in the high mutation limit, recovering the inference formula (6) of the epistatic interactions introduced above. We start by assuming the distribution P to be of the same form as in equation (4):

$$P(g, t) = \frac{1}{Z(t)} \exp\left[\sum_i \phi_i(t)s_i + \sum_{i<j} J_{ij}(t)s_i s_j\right], \tag{12}$$

where $Z(t)$ is the normalization factor. Following Neher and Shraiman in [12], we now inject this ansatz in the master equation for the evolution of $\log P(g, t)$ in the presence of mutations and recombination and obtain

J. Stat. Mech. (2021) 083501

$$\frac{\mathrm{d}\,\log P(g,t)}{\mathrm{d}t} = -\frac{\mathrm{d}}{\mathrm{d}t}\log Z(t) + \sum_i \dot{\phi}_i(t)s_i + \sum_{i<j}\dot{J}_{ij}(t)s_i s_j$$
$$= F(g) - \langle F\rangle + \mu\underbrace{\sum_i\left[\frac{P(M_i g,t)}{P(g,t)} - 1\right]}_{M(g,t)}$$
$$+ r\underbrace{\sum_{\{\xi_i\}\{s'_i\}} C(\{\xi\})P(g',t)\left[\frac{P(g^{(m)},t)P(g^{(f)},t)}{P(g,t)P(g',t)} - 1\right]}_{R(g,t)}. \tag{13}$$

Now, we separate the mutation and recombination term ($M(g,t)$ and $R(g,t)$, respectively) from the rhs of the last equation and compute them separately. Starting from the recombination term, we may rewrite it as

$$R(g,t) = \sum_{\{\xi_i\}\{s'_i\}} C(\{\xi\})P(g',t)\left[\mathrm{e}^{\sum_{i<j}J_{ij}\left[(\xi_i\xi_j+\overline{\xi}_i\overline{\xi}_j-1)(s_is_j+s'_is'_j)+(\xi_i\overline{\xi}_j+\overline{\xi}_i\xi_j)(s_is'_j+s_is'_j)\right]} - 1\right], \tag{14}$$

where $\overline{\xi}_i = (1-\xi_i)$. Now, in the high recombination limit considered in [12] the authors suppose that the interactions J_{ij} are small and can be expanded from the exponential. We note that the same argument should also hold when mutations are dominant in the evolution. Hence, we write

$$R(g,t) \sim \sum_{\{\xi_i\}\{s'_i\}} C(\{\xi\})P(g',t)\left[\sum_{i<j}J_{ij}\left[(\xi_i\xi_j+\overline{\xi}_i\overline{\xi}_j-1)(s_is_j+s'_is'_j)\right.\right.$$
$$\left.\left.+(\xi_i\overline{\xi}_j+\overline{\xi}_i\xi_j)(s_is'_j+s_is'_j)\right]\right]$$
$$= \sum_{i<j}c_{ij}J_{ij}\left[(s_i\langle s_j\rangle + s_j\langle s_i\rangle) - (s_is_j + \langle s_is_j\rangle)\right], \tag{15}$$

where $c_{ij} \equiv \sum_\xi C(\{\xi\})\left[\xi_i\overline{\xi}_j + \overline{\xi}_i\xi_j\right]$ represents the probability that loci i and j arrive from different parents.

Now we turn to the mutation term $M(g,t)$ that can be written as follows:

$$M(g,t) = \sum_i\left[\mathrm{e}^{-2\phi_is_i - 2\sum_j J_{ij}s_is_j} - 1\right]. \tag{16}$$

In the high mutation regime, we suppose that both the interactions and the fields are small and can be expanded from the exponential:

J. Stat. Mech. (2021) 083501

$$M(g,t) \sim -2\sum_i \phi_i s_i - 4\sum_{i<j} J_{ij} s_i s_j. \tag{17}$$

Injecting these results for $M(g,t)$ and $R(g,t)$ into equation (13) and separating the dependencies on s_i and $s_i s_j$, we can obtain equations for $\dot{\phi}_i$ and $\dot{J}_{ij}$ similarly to what has been done by Neher and Shraiman in [12]. In particular, we find:

$$\dot{\phi}_i = f_i + r\sum_{j\neq i} c_{ij} J_{ij} \langle s_j \rangle - 2\mu\phi_i \tag{18}$$

$$\dot{J}_{ij} = f_{ij} - (4\mu + r c_{ij}) J_{ij}. \tag{19}$$

Hence, the interactions J_{ij} will quickly evolve through the stationary solution $J_{ij}^{\mathrm{st.}} = f_{ij}/(4\mu + r c_{ij})$. Inverting the latter equation, we recover the inference formula (6).

4. The argument by Gaussian closure

Going forward, we want to parameterize the distribution $P(\mathbf{g}, t)$ by its cumulants. In particular, we define the cumulants of first and second order as $\chi_i \equiv \langle s_i \rangle$ and $\chi_{ij} \equiv \langle s_i s_j \rangle - \chi_i \chi_j$. Note that in this way $\chi_{ii} = 1 - \chi_i^2$. Using equation (11), we can write the time evolution for these cumulants as follows:

$$\dot{\chi}_i = \langle s_i [F(\mathbf{g}) - \langle F \rangle] \rangle - 2\ \mu \chi_i \tag{20}$$

$$\dot{\chi}_{ij} = \langle (s_i - \chi_i)(s_j - \chi_j)[F(\mathbf{g}) - \langle F \rangle] \rangle - (4\mu + r c_{ij})\chi_{ij}, \tag{21}$$

with $i \neq j$ in the second line and c_{ij} defined as equation (9).

In general, equations (20) and (21) are not a closed set of equations since they would also depend on higher order cumulants χ_{ijk}, χ_{ijkl}, etc. The Gaussian closure that we introduced recently [19] aims to overcome this problem by neglecting those higher order cumulants (connected correlation functions) under the assumption that at high recombination and/or mutations rate their influence on the global dynamics is weak. For a Gaussian distribution, all cumulants of order higher than two vanish.

With this approximation, (20) and (21) define a closed set of $L(L+1)/2$ dynamical equations only depending on χ_i and χ_{ij}.

$$\dot{\chi}_i = \sum_j \chi_{ij} \left(f_j + \sum_k f_{jk} \chi_k - 2 f_{ij} \chi_i \right) - 2\mu\chi_i \tag{22}$$

$$\begin{aligned} \dot{\chi}_{ij} = &-2\chi_{ij} \sum_k \left[f_{ik}(\chi_{ik} + \chi_i \chi_k) + f_{jk}(\chi_{jk} + \chi_j \chi_k) \right] \ + 2 f_{ij} \chi_{ij} (\chi_{ij} + 2\chi_i \chi_j) \\ &+ \sum_{k,l} f_{kl} \chi_{ik} \chi_{jl} \ - (4\mu + r c_{ij})\chi_{ij} - 2\chi_{ij}(f_i \chi_i + f_j \chi_j) \end{aligned} \tag{23}$$

J. Stat. Mech. (2021) 083501

In principle, equations (22) and (23) could be simultaneously solved in order to determine the stationary state, which is of our interest, and this in turn would allow us to determine the $L(L+1)/2$ quantities $\{f_i\}, \{f_{ij}\}$ as a function of the $\{\chi_i\}, \{\chi_{ij}\}$. Unfortunately, considering the size of the system, this is analytically not feasible.

Nevertheless, equation (23) suggests another route to infer f_{ij} according to the following argument: when studying the stationary state, we can assume self-consistently that all the off-diagonal χ_{ij} are small, so that we can expand χ_{ij}, with $i \neq j$, as a power series of $1/(4\mu + rc_{ij})$:

$$\chi_{ij} = \frac{\chi_{ij}^{(1)}}{4\mu + rc_{ij}} + \mathcal{O}((4\mu + rc_{ij})^{-2}). \tag{24}$$

Inserting this in equation (23), we obtain

$$\chi_{ij}^{(1)} = f_{ij}(1 - \chi_i^2)(1 - \chi_j^2). \tag{25}$$

We therefore conclude that, to the first order,

$$\chi_{ij} = \frac{f_{ij}}{4\mu + rc_{ij}}(1 - \chi_i^2)(1 - \chi_j^2). \tag{26}$$

Turning around this into an inference formula for fitness, we arrive at (8).

5. Simulation strategies and results

The basic idea is to simulate the states of a population with N individuals (genome sequences) evolving under mutation, selection and recombination and genetic drift. As in previous work, we have used the FFPopSim package developed by Zanini and Neher for this purpose [20]. Simulation and parameter settings are given in appendix B.

In a QLE phase, the outcomes of such simulations are trajectories of means $\chi_i(t)$ and correlations $\chi_{ij}(t)$, which in principle can be computed from the configuration of the population $\mathbf{g}^{(s)}(t)$ at generation t. After a suitable relaxation period, we take the set $\mathbf{g}^{(s)}(t)$ to be independent samples from a distribution (4) with unknown direct couplings J_{ij}. We will use the DCA algorithm nMF throughout [23] to infer parameters J_{ij} from data for original KNS (for descriptions, see appendix C).

The principle of the numerical testing is to infer epistatic fitness parameters from the data by (5) and (8), and then compare to the underlying parameters f_{ij} used to generate the data. Here, the testing epistatic fitness is Sherrington–Kirkpatrick model [24] with different variations. The additive fitness f_i follows Gaussian distribution with zero means and the standard deviation $\sigma(\{f_i\}) = 0.05$ in our simulations. We note that (5) is proposed to hold for weak selection and high recombination, and has already been tested in [8]. Data availability is an issue. As in [8] we have used all-time versions of the algorithms, where samples $\mathbf{g}^{(s)}(t)$ at different t are pooled. This is primarily to mitigate the effect that in a real-world population the number of individuals N is very large, but in the simulations it is only moderately large. All DCA methods as well as empirical correlations can be more accurately estimated with more samples.

J. Stat. Mech. (2021) 083501

5.1. Mutation vs recombination rate

We start by taking a fixed fitness landscape (same f_{ij}) and systematically vary mutation and recombination (μ and r). Each sub-figure in figure 1 shows scatter plots for the KNS fitness inference formula (5) and formula (8) based on Gaussian closure vs the model parameter f_{ij} used to generate the data. These model parameters were independent Gaussian random variables specified by their standard deviation $\sigma(\{f_i\})$ and $\sigma(\{f_{ij}\})$ as hyper-parameters. The parameters J^*_{ij} that enter (5) are inferred by nMF.

The variations in figure 1 are such that each column has the same recombination rate in the order low-medium-high from left to right, and each row has the same mutation rate in the order low-medium-high from top to bottom. In the top row, both inference formulae work well, particularly for high recombination rate at the top right. In the middle and bottom rows, the KNS formula does not work, while the formula based on Gaussian closure still performs well, and in particular does not have systematic errors.

For comparison in more extensive parameter ranges, we have quantified inference performance by normalized root of mean square error

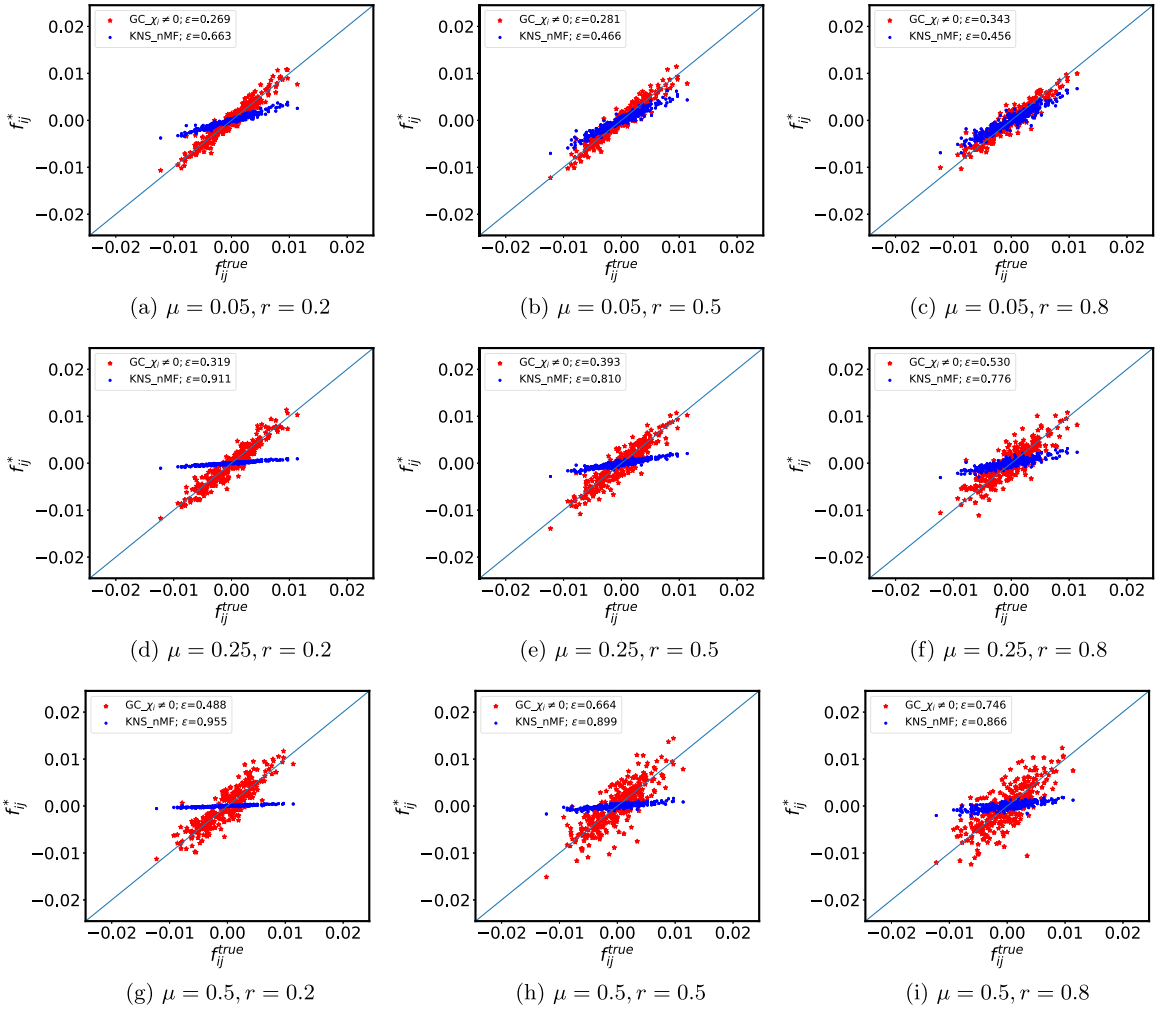
$$\epsilon = \sqrt{\frac{\sum_{ij}\left(f^*_{ij} - f_{ij}\right)^2}{\sum_{ij} f^2_{ij}}}. \tag{27}$$

We note that this reduces all of the information in the scatter plots in figure 1 to one single number. Although we have not observed such behaviour, it is conceivable that inference could be very accurate for most pairs (i, j) such that ϵ is small, but still has large errors for some few pairs. An overall value ϵ much less than one hence does not guarantee that fitness inference is accurate for all pairs. On the other hand, a large mean square error could correspond to either systematic or random errors in the scatter plots. We have observed both behaviours.

Anticipating a discussion that we will have in appendix B, we chose to visualize the dependence of ϵ on variation of μ by incorporating the coalescence time $\langle T_2 \rangle$, previously used in theoretical discussions of problems of the kind studied here [25, 26]. Figure 2 shows that is, at least for low epistatic fitness, a tendency reconstruction error to grow with μ. For the tests that have been carried out, it also appears that the threshold between the phase where fitness inference is possible takes place when there is about one mutation per coalescence time and a number of pairs of loci.

Returning now to the mapping out of regions where parameter inference is possible or not possible, we point to phase diagrams of ϵ shown in figure 3, for the KNS formula with nMF and the formula from Gaussian closure, respectively. The number of generations in simulations is set as $T = 10\,000$ and kept as a constant for all combinations of parameters. As in the scatter plots, we observe large differences as to two epistatic fitness inference formulae. In short, for linear structure of genomes, the KNS formula (5) works only for low mutation rate and high recombination rate (figure 1(c)). The new formula (8) from Gaussian closure instead works for a much larger region with weak fitness. The standard deviation of epistatic fitness $\sigma\left(\{f_{ij}\}\right) = 0.004$ in figure 3. We comment on the reasons for this effect in section 6. For a stronger mutation rate and larger recombination rate (data not shown), the root mean square error (ϵs) of inference based on the Gaussian closure formula increases, i.e. in that range this formula does not

J. Stat. Mech. (2021) 083501



(a) $\mu = 0.05, r = 0.2$ (b) $\mu = 0.05, r = 0.5$ (c) $\mu = 0.05, r = 0.8$

(d) $\mu = 0.25, r = 0.2$ (e) $\mu = 0.25, r = 0.5$ (f) $\mu = 0.25, r = 0.8$

(g) $\mu = 0.5, r = 0.2$ (h) $\mu = 0.5, r = 0.5$ (i) $\mu = 0.5, r = 0.8$

Figure 1. Scatter plots for testing and recovered f_{ij}s with mutation rate μ and recombination rate r. r increases from left to right columns (0.2, 0.5 and 0.8, respectively,) while μ enlarge from top to bottom (0.05, 0.25 and 0.5, respectively). The red stars for the Gaussian closed KNS $f^*_{ij} = \chi_{ij} \cdot (4\,\mu + rc_{ij}) / \left((1-\chi_i^2)(1-\chi_j^2)\right)$; blue dots for original KNS $f^*_{ij} = rc_{ij} \cdot J^{*,nMF}_{ij}$. Other parameters: $\sigma(\{f_i\}) = 0.05$, $\sigma(\{f_{ij}\}) = 0.004$, cross-over rate $\rho = 0.5$, number of loci $L = 25$, carrying capacity $N = 200$, number of generations $T = 10\,000$. Inference by Gaussian closed KNS works in a much wider parameter range than original KNS. One realization of the fitness terms f_{ij} and f_i for each parameter value.

work either. Specifically, the KNS formula (5) has severe systematic error, while formula (8) with Gaussian closure performs worse with heavier noise.

5.2. Fitness variations vs recombination rate

We continue by varying recombination r and the dispersion in the fitness landscape (f_{ij} drawn from Gaussian distributions with different hyper-parameters $\sigma(\{f_{ij}\})$). Each

J. Stat. Mech. (2021) 083501

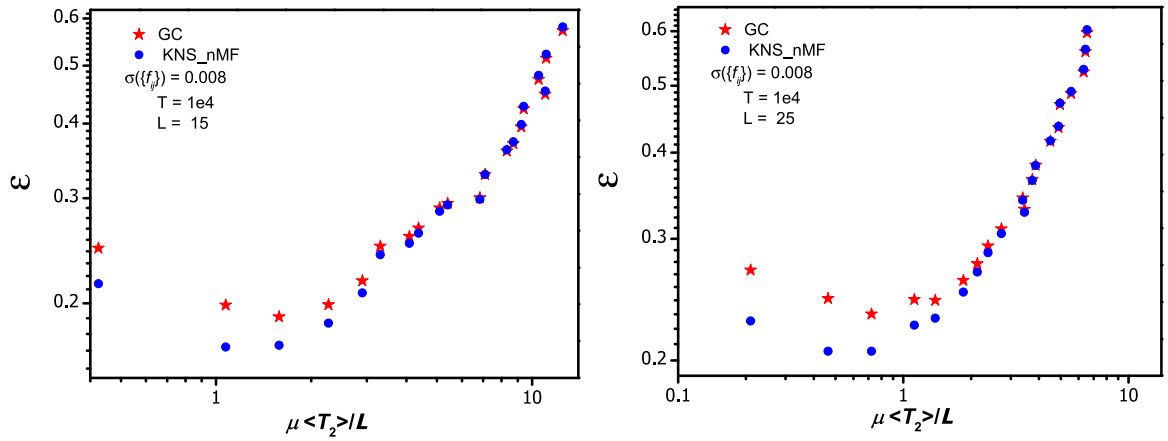


Figure 2. Epistasis reconstruction error ϵ versus $\mu\langle T_2\rangle/L$. Red stars for $f^*_{ij} = \chi_{ij} \cdot (4\mu + rc_{ij})/((1-\chi_i^2)(1-\chi_j^2))$ while blue dots for $f^*_{ij} = J^{*,nMF}_{ij} \cdot (4\mu + rc_{ij})$. Epistasis f_{ij} are inferred best with $\mu = 0.05$. The other parameter values: $\sigma(f_i) = 0.05$, $\sigma(f_{ij}) = 0.008$, carrying capacity $N = 200$, out-crossing rate $r = 0.5$, cross-over rate $\rho = 0.5$, number of loci $L = 15$, generations $T = 10\,000$. Ten realizations of the fitness terms f_{ij} and f_i for each parameter value.

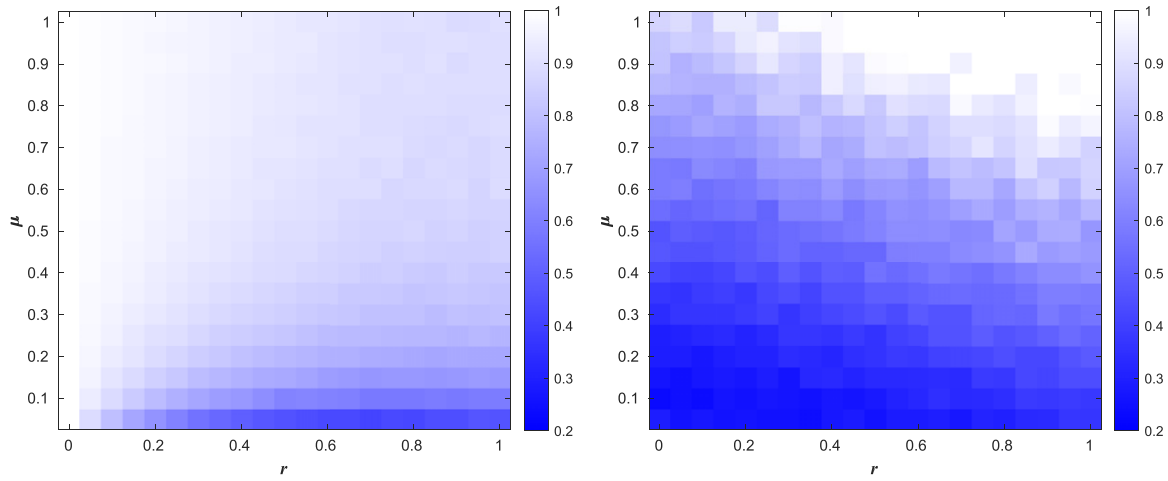
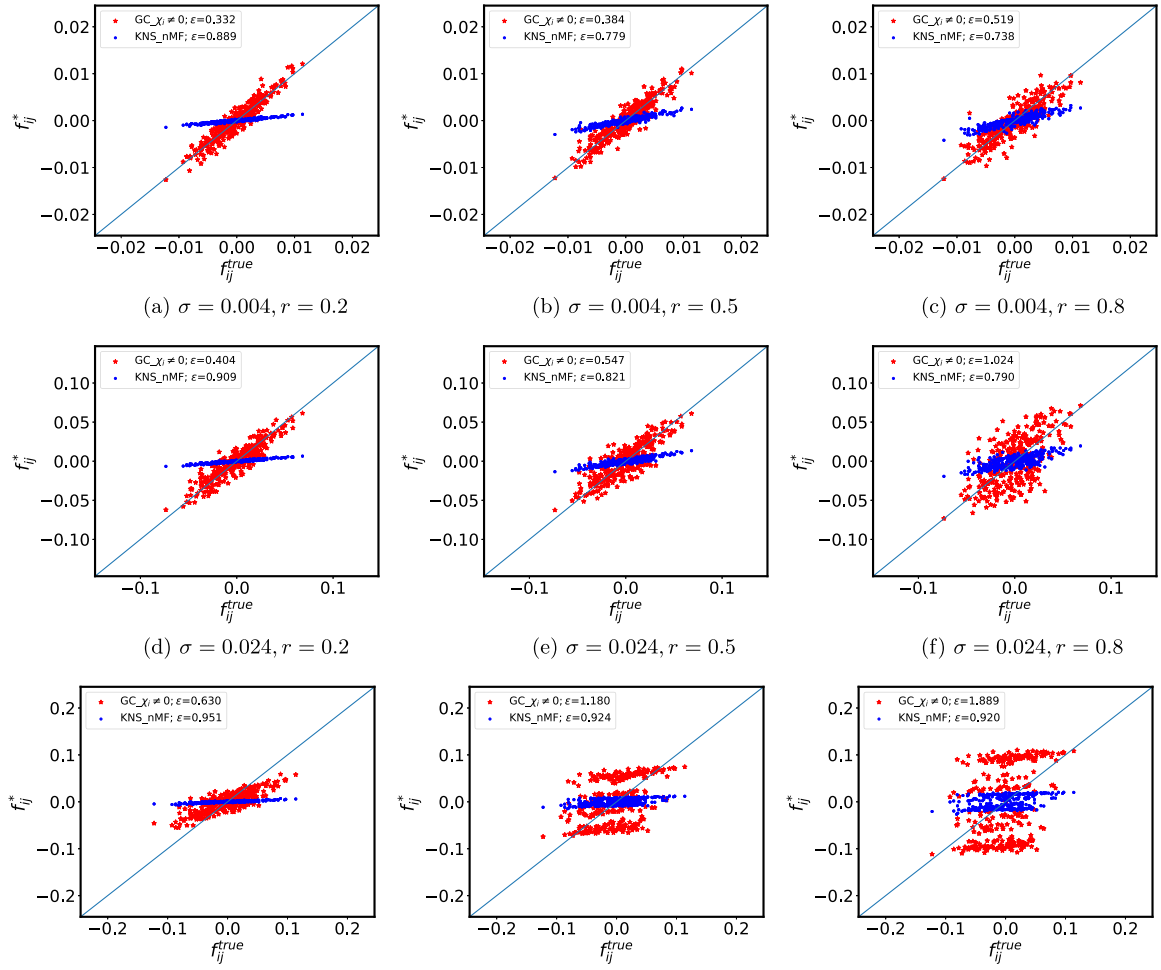


Figure 3. Phase diagram for mutation rate μ versus recombination rate r. The color is encoded by the reconstruction error ϵ given in equation (27). (Left) KNS theory $f_{ij} = J^{*,nMF}_{ij} \cdot rc_{ij}$. (Right) Gaussian closed KNS theory $f_{ij} = \chi_{ij} \cdot (4\mu + rc_{ij})/\left((1-\chi_i^2)(1-\chi_j^2)\right)$. Parameters: $\sigma(\{f_i\}) = 0.05$, $\sigma(\{f_{ij}\}) = 0.004$, cross-over rate $\rho = 0.5$, number of loci $L = 25$, carrying capacity $N = 200$, generations $T = 10\,000$. One realization of the fitness terms f_{ij} and f_i for each parameter value.

sub-figure in figure 4 shows scatter plots for the two epistatic fitness inference formulae for the model parameter $\sigma(\{f_{ij}\})$ vs the recombination rate r. The order in figure 4 is increasing recombination rate r in the columns from left to right, and increasing

J. Stat. Mech. (2021) 083501



(a) $\sigma = 0.004, r = 0.2$ (b) $\sigma = 0.004, r = 0.5$ (c) $\sigma = 0.004, r = 0.8$

(d) $\sigma = 0.024, r = 0.2$ (e) $\sigma = 0.024, r = 0.5$ (f) $\sigma = 0.024, r = 0.8$

(g) $\sigma = 0.04, r = 0.2$ (h) $\sigma = 0.04, r = 0.5$ (i) $\sigma = 0.04, r = 0.8$

Figure 4. Scatter plots for testing and reconstructed f_{ij}s. The standard deviation $\sigma(\{f_{ij}\}^{\text{true}})$ increases from top to bottom rows $(0.004, 0.024$ and 0.04 respectively) and recombination rate r enlarges in columns from left to right $(0.2, 0.5$ and 0.8, respectively). Red stars for $f_{ij}^* = \chi_{ij} \cdot (4\,\mu + r c_{ij})/((1-\chi_i^2)(1-\chi_j^2))$ and blue dots for $f_{ij}^* = J_{ij}^{*,\text{nMF}} \cdot r c_{ij}$. Both inference formulae do not work for large σ and high r, where strong correlations emerge between loci that drive the system out of the QLE phase [27, 28], as shown in (g), (h) and (i). The other parameter values: standard deviation $\sigma(\{f_i\}) = 0.05$, mutation rate $\mu = 0.2$, cross-over rate $\rho = 0.5$, number of loci $L = 25$, carrying capacity $N = 200$, generations $T = 10\,000$. One realization of the fitness terms f_{ij} and f_i for each parameter value.

$\sigma(\{f_{ij}\})$ in the rows from top to bottom. Here, the mutation rate $\mu = 0.2$ and the other parameters are the same as those tested in figure 1.

Overall, the KNS formula (5) does not work for any of the parameter values shown in figure 4 with mutation rate $\mu = 0.2$.

This either because of systematic errors as in the top row (low fitness dispersion) and left column (low recombination), or due to the emergence of strong correlation between

J. Stat. Mech. (2021) 083501

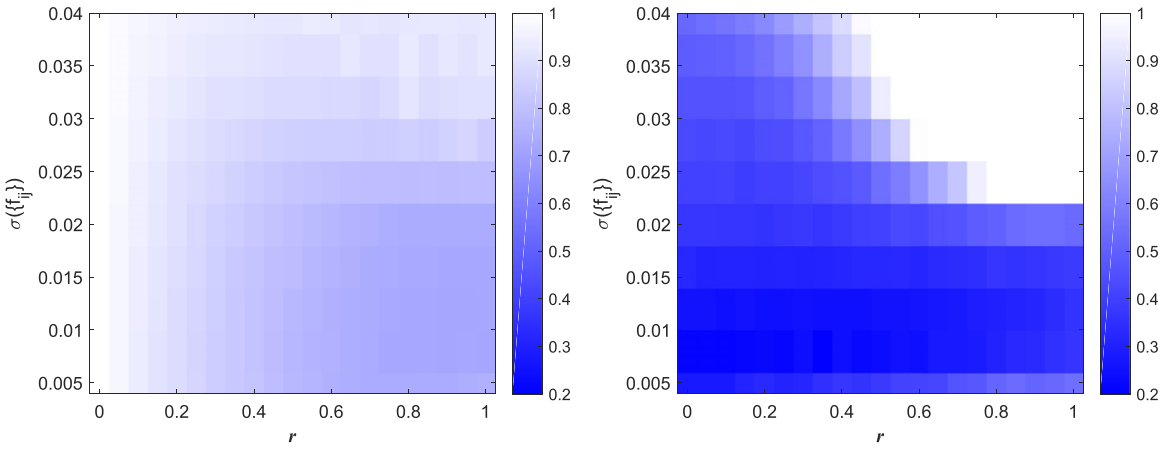


Figure 5. Phase diagram for the standard deviation $\sigma(\{f_{ij}\})$ versus recombination rate r. (Left) KNS theory $f^*_{ij} = J^{\mathrm{nMF}}_{ij} \cdot rc_{ij}$. (Right) Gaussian closed KNS theory $f^*_{ij} = \chi_{ij} \cdot (4\mu + rc_{ij})/((1-\chi_i^2)(1-\chi_j^2))$. Parameters: mutation rate $\mu = 0.2$, crossover rate $\rho = 0.5$, number of loci $L = 25$, carrying capacity $N = 200$, generations $T = 10\,000$. One realization of the fitness terms f_{ij} and f_i for each parameter value.

loci that drive the evolution out of the QLE regime [27, 28], as in the bottom right corner in figures 4(h) and (i). The Gaussian closure formula (8) in contrast works well for low recombination or low fitness dispersion or both, but fails as well for sufficiently high recombination and fitness strength.

As above, we have quantified inference performance in larger parameter ranges by the root of mean square error ϵ. The phase diagrams in figure 5 show again that the Gaussian closure formula works except when r and $\sigma(\{f_{ij}\})$ are both large, while the KNS formula does not work in any range with mutation rate $\mu = 0.2$.

6. Discussion

In this paper, we have pursued the investigations started by Kimura in 1965 [9] on how epistatic contributions to fitness is reflected in the distribution over genotypes in a population. Our perspective is that of fitness inference: we assume that the distribution is observable from many whole-genome sequences of an organism and ask what we can learn about synergistic effects on fitness from concurrent allele variations at different loci, i.e. about epistasis. Our benchmark has been the generalization of the Kimura theory by Neher and Shraiman to a phase of genome-scale QLE [7, 11, 12]. In recent work, we showed in numerical testing that a central formula describing the QLE phase allows us to retrieve epistatic contributions to fitness in the limit of high recombination [8].

Here, we have extended these considerations to the regime where mutation can be a stronger (faster) process than recombination. We have done this on one hand by generalizing the derivation from the assumptions of a QLE state in [12], and on the other

J. Stat. Mech. (2021) 083501

by following the analogy to physical kinetics and approximations of the Boltzmann equation, recently developed by three of us [18, 19]. In the second approach, we consider the evolution equations for single-locus and two-loci frequencies in a population evolving under selection, mutation, genetic drift and recombination, and then close those equations by setting higher-order cumulants to zero (Gaussian closure).

We hence do not need to make explicit assumptions on the functional form of the distribution over genotypes in a population, only that it is possible to treat it as a Gaussian for the purpose of evaluating higher moments. Methods of a similar type were earlier developed for a more macroscopic level approach that does not assume access to individual genotypes, and therefore also cannot be tested on that level [29].

The phase diagrams in figures 3 and 5 in section 5 show that the new inference formula dramatically outperforms previous one, with the exception of a region at strong fitness and high recombination rate. Since in many biological systems mutations can be at least as strong as recombination, we have extended the range of epistasis inference methods considerably. This is the main result of the current work. We have also performed preliminary investigations of a possible dependence on the phase boundary on coalescence time, a quantity which roughly measures the time to a common ancestor for the whole population. In earlier theoretical work based either on the infinitesimal model of genetics or on more macroscopic considerations, it has been predicted that the transition occurs at about one mutation per coalescence time per locus pair [25, 26]. To the extent that we have been able to test this hypothesis, we find that it holds also for our procedure where inference is assessed for each epistatic pair and its associated epistatic fitness parameter separately (see figure 2).

A theoretical advantage of the extension presented here follows from the fact that the previous inference formula (5) is obtained by perturbation in the inverse of recombination rate, i.e. equation (23) and appendix B in [12]. In a finite population this derivation requires that mutations be so much weaker compared to recombination that they can be neglected for quantitative properties in QLE, while still being non-zero. The latter restriction is necessary as otherwise the fittest genotype will eventually take over the population, and the QLE phase will only be a long-lived transient [7, 8]. One consequence of a low mutation rate is that any imbalance in total epistatic fitness will lead to almost fixated alleles. In the QLE phase where equation (5) can be used quantitatively, the first order moments χ_i are therefore typically different than zero. Inference formula (8) is on the other hand obtained by expanding the equations of Gaussian closure under conditions appropriate for high mutation rate, and χ_i does not necessarily need to be zero either. A further assumption to arrive at (8) is that epistatic fitness variations are not too strong, qualitatively $L\sigma(\{f_{ij}\}) < 1$. Moreover, the additive fitness should be sufficiently weak as well to make sure the population is strictly mono-clonal, which is one of the assumptions of the Gaussian closure [18, 19]. Data shown in bottom row of figures 4(g), (h) and (i) have $L\sigma(\{f_{ij}\}) \approx 1$.

In conclusion, we have presented an extension of the classic KNS theory, which allows us to reliably infer epistatic contributions to fitness. In a separate work, three of us recently applied the method to more than 50 000 full-length genomes of the SARS-CoV-2 virus [30], and were able to predict new epistatic interactions between eight viral

J. Stat. Mech. (2021) 083501

genes, many involving ORF3a, a protein implicated in severe manifestations of COVID-19 disease. Methodological development of epistasis analysis using DCA, as we have discussed here, may hence also have practical applications of some impact on society.

Acknowledgments

We are grateful to Guilhem Semerjian for valuable input and suggestions. We also acknowledge constructive criticism of an anonymous referees, which allowed us to significantly improve our work. The work of H L Z was sponsored by National Natural Science Foundation of China (11705097), Natural Science Foundation of Jiangsu Province (BK20170895), Jiangsu Government Scholarship for Overseas Studies of 2018. The work of E M was supported by ICFP fellowship and École Normale Superieure (Paris). The work of V D was supported by Extra-Erasmus Scholarship (Department of Physics, University of Trieste) and Collegio Universitario 'Luciano Fonda'. He also warmly thanks Nordita (Stockholm, Sweden) and K T H (Stockholm, Sweden) for hospitality. S C and R M acknowledge financial support from the Agence Nationale de la Recherche projects RBMPro (ANR-17-CE30-0021) and Decrypted (ANR-19-CE30-0021). E A acknowledges the Science for Life Labs (Solna, Sweden) 'Viral sequence evolution research program'.

Appendix A. Higher order corrections to the Gaussian closure inference formula

Starting from equation (23) and exploiting the same argument as in the main body of the paper, it is straightforward to compute higher order terms in the expansion for χ_{ij}. Let us define for simplicity $\epsilon = 1/(4\mu + rc_{ij})$. In the limit $\epsilon \to 0^+$, we write

$$\chi_{ij} = \epsilon\chi_{ij}^{(1)} + \epsilon^2\chi_{ij}^{(2)} + \epsilon^3\chi_{ij}^{(3)} + \mathcal{O}(\epsilon^4), \tag{A.1}$$

and in the case where $f_i = 0$ for all i we find

$$\chi_{ij}^{(1)} = f_{ij} \tag{A.2}$$

$$\chi_{ij}^{(2)} = 2\sum_k f_{ik}f_{jk} \tag{A.3}$$

$$\chi_{ij}^{(3)} = \sum_{k<l} f_{kl}(f_{ik}f_{jl} + f_{jk}f_{il}) - f_{ij}^3 + \sum_k \left[f_{ik}\left(2\sum_l f_{kl}f_{jl} - f_{ij}f_{ik}\right) + f_{jk}\left(2\sum_l f_{kl}f_{il} - f_{ij}f_{jk}\right)\right]. \tag{A.4}$$

We observe that each correction is of order $L \times \sigma(\{f_{ij}\})$ with respect to the lower one, therefore we expect the expansion to be accurate only if $L \times \sigma(\{f_{ij}\}) \ll 1$.

J. Stat. Mech. (2021) 083501

Table 1. Main default parameters of FFPopSim used in the simulations.

Number of loci (L)	25
Number of traits	1
Circular	False
Carrying capacity (N)	200
Generation	10 000
Recombination model	CROSSOVERS
Crossover rate (ρ)	0.5
Fitness additive (coefficients)	Gaussian random number with $\sigma(\{f_i\}) = 0.05$

In the more general case where $f_i \neq 0$ for all i, we find the first two orders in equation (A.1) to be:

$$\chi_{ij}^{(1)} = f_{ij}(1-\chi_i^2)(1-\chi_j^2) \tag{A.5}$$

$$\chi_{ij}^{(2)} = \sum_k f_{ik}\left(\chi_{jk}^{(1)} + \chi_{ik}^{(1)}\chi_i\chi_j - \chi_i\chi_k\chi_{ij}^{(1)}\right) - \sum_{k,l} f_{kl}\chi_i\chi_l\chi_{jk}^{(1)} - \sum_l f_{il}\chi_l\chi_i\chi_{ij}^{(1)} + \sum_{k<l} f_{kl}\left(\chi_{ik}^{(1)}\chi_j\chi_l + \chi_{il}^{(1)}\chi_j\chi_k - 2f_i\chi_i\chi_{ij}^{(1)} + f_{ij}\chi_i\chi_j\chi_{ij}^{(1)}\right) + \{i \leftrightarrow j\} \tag{A.6}$$

where, for the sake of clarity, in the last equation we have left implicit the terms like $\chi_{ij}^{(1)}$ as specified in equation (A.5).

Appendix B. FFPopSim settings

The FFPopSim package, written by Zanini and Neher simulates a population evolving due to mutation, selection and recombination [20].

We use the class `haploid_highd`, i.e. individual-based simulations that handle the population as a set of clones $(g_i, n_i(t))$, where g_i is a genotype and $n_i(t)$ is the number of individuals with genotype g_i at time t (only existing clones are tracked). At each generation, the size of each clone is first updated $n_i(t) \to n_i(t+1) \sim \mathscr{P}_\lambda$, where $\mathscr{P}$ is the Poisson distribution with parameter $\lambda = \frac{1}{\langle e^F \rangle} e^{F(g_i)+1-\frac{1}{N}\sum_j n_j(t)}$, N is the carrying capacity and $F(g)$ is the fitness function. A fraction r^* (outcrossing rate) of the resulting offspring is destined to the recombination step, paired and reshuffled. Finally, each individual is allowed to mutate with probability $1 - e^{-L\mu}$, where μ is the recombination rate, the exact number of mutations being Poisson distributed $\mathscr{P}_{L\mu}$.

We have used FFPoSim in a similar manner as in [8] and we will only list the settings here. Parameters that are the same in all simulations reported in this paper are listed in table 1. Parameters that have been varied (not all variations reported in the paper) are listed in table 2.

J. Stat. Mech. (2021) 083501

Table 2. Variable parameters of FFPopSim used in the simulation.

Initial genotypes	Binary random numbers
Out-crossing rate (r)	$[0, 1.0]$
Mutation rate (μ)	$[0.05, 0.5]$
Epistatic fitness	Gaussian random number with $\sigma(\{f_{ij}\}) \in [0.004, 0.04]$

It is important to notice that the out-crossing rate r^* in FFPopSim *a priori* differs from our recombination rate, r, appearing in equation (8). In the simulation package, dynamics are discrete in time (with time step of one generation) and r^* is a probability taking value between 0 and 1. In our theory, r is a rate that can take any positive value. In the examples given in [20], e.g. figure 2 in the article's main text and figure 2 in its supplementary information, the out-crossing probability does not exceed 10^{-2}. For such low values r^* coincides with a rate (since the time step is equal to unity), which justifies its denomination. We use this correspondence $r^* = 1 - \mathrm{e}^{-r} \sim r$ between the out-crossing rate r^* in FFPopSim and our recombination rate r, valid for small values, to produce the scatter plots in figures 1 and 4.

Notice that this correspondence breaks down for large recombination rates. Indeed, even for out-crossing rate $r^* = 1$ in the simulation package, mutations and fitness effects can still be quite large, depending on the values of the f_{ij}'s and of μ, and QLE is not recovered. In the theory, however, all fitness and mutation effects become relatively weak, of the order of $1/r$.

In addition to forward simulations, a subsequent release of the original FFPopSim package allows for the possibility of tracking the genealogy of loci, e.g. that of the central locus. Such information can be used in the first place to draw a coalescent tree [25]: technically, this is done by converting the genealogy in a `BioPython` tree and using the module `Bio.Phylo` for plotting purposes.

A quantitative analysis of such trees can be carried out, for instance the time to the most recent common ancestor (MRCA) T_{MRCA} of a group of individuals at time t is nothing but the temporal distance of the leaves (individuals) from the root (common ancestor) in the corresponding coalescent tree $\mathcal{CT}$, as shown in figure 6. In the same vein, we are able to evaluate the average pair coalescent time $\langle T_2 \rangle$: we sample $n_2 = 10$ pairs of leaves. For each of them we extract the information about their subtree $\mathcal{CT}_2$ and evaluate the $T_{\mathrm{MRCA}}(\mathcal{CT}_2)$, which now corresponds to the difference between the present and the time in the past when the two branches stemming from the chosen leaves merge. Averaging over the sample of size n_2 gives an estimate of the desired quantity.

Appendix C. Naive mean-field (nMF)

nMF is based on minimizing the reverse Kullback–Leibler distance between an empirical probability distribution and a trial distribution in the family of independent (factorized) distributions. This leads to the inference formula $J_{ij}^{*,n\mathrm{MF}} = (\chi^{-1})_{ij}$. If the correlation χ_{ij}

J. Stat. Mech. (2021) 083501

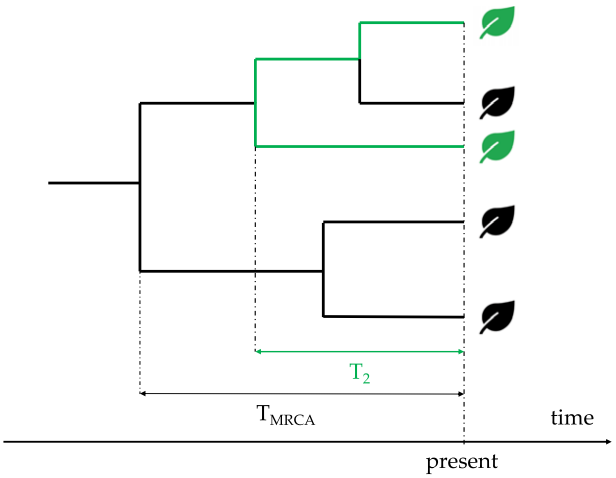


Figure 6. Illustrative coalescent tree. The time to the MRCA T_{MRCA} of a tree is the difference between the current time and the time point where all of the branches merge. The pair coalescent time T_2 for two chosen leaves (individuals) is the T_{MRCA} with respect to their subtree (highlighted in green).

Algorithm 1. Epistatic fitness inference by KNS formula (5) with J^*_{ij} reconstructed by nMF procedure: f^{nMF}_{ij}.

Input: mean correlations: $\langle \chi_{ij} \rangle$
Output: inferred epistatic fitness: f^{nMF}_{ij}
1: import **scipy**
2: from **scipy** import **linalg**
3: $J^{\mathrm{nMF}}_{ij} = -\mathrm{linalg.inv}(\langle \chi_{ij} \rangle)$
4: $f^{\mathrm{nMF}}_{ij} = J^{\mathrm{nMF}}_{ij} * rc_{ij}$

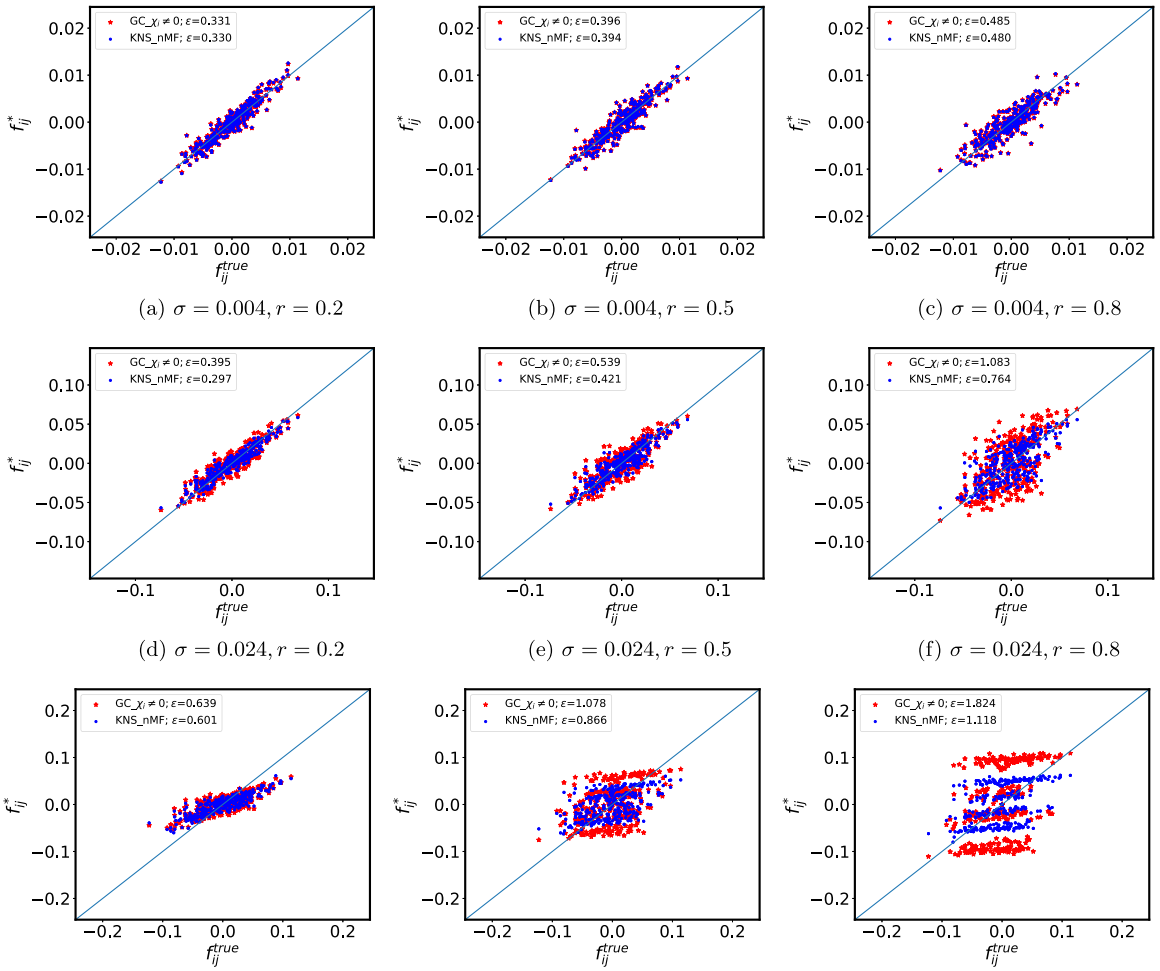
is computed as an average over the population at a single time, we call it single-time-nMF. If on the other hand χ_{ij} is computed by additionally averaging over time, we call it all-time-nMF.

The pseudo-code for nMF inference taking χ_{ij} as input is presented in algorithm 1.

Appendix D. Numerical comparison between equations (6) and (8)

To compare the results of epistasis inference by equation (6) from KNS theory and equation (8) through Gaussian closure, we present the numerical simulations in figure 7 with a fixed mutation rate $\mu = 0.2$ while different $\sigma(\{f_{ij}\})$s and recombination rate r. The blue dots are for equation (6), while the red stars for equation (8). As shown in the top row of figures 7(a)–(c), two methods perform almost the same for weak epistatic fitness $\sigma(\{f_{ij}\}) = 0.004$. When increasing $\sigma(\{f_{ij}\})$ and for sufficiently low recombination rates as in figures 7(d), (e) and (g), we observe that (6) works considerably better than equation (8), as is evident from the smaller reconstruction error of the former with respect to the latter. Finally, none of them work for large $\sigma(\{f_{ij}\})$ and high r as shown

J. Stat. Mech. (2021) 083501



(a) $\sigma = 0.004, r = 0.2$ (b) $\sigma = 0.004, r = 0.5$ (c) $\sigma = 0.004, r = 0.8$

(d) $\sigma = 0.024, r = 0.2$ (e) $\sigma = 0.024, r = 0.5$ (f) $\sigma = 0.024, r = 0.8$

(g) $\sigma = 0.04, r = 0.2$ (h) $\sigma = 0.04, r = 0.5$ (i) $\sigma = 0.04, r = 0.8$

Figure 7. Scatter plots for testing and reconstructed f_{ij}s. The standard deviation $\sigma(\{f_{ij}\}^{\text{true}})$ increases from the top to bottom rows $(0.004, 0.024$ and 0.04 respectively) and recombination rate r enlarges in columns from left to right $(0.2, 0.5$ and 0.8 respectively). Red stars for $f^*_{ij} = \chi_{ij} \cdot (4\mu + rc_{ij})/((1-\chi_i^2)(1-\chi_i^2))$ and blue dots for $f^*_{ij} = (4\mu + rc_{ij}) \cdot J^{*,nMF}_{ij}$. The other parameters are the same to those in figure 4. In the regime of weak σ and r, the reconstructions are equivalent. Increasing σ for sufficiently small r as in (d), (e) and (g) the mean field reconstruction outperforms the Gaussian one. However, both reconstructions fail for sufficiently high σ, r, as in (f)–(i), where strong correlations emerge between loci that drive the system out of the QLE phase [27, 28]. One realization of the fitness terms f_{ij} and f_i for each parameter value.

in figures 7(f), (h) and (i). The parameters for these cases are located in the white area of figure 5, where the system may not be in the QLE state and both of the reconstructions (Neher–Shraiman and Gaussian closure) fail. This part with strong correlations has been studied extensively in V D's Master's Thesis [27, 28].

J. Stat. Mech. (2021) 083501

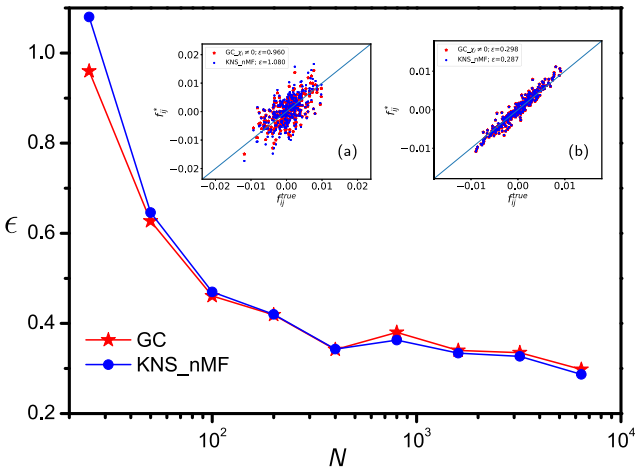


Figure 8. Semi-log plot for epistasis reconstruction error ϵ versus the average size of population N. (a) Scatter plot for testing and reconstructed f_{ij}s with $N = L = 25$. (b) Scatter plot with $N = 6400$. Red stars for $f^*_{ij} = \chi_{ij} \cdot (4\mu + rc_{ij})/((1-\chi_i^2)(1-\chi_j^2))$ and blue dots for $f^*_{ij} = J^{*,nMF}_{ij} \cdot (4\mu + rc_{ij})$. Epistasis f_{ij} are recovered roughly better with increasing N. The other parameter values: $\sigma(f_i) = 0.05$, $\sigma(f_{ij}) = 0.004$, mutation rate $\mu = 0.25$, out-crossing rate $r = 0.5$, cross-over rate $\rho = 0.5$, number of loci $L = 25$, carrying capacity $N = 200$, generations $T = 10\,000$. One realization of the fitness terms f_{ij} and f_i for each parameter value.

Appendix E. Effects of genetic drift

The effects of genetic drift on epistasis effects are studied through the inference error ϵ with different population sizes N. It is presented in a semi-log plot, as shown in the main panel of figure 8. The red stars are for the epistasis inference error given by equation (8) $f^*_{ij} = \chi_{ij} \cdot (4\mu + rc_{ij})/((1-\chi_i^2)(1-\chi_j^2))$ while blue dots for equation (6) $f^*_{ij} = J^{*,nMF}_{ij} \cdot (4\mu + rc_{ij})$. There is a clear trend that both methods work better with increasing population sizes. However, equation (8) works slightly better when the population size is less than 400 while equation (6) recovers the epistasis better when $N > 400$. The inserts (a) and (b) of figure 8 show the scatter plots for the recovered and testing epistasis f_{ij}s with $N = 25$ (an equal number to that of locus in an individual sequence) and $N = 6400$, respectively. Clearly, both equations recover the epistasis better with large population size when compared to those with small ones.

Appendix F. Epistasis inference with directional selections

This appendix summarizes the effects of non-zero additive fitness on epistasis inference through numerical simulations. Here, the additive effects f_is are Gaussian distributed with non-zero means and the standard deviations are fixed as $\sigma(\{f_i\}) = 0.05$. The red stars are for the epistasis inference with Gaussian closure in equation (8), while the blue dots are for the revised KNS method in equation (6). The inserts of figure 9 show the scatter plots for the recovered and testing epistasis effects with (a) $\langle f_i \rangle = 0.001$ and (b) $\langle f_i \rangle = 0.01$, respectively. The other parameters for each point in the main panel are as

J. Stat. Mech. (2021) 083501

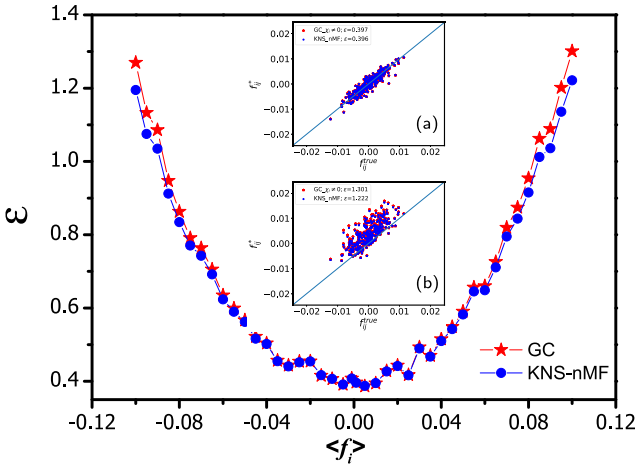


Figure 9. Epistasis reconstruction error ϵ versus the means of Gaussian distributed additive fitness $\langle f_i \rangle$. (a) Scatter plot for testing and reconstructed f_{ij}s with $\langle f_i \rangle = 0.001$. (b) Scatter plot with $\langle f_i \rangle = 0.01$. Red stars for $f^*_{ij} = \chi_{ij} \cdot (4\mu + rc_{ij})/((1-\chi_i^2)(1-\chi_j^2))$ and blue dots for $f^*_{ij} = J^{*,\mathrm{nMF}}_{ij} \cdot (4\mu + rc_{ij})$. The epistasis reconstructions are getting worse with stronger directional fields. The other parameter values: standard deviation $\sigma(\{f_{ij}\}) = 0.004$, mutation rate $\mu = 0.25$, out-crossing rate $r = 0.5$, cross-over rate $\rho = 0.5$, number of loci $L = 25$, carrying capacity $N = 200$, number of generations $T = 10\,000$. One realization of the fitness terms f_{ij} and f_i for each parameter value.

follows: standard deviation of the pairwise epistasis fitness $\sigma(\{f_{ij}\}) = 0.004$ and that of the single-locus additive fitness $\sigma(\{f_i\}) = 0.05$, mutation rate $\mu = 0.25$, out-crossing rate $r = 0.5$, cross-over rate $\rho = 0.5$, number of loci $L = 25$, carrying capacity $N = 200$, generations $T = 10\,000$.

Both methods recover the tested epistasis better with weaker means of additive fitness compared to that following stronger directional selections. It is notable that the reconstructed epistasis have a roughly corrected trends with large additive fitness, as shown in figure 9(b) for $\langle f_i \rangle = 0.01$. This may indicate the revision of the epistasis inference formulae in our work for stronger directional selections.

References

[1] de Visser J A G M and Krug J 2014 *Nat. Rev. Genet.* **15** 480–90
[2] Smith J M 1982 *Evolution and the Theory of Games* (Cambridge: Cambridge University Press)
[3] Chastain E, Livnat A, Papadimitriou C and Vazirani U 2014 *Proc. Natl Acad. Sci.* **111** 10620–3
[4] Shu Y and McCauley J 2017 *Euro Surveill.* **22**
[5] Bedford T and Neher R 2015–2020 Nextstrain (https://nextstrain.org/)
[6] Hadfield J, Megill C, Bell S M, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T and Neher R A 2018 *Bioinformatics* **34** 4121–3
[7] Gao C-Y, Cecconi F, Vulpiani A, Zhou H-J and Aurell E 2019 *Phys. Biol.* **16** 026002
[8] Zeng H L and Aurell E 2020 *Phys. Rev.* E **101** 052409
[9] Kimura M 1965 *Genetics* **52** 875–90
[10] Kirkpatrick M, Johnson T and Barton N 2002 *Genetics* **161** 1727–50

J. Stat. Mech. (2021) 083501

[11] Neher R A and Shraiman B I 2009 *Proc. Natl Acad. Sci.* **106** 6866–71
[12] Neher R A and Shraiman B I 2011 *Rev. Mod. Phys.* **83** 1283–300
[13] Roudi Y, Aurell E and Hertz J A 2009 *Front. Comput. Neurosci.* **3** 1–15
[14] Nguyen H C, Zecchina R and Berg J 2017 *Adv. Phys.* **66** 197–261
[15] Morcos F *et al* 2011 *Proc. Natl Acad. Sci.* **108** E1293–301
[16] Stein R R, Marks D S and Sander C 2015 *PLoS Comput. Biol.* **11** e1004182
[17] Cocco S, Feinauer C, Figliuzzi M, Monasson R and Weigt M 2018 *Rep. Prog. Phys.* **81** 032601
[18] Mauri E 2019 Population genetics and epistasis: a Gaussian approximation for allele dynamics *Master ENS ICFP Internship Report* École Normale Superieure
[19] Mauri E, Cocco S and Monasson R 2021 *Europhys. Lett.* **132** 56001
[20] Zanini F and Neher R A 2012 *Bioinformatics* **28** 3332–3
[21] Fisher R A 1930 *The Genetical Theory of Natural Selection* (Oxford: Clarendon)
[22] Blythe R A and McKane A J 2007 *J. Stat. Mech.* P07018
[23] Kappen H J and Rodríguez F B 1998 *Neural Comput.* **10** 1137–56
[24] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792–6
[25] Neher R A, Kessinger T A and Shraiman B I 2013 *Proc. Natl Acad. Sci.* **110** 15836–41
[26] Held T, Klemmer D and Lässig M 2019 *Nat. Commun.* **10** 2472
[27] Dichio V 2020 Statistical genetics and DCA inference beyond the Quasi linkage Equilibrium *Master's Thesis* University of Trieste, Italy
[28] Dichio V, Zeng H L and Aurell E 2021 Statistical genetics within and beyond the quasi-linkage equilibrium (arXiv:2105.01428)
[29] Nourmohammad A, Schiffels S and Lässig M 2013 *J. Stat. Mech.* P01012
[30] Zeng H-L, Dichio V, Rodríguez Horta E, Thorell K and Aurell E 2020 *Proc. Natl Acad. Sci.* **117** 31519–26

J. Stat. Mech. (2021) 083501